ACTAS DEL III CONGRESO INTERNACIONAL DE LINGÜÍSTICA DE CORPUS

LAS TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES: PRESENTE Y FUTURO EN EL ANÁLISIS DE CORPUS

Editores:
María Luisa Carrió Pastor
Miguel Ángel Candel Mora

Editores
María Luisa Carrió Pastor
Miguel Ángel Candel Mora

ACTAS DEL III CONGRESO INTERNACIONAL DE LINGÜÍSTICA DE CORPUS.

LAS TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES: PRESENTE Y FUTURO EN EL ANÁLISIS DE CORPUS

EDITORIAL

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Esta editorial es miembro de la UNE, lo que garantiza la difusión y comercialización de sus publicaciones a nivel nacional e internacional.

Primera edición, 2011

ISBN: 978-84-694-6225-6

Ref. editorial: 6032

ÍNDICE

PRÓLOGO

Este volumen refleja las ramificaciones que existen en la actualidad de los análisis de corpus, dentro de un entorno en el que se han intercambiado opiniones e ideas. Las contribuciones incluidas en estas actas del III Congreso Internacional de la Asociación Española de Lingüística de Corpus son una muestra del interés que está suscitando en la actualidad este tipo de estudios centrados en demostrar que la lengua se puede medir y el campo de las humanidades se puede equiparar con el de las ciencias.

Tras una concienzuda selección de los artículos presentados en este congreso, hemos respetado su división en los distintos paneles que lo componen y que están liderados por renombrados investigadores. Esperamos que los artículos, ideas, proyectos y novedades incluidos en este volumen sirvan como muestra del alto nivel de las presentaciones que tuvimos el placer de presenciar durante la celebración del congreso.

Nuestro deseo con esta publicación es que se divulgue el conocimiento relacionado con los estudios del corpus y que nazcan nuevas ideas al amparo de esta línea de investigación. Con ello deseamos que los autores se vean recompensados y que sirva como precedente para reuniones futuras de este campo.

María Luisa Carrió Pastor

Miguel Ángel Candel Mora

Editores

Diseño, elaboración y tipología de corpus

PEPCo: Designing a parallel and comparable translational corpus in Brazil

Lautenai Antonio Bartholamei Junior

Universidade Federal de Santa Catarina, Brazil

Abstract

PEPCo a tool developed to help researchers in the corpora exploration task. PEPCo has two main steps: (i) corpus design, i.e., text selection, representativeness; and (ii) development of tools. Tools provided by PEPCo are parallel concordances, monolingual concordances, word-lists, n-grams, and PEPCo Builder. The result (in progress) is a parallel corpus of about 3 million words and a comparable corpus of about 5 million words which could be useful for many researchers in translation studies in Brazil. Most researches using PEPCo are related to translation studies and translational phenomena emerging from a compiled corpus. Popular genres in PEPCo are Fantasy, Science-Fiction, Medical and Academic Texts. Corpus tools provide filters to user search for specific texts, genres, period, authors, translators, publishers. Also, users can specify to query only on source text, target text or both. PEPCo is used by students and teachers to researches and translator's training in Southern Brazil.

Keywords: corpus development, corpus-linguistics, corpus-based translation studies.

Resumo

PEPCo é projetado para auxiliar pesquisadores na exploração de corpora. O processo de desenho do PEPCo foi realizado em duas etapas: (i) o projeto corpus, ou seja, a seleção de texto, de representatividade, e (ii) desenvolvimento de ferramentas. Os recursos mais importantes no PEPCo são as concordâncias paralelas, concordâncias monolíngues, listas de palavras, n-grama, e a criação do corpus pelo usuário. O resultado é um corpus paralelo de cerca de 3 milhões de palavras e um corpus comparável de cerca de 5 milhões de palavras utilizado por muitos pesquisadores em estudos da tradução no Brasil. A maioria das pesquisas usando PEPCo estão relacionadas aos estudos de tradução e fenômenos de tradução emergentes por meio de um corpus compilado. Entre os gêneros populares no PEPCo estão: Fantasia, Ficção Científica e Textos Acadêmicos. O PEPCo oferece recursos para busca simples e avançada e é usado por alunos e professores para pesquisas e formação de tradutores no sul do Brasil.

Palavras-chave: desenvolvimento de corpus, linguística de corpus, estudos da tradução baseado em corpus.

1. Introduction

New research approaches and methodologies in translation studies have been developed in recent years. Based on the interdisciplinarity of the translation studies discipline, technological tools have been introduced to help researchers on carrying out researches. One of these new methodologies is the use of corpus tools to analyze language phenomenon. Corpus-based translation studies methodology was firstly introduced by Baker (1993, 1995) and nowadays is used as a valuable tool used as approach, as well a methodology. As the translation studies discipline is focused mainly in search for phenomena related to translate text, theses tools need to work with both original and translated text. With this is mind, this paper proposes to design a tool to process parallel and comparable corpora for a particular environment, Brazilian researchers in translations studies. Since work with corpora it is not an easy task, PEPCo is a tool which heavy relies on the usability for building and processing corpora.

2. Existing Corpora in Translation Studies and Linguistics

Before start designing a new tool to compile and process parallel and comparable corpora, we explore corpora and corpora tools available on the internet. In terms of corpora we can freely access via web browser we have COMPARA (Frankenberg-Garcia, 2001), OPUS – An Open Source Parallel Corpus (Tiedemann & Nygaard, 2004), NATools (Almeida & Simões, 2002) which are the most important and similar to what we need. Based on these available corpora, we have some appointments to do:

- they do not provide multiple translation of a source text.
- they do not allow user to create his/her own corpora.
- they do not provide both parallel and comparable corpora.
- they do not provide features like user history of queries.

Considering these initial features, existing corpora do not provide the tools require for most researchers in our context. Studies have been developed using more than one translation of a source text, and most users need to create their own corpora to carry out researches.

3. Initial Ideas

Design a new corpus tools based on the lack of features of other corpora was the initial idea we have in mind. Integrate existing tools and add required features were the main objective. Propose a corpus environment that enables users to use features presented in traditional corpora tools and add these others features could make possible to make use of new methodologies for corpus-based translation studies researches.

4. Main Features

For designing a new tool to work with corpora, we consider these main features:

- parallel corpora queries.
- comparable corpora queries.
- combined parallel and comparable queries.
- multiple translations of a source text.
- user own corpus creation.
- simple and advanced queries with history of queries.

All these features allow researchers to have more flexibility in working with corpora. One of the most important features for our researches is the use of more than only one translation of a source text mostly because they are working with translated and re-translated texts, then they need to have a tool that enables this option. For others researchers, the feature of create their own corpus could enable the possibility to work with new texts inserted in the corpus. Enabling users to insert their own text can also increase the collection of text and allows for more possibilities of researches.

5. Developing the Corpus Platform

In the stage of developing the corpus processor, the first step was the database design. For the database, we decide to use a relational database and the structured query language method. MySQL was the choice for the reason we can have a free open source database tool and can easily be integrated with the PHP scripting language. PHP was used as the scripting language which processes the interface and the interaction between the platform and the database. Second step was the development of a user-friendly interface when users interact with the post and get method in the corpora.

6. Database Design

In the process of developing the corpus platform, this was considered the most important task. Database design will define how the data was structured and how to retrieve it. For advanced queries, we used relational tables in which the data was organized for easy retrieval.

We used the following tables to store the corpora data:

- authors
- corpora
- dates
- publishers
- genres
- books
- countries

- genders
- sub corpora
- texts
- text additional
- translators
- users
- user history

By now we define each table in the database design:

- Authors: table authors stores all data related to all authors in the corpora. Main fields in this table are: auhor_id, author_name, country_id and gender_id.
- Corpora: table corpora stores all data related to all corpora in the system. Main fields in this table are: corpus_id, subcorpus_id, genre_id, book_id, author_id, and translator_id.
- Dates: table dates stores an index of dates which can be relational to original and translated date of publication and enables the search for specific period of time in the corpora. Main fields in this table are: date_id and date_name.
- Publishers: table publishers stores all data related to publishing houses. Main fields in this table are: publisher_id and publisher_name.
- Genres: table genres stores all data related to gender and enables to specific the gender of authors and translators. Main fields in this table are: genre_id and genre_name.
- Books: table books stores all data related to books include in the corpora and based on this table we can get information related to the header of the sub corpus. Main fields in this table are: book_id, book_name, book_date_of_publication.
- Countries: table countries stores an index of all countries and all data is related to countries in which authors, translators and publishers are located. Main fields in this table are: iso, iso3, numcode and country_name;
- Genders: table genders stores an index of genders and is related to authors and translators genders. Main fields in this table are: gender_id and gender_name.
- Subcorpora: table subcorpora stores data related to all subcopora in the corpora.
- Texts: table texts stores the translations unities, both original and translated texts. Main fields in this table are:
- Texts Additional: table texts additional store other translation of a source text. Using this table make possible to have more than only one translation for a source text. Main fields in this table are: text_id and text_translation.

- Translators: table translators stores all data related to translators. Main fields in this table are: translator_id and translator_name.
- Users: table users stored all data related to system users. Main fields in this table are: user_id and user_name.
- User History: table user history stores all data related to query history for every user. Main fields in this able are: user_id and user_query.

Using all theses tables in the database design is possible to have a complete management corpora system with all features require in most of our researches. Also, using a relational database is possible to add more tables and fields to support future features.

7. Interface Developments

For the interface development, we use technology most used in a web platform. For presenting data to user's browser, HTML5 was used integrate with CSS3 and Jquery Javascript Library for providing better usability.

In the front-end interface, we provide two main features, simple and advanced search. Simple search feature simply provide a search in the entire corpora and them send results to the browser. Advanced search provide a more detailed search in the corpus. Using advanced search interface, users can specify a list of constraints when querying the corpora. Combination of constraints could specify queries in this way:

- node word.
- node word in translation(s).
- node words starting with, any or exact query.
- multiple selections for genre.
- multiple selections for books.
- multiple selections for authors.
- multiple selections for translators.
- multiple selections for publishers.
- specify gender for authors and translators.
- specify authors and translators nationality.
- specify period of publication for the original and translated texts.

Using all of these presented possibilities for querying corpora, it could be useful for most users and researchers.

All queries performed by a user are stored in the system to provide the same query for future research and/or using in events such as conferences, congress and at classroom.

PEPCo uses a back-end interface for administration of the corpora. In the back-end interface we can provide an user-friendly interface to interact and manage data in the corpora. Back-end interface is password protected using encryption methods.

For the interface generation PHP scripting language was used.

8. PEPCo Data

PEPCo provide parallel corpora of about 2 million words categorized for simple an advanced search. In addition to it, comparable corpora of about 5 million words is also provided.

Most genres included in the corpora are fantasy and science fiction. For every completed work using PEPCo, all material is included in the official corpora. Before researches are completed, users have access to another feature provided by PEPCo, the PEPCo User Corpora.

9. PEPCo User Corpora

PEPCo User Corpora is a tool that enables users to add their own corpus and use all features presented in the system, such as concordances, wordlists, user history and export to file.

In order to create their own corpus, users need to have both original and translated text aligned by matching all lines for every fine. As we require a more qualitative than quantitative criterion, the files need to have an exact alignment. Also, they need to submit all extra-linguistic information about the corpus.

When all data are prepared, PEPCo provided an interface to submit parallel texts and a form to fill about extra-linguistic information. Extra-linguistics information is required to create a header file to each corpus.

Each user has an account and he/she can have access to his/her own corpora by accessing the system.

10. Results

Design a corpus system to support local researches and researchers show that is possible to achieve good results in when proposing new approaches to corpus-based methodologies.

Since PEPCo was available online, researchers have been using data extracted from queries to analyze and propose projects and researches. In addition, we have completed researches in both graduate and master level using data from PEPCo full corpora or User Own Corpora feature.

When working with PEPCo, a bunch of problems, bugs and new features requests was found. For problems and bugs, most of them ware fixed. For new features request, some of them ware included and others are coming soon.

11. Conclusions

For most researchers when start to working with corpora can find some systems available. But in most cases these systems do not provide all features require by that researcher or a group.

When proposing a new system we tried to look at existing systems available and extract most important features and then add the group required features to it. PEPCo provide the most used features and new important ones such as combine parallel and comparable corpora and also the possibility for users to create their own corpora.

12. Future Works

Future works on designing a new corpora system is the development of an interface for compiling and processing subtitles with support to video (Bartholamei Jr., 2011).

Working on the retrieval system to produce more complex queries and adding support to tagged corpora is the next level to achieve in PEPCo.

References

Almeida, José João; Simões, Alberto Manuel; Castro, José Alves. 2002. *Grabbing parallel corpora from the web*. Number 29, pages 13–20. Sociedad Española para el Procesamiento del Lenguaje Natural, Sep.

Baker, Mona (1993). *Corpus Linguistics and Translation Studies: Implications and Applications*, in Mona Baker, Gill Francis and Elena Tognini-Bonelli (eds) Text and Technology: In Honour of John Sinclair, Amsterdam & Philadelphia: John Benjamins, 233-250.

Baker, Mona (1995). *Corpora in Translation Studies. An Overview and Suggestions for Future Research*.Target 7(2): 223-43.

Bartholamei, Jr. (2011). *Construção de um Corpus de Legendas Paralelo Multilíngue com Suporte a Vídeo*. CCE/UFSC.

Frankenberg-Garcia, A. (2001) *COMPARA: the Portuguese-English Parallel Corpus*. In I Congresso Internacional de Estudos Anglo-Portugueses, Universidade Nova de Lisboa, 6-8 Maio 2001, pp 45-57.

Jörg Tiedemann, Lars Nygaard, 2004. *The OPUS corpus - parallel & free*. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal, May 26-28.

Nip & Tuck: A corpus-based qualitative typology for concision in scientific writing

Marta Conejero, Asunción Jaime, Debra Westall

Universidad Politécnica de Valencia

Abstract

One challenge facing researchers, especially those who use English as an Academic Language (EAL), is that of publishing in high-impact journals. Recently, acceptability has become extremely demanding as editors expect 'native-like' stylistic patterns to facilitate comprehension. In this research we analyzed a sample of 12 scientific articles (manuscripts and publication), written directly in English researchers at the Universidad Politécnica de Valencia *(UPV), thoroughly revised by one of the present authors, and available online in scientific and engineering journals. In this paper, we shall explain the unique features of this sample, the results of our analysis, and a typology based on concision-achieving in scientific writing.*

Keywords: *corpus analysis, typology, concision, academic English*

Resumen

Publicar en revistas de alto impacto en lengua inglesa es un reto permanente para los investigadores, sobre todo, para aquellos cuya lengua materna no es la inglesa. Actualmente las normas de aceptación requieren patrones de estilo que faciliten la comprensión. Para el presente estudio analizamos una muestra de 12 artículos científicos (manuscritos y publicaciones) escritos directamente en inglés por investigadores de la Universidad Politécnica de Valencia (UPV) y disponibles online en revistas científicas en lengua inglesa. Previa a su publicación, los manuscritos originales se sometieron a una cuidadosa revisión lingüística a cargo de nuestro equipo. En este trabajo explicamos las peculiaridades de la muestra, los resultados del análisis realizado y un modelo de concisión basado en la muestra.

Palabras Clave: *análisis de corpus, tipología, concisión, inglés académico*

1. Introduction

Corpus analysis is fast becoming the most essential and productive technique for theoretical and computational linguistics research. Over the past few years there has been a tremendous growth and interest in corpus building and analysis. Among the most popular corpora are the *Survey of English Usage* (1957), the well-known *Brown University Standard Corpus of Present-Day American English (1961);* the *Birmingham Collection of English Text (1980-85),* the *Lancaster-Oslo/Bergen Corpus* (LOB) (1970), the *British National Corpus (BNC)* (1994) or the more recent *International Corpus of English* (*ICE*) and *The Michigan corpus of academic spoken English (MICASE)*. As a general rule, these are bodies of texts assembled according to explicit design criteria for specific purposes, creating a rich variety of corpora that reflects the diversity of their designers' objectives (Atkins, Clear & Ostler, 1991). Krieger (2003) envisions Corpus Linguistics as the discovery of "patterns in authentic language use through the analysis of actual usage". Likewise, WordSmith creator Mike Scott recently affirmed that "Although it might seem that the main utility of Corpus Linguistics lies in the power software brings of ploughing through large amount of text in order to find examples that is in truth secondary. The main purpose is to identify textual or linguistic patternings" (Scott, 2011).

The present work is the result of our 'corpus-driven' research into human-based discourse analysis, as opposed to machine-based analysis, starting with a selection of scientific articles, written directly in English by researchers at the *Universidad Politécnica de Valencia* (UPV) and then fully revised by a native English speaker skilled in the edition of scientific papers (author 3) before being submitted for review and possible publication in relevant specialist journals. This research was inspired by the requests for linguistic consultation from these members of the UPV community who use English as an Academic/Additional Language (EAL) to disseminate their findings in high-impact journals publications, but do not always meet the required publishing standards, with their manuscripts being accepted only after fulfilling reviewers' or editors' demands that the texts be 'revised by a native speaker'. In this sense, Flowerdew (2008) offers insight into the process:

> […] one can well imagine cases where a manuscript produced by an EAL research team might be perfectly acceptable to them and to the team's peers (or that it is the best they are capable of, knowing it is still not written in what might be considered to be the appropriate form) but when it is submitted to a journal editor and reviewers, it becomes problematic because different standards are applied. (Flowerdew, 2008: 80)

Thus it is necessary for us to view the context in which the texts are written, why the texts are written as they are and authors' motivation, in Paltridge's (2008: 21) words: "what can be said and cannot within a particular genre as well as the goals, assumptions and values that are presupposed by expert readers", in particular "discourse community expectations, conventions and requirements for the text" (Paltridge, 2008: 5).

As reviewers and editors tend to target 'word choice and grammar' and 'the writing style' of our UPV colleagues, our analysis later revealed, among other aspects, that many changes

made to the original manuscripts entailed reduction strategies that led to greater concision and improved readability. The need for greater concision was also detected by Iscla and Benavent (2003: 63), whose study of titles in a medical journal revealed: "El defecto más frecuente (21%) ha sido la falta de concisión por el uso de palabras o expresiones que no aportan información". Therefore, we decided to compile a specific working corpus with some 75 of the original, revised and published papers, written directly in English by UPV researchers, faculty and engineers, and revised by the linguistic consultant over a five-year period (2001/2-2006/7).

Further examination of the revised manuscripts confirmed the repeated occurrence of wordy patterns which had been changed to shorten phrases and simplify expression. As Bem (2003: 3) pionted out: "The primary criteria for good scientific writing are accuracy and clarity. […] The first step towards clarity is good organization. […] The second step toward clarity is to write simply and directly". The examination shed light on these very specific and somewhat complex aspects of scientific discourse production, which have yet not been fully examined in the literature and which we consider to be a research topic worth pursuing. In this sense, Paltridge (2008: 18) refers to academic settings and knowledge of specific discourse communities 'beyond the text'. As linguistic researchers, it is our aim is to identify, classify, exemplify and describe what students and researchers need to know to succeed as well as facilitate the effective acquisition of skills, strengthen their written expression and provide training in particular writing tasks.

2. Corpus design and methodology

For this study, twelve manuscripts were selected from the initial 75-item set of original and published papers (Table 1).

Table 1. Articles selected for the sample and analyzed for concision patterns.

1. Modeling crop regional production using positive mathematical programming. *Mathematical and Computer Modelling,* 2002, 35(1), 77-86.

2. Optimization of Touristic Distribution Networks using Genetic Algorithms. *Statistics and Operations Research Transactions*, 2003, 27(1), 95-112.

3. Vegetable trade flows between the European Union and its Mediterranean partners. *Mediterranean Journal of Economics, Agriculture and Environment*), 2005, 4(2), 4-10.

4. Mechanical harvesting of processed apricots cv. Búlida in Spain. *Applied Engineering in Agriculture,* 2006, 22(4), 499-506.

5. EMG assessment of chewing behaviour for food evaluation: Influence of personality characteristics. *Food Quality and Preference*, 2007, 18(3), 585-595.

6. Effects of dietary soybean oil concentration on growth, nutrient utilization and muscle fatty acid composition of gilthead sea bream (*Sparus aurata L.*). *Aquaculture Research*, 2007, 38(1), 76-81.

7. A parametric study of optimum earth retaining walls by simulated annealing. *Engineering Structures,* 2008, 30(3), 821-830.

8. Charge optimisation study of a reversible water-to-water propane heat pump. : *International Journal of Refrigeration,* 2008, 31(4), 716-726.

9. Efficiency of repeated *in vivo* oocyte and embryo recovery after rhFSH treatment in rabbits *Reproduction in Domestic Animals,* 2008, 18 [Epub ahead of print]

10. Analytical prospect of compact disk technology in immunosensing. *Analytical and Bioanalytical Chemistry,* 2008, 391(8), 2837-2844.

11. Mechanical harvesting of processed peaches. *Applied Engineering in Agriculture,* 2008, 24(6): 723-729.

12. Review of standards for the use of hydrocarbon refrigerants in A/C, heat pump and refrigeration equipment. International Journal of Refrigeration, 2008, 31, 748-756.

All 12 articles were written directly in English by one of the co-authors, most of whom have an advanced level of English (C1-C2). With their years of professional experience, the authors are fully competent in terms of the specific and specialized vocabulary, and they have acquired a relatively-sufficient understanding of what 'sounds' correct, even though they may not know the grammatical or stylistic explanations for this. Despite the years of comprehensive reading in their respective fields of specialization, they often find it difficult to model native writing styles. However, they much prefer to write directly in English and avoid machine translations or professional translators which often produce unsatisfactory results or delay the manuscript preparation process. Although there are no guarantees that the initial draft of a manuscript will be acceptable for publication without further changes, the authors consider that the meticulous revision carried out by a linguistic consultant (author 3) to be crucial for enhancing the readability of their papers, and the suggestions made contribute to success in the overall publication process. These suggestions generally involved correcting grammar and word choice, awkward constructions, wordiness, vague references, semantic ambiguity and the like.

Moreover, as all 12 articles were published online in English language journals of relevance in the fields of science and engineering, we consider that this aspect certainly adds value to the research proposed herein. We were able to select papers which are easily accessible and cited in many of the top science journals, which is highly relevant given the qualitative nature of our selection. Naturally, the articles selected were those for which one or more of the original manuscripts were available for comparison and contrast with the corresponding published texts. We compared the first versions to second or final ones, which proved to be enlightening when extracting the changes (linguistic differences, readability improvements) and examining the overall evolution of articles until their publication.

The procedure for analyzing the manuscripts involved three steps. First, we coded what constituted each type of suggestion; then we categorized the corrections and classified the results into a working typology. To this end, the entire text of each manuscript was examined for the types of modifications suggested and it was then compared to the published article given that the researchers/authors were responsible for introducing any or all the changes they deemed necessary and/or appropriate. No attempt was made to identify every single modification in every set of articles as we quickly detected areas which were more subtle and complex than those identified in traditional 'error analysis' studies.

In addition to studying repeated modifications, our objective was to develop a typology based on qualitative rather than quantitative representative examples so it was necessary to classify the proposed modifications and to design a relevant typology of general readability and style problems, on the one hand, and of more particular errors (wordiness [cumbersome] and word choice [awkward]), on the other. It was observed that both general and particular problems were classifiable under one heading: "Concision-achieving patterns" or "nip-and-tuck" procedures.

3. Results

The sample analyzed contained 12 original manuscripts and the corresponding articles published in the fields of Applied Mathematics, Agricultural Economics, Food Technology, Thermodynamics, Agricultural Machinery, Animal Science, Crop Production, Biotechnology, Analytical Chemistry and Civil Engineering (see Table 1, above).

The primary result of this research so far is, as the title of this paper suggests, a "Nip & Tuck" corpus-based qualitative typology for concision in scientific writing. Despite the variety of concision types observed in the manuscripts analyzed, this qualitative typology reflects the most frequent instances identified. This typology may be useful for Spanish researchers when writing their manuscripts for publication in international scientific journals since it provides, what we consider, simple instructions along with authentic examples of these problematic areas of scientific expression. For illustrative purposes, here we offer three procedures to minimize wordiness and avoid awkwardness (*original phrasing* > revised phrasing) and, thus, improve the readability of the original manuscripts.

A) To *remove* unnecessary words

i) We can eliminate the subject and auxiliary verb when used with *as* or *if*:

- *as can be observed* > as observed
- *as we said before* > as stated
- *as it will be discussed* > as discussed
- *as can be seen in figure 6* > as figure 6 indicates

ii) We can shorten introductory and relative clauses:

- *In order to summarize the results* > Summarizing the results

- *With the aim of assuring that the possible detected differences…* > To assure that possible detected differences …
- *An optimum subcooling exists which is related to the temperature* > An optimum subcooling is related to the temperature

iii) We can rewrite noun phrases using a verb or changing the order of nominal group components:

- *the analysis of the refrigerant* > analysing the refrigerant
- *the rest of the refrigerant* > the remaining refrigerant
- *the performance of discs* > disc performance
- *the layout of the system* > the system layout

B) To *reduce* redundant or superfluous ideas, we can focus attention on the main idea of the phrase or sentence:

- *This paper presents a theoretical study performed with a mathematical model* …> The theoretical study is discussed with a mathematical model…
- *As regards calculations of the wall as a structure, they are performed …* > The calculations of the wall as a structure are performed ….
- *[…] muscle fatty acid composition of the fish was highly dependent of the experimental diet given* > muscle fatty acid composition differed with the fish diet
- *Nevertheless, oocyte recovery is attractive since it is essential for some reproductive technologies* > Nevertheless, oocyte recovery is essential for certain reproductive technologies

C) To *rework* the complex phrasing, especially with statements of purpose:

- *The purpose of this research was to design, construct and test catching systems for picking the peaches detached from the trees by shaking* > The purpose of this research was to evaluate three catching systems to collect peaches detached from trees.

4. Conclusions

In this study we described a sample created with 12 scientific articles (manuscripts and publication), written directly in English by UPV researchers, thoroughly revised by a linguistic consultant and now available online in relevant specialized journals. The unique features of this sample and the results of our analysis led to the development of a corpus-based qualitative typology for concision in scientific writing. The three main procedures (remove, reduce and rework) are described herein and authentic examples of these problematic areas of scientific expression are illustrated.

This typology may be useful for Spanish authors when faced with the challenges of preparing their research for publication in international scientific journals. Future research will certainly require the continued belief in the benefits of interdisciplinary collaboration, especially as the UPV community becomes fully integrated in the European Higher Education Area. Work is currently underway to broaden the typology to include more concision-achieving strategies as well as other complex EAL constructions or patterns like those typically featured in written scientific and engineering discourse. Finally, it is our hope that this research can be transformed into self-study materials available online for engineers and researchers, teachers and students, translators as well as EAL material designers and evaluators.

5. Acknowledgements

The authors are most grateful to the UPV researchers who have graciously shared their writing experiences with us for this study.

6. References

Alcaraz Varó, E. (2000). *El Inglés Profesional y Académico*. Madrid: Alianza Editorial.

Atkins, S., Clear, J. & Ostler, N. (1992). Corpus Design Criteria. *Literary & Linguistic Computing,* 7(1), 1-16.

Bem, D.J. (2003). Writing the Empirical Journal Article. In J.M. Zanna & H. L. Roediger (Eds.), *The Complete Academic: A Practical Guide for the Beginning Social Scientist.* Washington: American Psychological Association.

Biber, D., Conrad, S. & Reppen, R. (1998). *Corpus linguistics: investigating language structure and use*. Cambridge: CUP.

Biber, D. & Conrad, S. (2001). Corpus based research in TESOL. *TESOL Quarterly,* 35(2), 331-335.

Conejero, M. & Westall, D. (2009). European Convergence, Life Sciences and Writing Research. In P E. Caridad de Otto y A.F.López de Vergara Méndez (Eds.), *Las lenguas para fines específicos ante el reto de la convergencia Europea.*. Tenerife: Servicio de Publicaciones ULL (Colección: Documentos congresuales/24). (pp. 200-215).

Conejero, M., Jaime, A. & Westall, D. (in press, 2011). Tools for strengthening scientific writing: a new approach to autonomous learning. In A. Gimeno (Ed.). *ReCall (Proceedings of EUROCALL2009: New trends in CALL: working together*. Macmillan ELT (pp.130-145).

Flowerdew, J. (2008). Scholarly writers who use English as an Additional Language: what can Goffman's "Stigma" tell us? *Journal of English for Academic Purposes* 7(2), 77–86.

Francis, W. N. & Kucera, H. (1964/1979). *Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English, for use with Digital Computers.* Department of Linguistics, Brown University.

HUNSTON, S. (2002). *Corpora in Applied Linguistics.* Cambridge: CUP.

ISCLA, A. & ALEIXANDRE BENAVENT, R. (2003). Defectos en el título de los artículos publicados en las revistas Piel. *Piel* 18, 63-69. Retrieved from http://www.doyma.es/revistas/ctl_servlet?_f=7012&articuloid=13043725&revistaid=21

JOHNSON, B. (2005). Concision: the art of linguistic liposuction. *Science Editor* 28(4),134-135. Retrieved from http://www.stcsig.org/sc/newsletter/html/2010-1.htm#item1

JOHANSSON, S., LEECH, G. N. & GOODLUCK, H. (1978). *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers*. Department of English, University of Oslo.

KRIEGER, D. (2003). Corpus Linguistics: What It Is and How It Can Be Applied to Teaching. *The Internet TESL Journal*, 9(3). Retrieved from http://iteslj.org/Articles/Krieger-Corpus.html

PALTRIDGE, B. (2008). Textographies and the researching and teaching of writing. *Iberica* 15, 9-24.

SCOTT, M. (2011). Investigating patterns. Paper presented at the *Congreso Internacional de Lingüística de Corpus CILC 2011.* Valencia, Spain, 7-9 April 2011. Abstract available at http://www.upv.es/upl/U0547372.pdf

Turigal: compilation of a parallel corpus for bilingual terminology extraction

Adonay Custódia Dos Santos Moreira

School of Technology and Management

Polytechnic Institute of Leiria

Turigal, a parallel corpus of tourism advertising material, has been devised to support the creation of a bilingual term bank on tourism. The corpus consists of texts – printed brochures, guidebooks and websites – in Portuguese and their translations into English, all of which were sourced from Portuguese Tourism Regions, Regional Tourism Boards and Regional Tourism Promotion Agencies, and stored as plain text. For the moment, it contains 1,285,764 words and is included in the Linguistic Corpus of the University of Vigo (CLUVI). This paper describes the methodology used in the compilation of Turigal. First, we examine the process of text collection and storage. Then, we discuss Pearson's (1998) set of criteria for corpus design and text selection which has been considered when compiling our corpus. Finally, we present the alignment and tagging of Turigal.

Keywords: parallel corpus, corpus design, alignment, tagging.

1. Introduction

These last few years have witnessed an increase in research involving the compilation of large quantities of texts and their respective translations, as well as the development of techniques for processing those bilingual term banks (Bowker & Pearson, 2002; Biber, Conrad & Reppen, 2004; McEnery & Wilson, 2004). The present study is an example of such research as it uses a Portuguese-English unidirectional parallel corpus as a starting point for the retrieval of terminology. The main goal of our research is to exploit one of the possibilities offered by parallel corpora: the compilation of bilingual term banks. *Turigal*, a parallel corpus of tourism advertising material, has been devised to support the creation of a bilingual term bank on tourism. Our comprehensive term bank is comprised of pragmatic (context of use, relative frequency of terms), linguistic (gender, number, grammatical category, lemmas, synonyms) and conceptual information (thematic tree, semantic relations). It can eventually be useful to assist translators working in this industry, tourism professionals who work in an increasingly multilingual society and would gain from access to a 'ready-made' bilingual list of terms, and tourism trade businesses that market products and services internationally through the use of printed or electronic multilingual texts.

First, we look at the collection and storage of texts. Then, we discuss Pearson's (1998) set of criteria for corpus design and text selection – namely size, constitution, publication, author, factuality, technicality, audience, intended outcome, setting and topic – which has been considered when compiling our corpus. And lastly, we present the alignment and tagging of *Turigal*.

2. Compilation of a parallel corpus for bilingual terminology extraction

2.1. Text collection and storage

The corpus on which the term bank is based consists of texts (printed brochures, guidebooks and websites) in Portuguese and their translations into English, all of which were sourced from Portuguese Tourism Regions, Regional Tourism Boards and Regional Tourism Promotion Agencies, and stored as plain text. For the moment, it contains 1,285,764 words (469,873 words in the leaflets and 815,891 words in the webpages; 632,193 words in Portuguese and 653,571 in English) and it is included in the *Linguistic Corpus of the University of Vigo* (Gómez Guinonart, 2003) and available for free consultation at http://sli.uvigo.es/CLUVI. Since our terminological approach is based on corpus – specifically a parallel corpus, where the meaning of terms arises from their context of use – and this corpus is determined by the purpose for which it will be used, we have named it "special purpose parallel corpus". This expression is adapted from Pearson (1998: 48)'s term – "special purpose corpus" – and it designates a parallel corpus built for specific purposes. Thus, within the present research, a "special purpose parallel corpus" consists of a corpus of original texts and their translations, which is used for terminological purposes.[1]

1 Within this research, a corpus is a set of texts of a given field, which have been written and used by specific groups of people and selected according to a specific purpose. In this case, the purpose is the extraction of bilingual terminology.

Due to the texts' format – brochures/guidebooks and hypertexts – the apparently simple task of storing them as plain text turned out to be time-consuming. On the one hand, the printed brochures and guidebooks had different formatting types (size, font, text layout and page configuration), different colours and quite often texts in Portuguese, English and other languages were kept side by side on the same page. All brochures and guidebooks had to be scanned. On the other hand, working with webpages, though more productive in terms of the quantity of texts obtained, also required substantial post-processing, since many webpages had formatting codes which prevented easy access. Webpages had multiple input formats and in the process of text conversion some chunks of text would sometimes disappear and hence had to be manually typed. Moreover, texts which were not translated or were only in English had to be disregarded. Finally, some newly implemented sites were extremely slow and one could only open a page at a time, which slowed down the storage process.

Both types of texts – printed texts and hypertexts – were then submitted to an Optical Character Recognition (OCR) programme and then to a spelling correction programme, in order to check the text generated by the OCR. The texts which remained practically illegible after the OCR was applied to them were manually typed.

All graphs, addresses and pictures have been removed as well as some information considered irrelevant for our terminographical purposes, such as proper names (people and companies) and addresses. As far as hypertexts are concerned, these have been saved sequentially, according to the "site map", whenever this was available. To allow for the alignment of texts – the process of matching each phrase in Portuguese to its English translation – all texts which have originally been saved individually were joined together in a single text, creating a larger text or "super text".

A total of 3,484 Portuguese and English webpages from 17 websites were stored as plain text and were subsequently aligned. It should be noted, however, that the number of stored webpages was substantially higher. Despite having similar titles, many webpages in Portuguese could not be matched to their English version, due to a lack of correspondence in content. Those webpages just in one language had to be disregarded as well.

With regard to printed promotional texts, a total of 110 brochures/guidebooks were stored. Table 1 shows the Internet addresses (URL) and the number of brochures/guidebooks gathered from Tourism Regions.

Table 1: List of websites and brochures/guidebooks from Tourism Regions.

PORTUGUESE TOURISM REGIONS	**URL USED IN *TURIGAL***	**NUMBER OF BROCHURES/ GUIDEBOOKS USED IN *TURIGAL***
Algarve	—	3
Alto Minho	http://www.rtam.pt	13
Alto Tâmega e Barroso	http://www.rt-atb.pt	2
Centro	http://www.turismo-centro.pt	22
Dão Lafões	http://www.rtdaolafoes.com	4
Douro Sul	—	—
Évora	http://www.rtevora.pt	—
Leiria / Fátima	http://www.rt-leiriafatima.pt	2
Nordeste Transmontano	—	2
Oeste	http://www.rt-oeste.pt	—
Planície Dourada	http://www.rt-planiciedourada.pt	2
Ribatejo	—	7
Rota da Luz	http://www.rotadaluz.pt	11
S. Mamede	http://www.rtsm.pt	1
Serra da Estrela	—	2
Serra do Marão	http://www.rtsmarao.pt	—
Setúbal / Costa Azul	—	15
Templários	http://www.rttemplarios.pt	6
Verde Minho	—	3

Table 1 only lists bilingual websites whose texts have been collected. At the time this research was undertaken there were 19 Tourism Regions, some of which – *Douro Sul*, *Ribatejo*, *Serra da Estrela*, *Setúbal/Costa Azul* e *Verde Minho* – did not yet have English websites. Sometimes, only the titles of the webpages were in English.

Turigal also comprises texts sourced from the Azores and Madeira Regional Tourism Boards, and the following Regional Tourism Promotion Agencies: ADETURN, ARTA, ATA Azores, ATA Algarve and ATL. Table 2 indicates the URL and the number of brochures/guidebooks collected from these organizations.

Table 2 – List of websites and brochures/guidebooks from Regional Tourism Boards and Regional Tourism Promotion Agencies.

REGIONAL TOURISM BOARDS AND REGIONAL TOURISM PROMOTION AGENCIES	**URL USED IN *TURIGAL***	**NUMBER OF BROCHURES / GUIDEBOOKS USED IN *TURIGAL***
ADETURN	http://www.visitportoenorte.com	2
ARTA	—	1
ATA Azores / Azores Regional Tourism Board	http://www.visitazores.org	7
ATA Algarve	http://www.visitalgarve.pt/	—
ATL	http://www.visitlisboa.com	—
Madeira Regional Tourism Board	http://www.madeiraislands.travel/pls/madeira/wsmwhom0.home	5

Most bilingual brochures/guidebooks displayed in Tables 1 and 2 were obtained in the aforementioned organizations or received by mail, after being requested by telephone or e-mail. Others were fetched from a Tourism Fair held in Lisbon in January 2007.

2.2. Corpus design and text selection

We have used Pearson's (1998: 58-62) set of criteria for corpus design and text selection – namely size, constitution, publication, author, factuality, technicality, audience, intended outcome, setting and topic – to outline our special purpose corpus.

At first glance, a 1,285,764 word-corpus is small; however, one should take into consideration its time-consuming conversion to electronic form and the shortage of bilingual promotional texts from official organizations. Like Pearson (1998: 57), we believe a special purpose corpus does not have to be as big as a general purpose corpus. *Turigal* is considered to be sufficiently representative of all bilingual (Portuguese-English) promotional materials published and distributed by the official organizations responsible for the internal and external tourism promotion of Portugal in 2007, the year the texts were collected. The fact that it is a corpus difficult to compile can make it even more interesting to be studied, since there will certainly be fewer people interested in spending time in its collection.

The corpus contains complete written informative/promotional texts, of different size, in Portuguese and their translations into English. These freely available texts are clearly consumer-oriented, since their purpose is to transmit information to potential buyers in

order to persuade them to buy or consume products or services. They come in multiple formats – brochures, guidebooks and websites – and they are all full texts (not extracts) from written sources. Most brochures and guidebooks have no publishing date, but the ones which do are mostly from 2005 and 2006.

All texts have been published by official tourism bodies. This validates texts as a potential source of terminology in the area of tourism.

The case of authorship is particularly complex in our corpus, since most websites, guidebooks and brochures do not mention the authors and translators of texts. Websites frequently give the name of the company responsible for creating the websites, but do not indicate the name of people responsible for creating their texts. However, since texts are published by official tourism bodies, one assumes their authors are experts in tourism or someone with technical qualifications in the area.

Pearson also considers the criterion of "factuality" (1998: 61). According to the author, texts must be factual or should represent what is known or believed to exist. In our study, this criterion is particularly ambiguous with respect to promotional texts: on the one hand, we can consider them factual, since they display a specific tourism product that can be purchased by consumers; on the other hand, the language that is used is not in any way factual. Dann (1996) and Buck (1977) identified some features of this language of tourism. According to Dann, the language of tourism only speaks in a positive manner of the services and attractions that it is promoting (1996: 65). Thus, it is a hyperbolic language, full of clichés, which is used to capture tourists' attention at all costs. Referring specifically to the language of tourism brochures, Buck remarks that these are naturally fraudulent, as they send preconceived messages that affect tourists' expectations and perceptions. Hence, we have considered that the criterion of "factuality" indicated by Pearson has no relevance to our work.

As for "technicality", Pearson makes a distinction between technical (written by specialists for specialists) and semi-technical texts (written by specialists for a specific target audience). Our texts fall into this second category. However, the audience of our texts is not the student or professional working in the discipline, but the average citizen with a lower level of expertise in the area.

The intended outcome of our texts is informative, but mostly promotional, and the setting corresponds to communication between relative experts and the uninitiated. In her work, Person rejected texts which fitted this communicative setting, on the grounds that they were not likely to contain a high density of terms.

Regarding the last criterion – topic – all our promotional texts belong to the area of tourism.

It should be noted that the criteria selected by Pearson to classify her corpus were not considered in the same way for our project. Such criteria should suit the objectives of each research project and our objectives were distinct from hers. Pearson wanted to select specialized texts likely to contain metalinguistic statements which could be used to formulate definitions (1998: 62). Her aim was to provide specific groups of researchers

with useful definitions of terms. Ours is to create a descriptive terminological resource designed to meet the needs of a specific group of users – translators.

2.3. Text alignment and tagging

Texts were aligned with the program *TRANS Suite 2000 Align* (Cypresoft, 2000) and the format chosen for storing the aligned parallel texts is an adaptation of the TMX format (Translation Memory eXchange), as this is the XML encoding standard for translation memories and parallel corpora (Savourel, 2005). Each text has a header with information about text type (brochure, guidebook or website), its title in Portuguese and English, author, translator, publisher, year, URL, and date of access to the website and to the brochure, whenever the latter had no indication of publishing date.

As for the alignment itself, although a source sentence usually corresponds to a sentence in the translation, on some occasions one source sentence corresponds to two or more translated sentences or vice-versa, i.e., two or more source sentences correspond to one sentence in the translation. The alignment always starts with the source sentence, which means that the translation sentences were split or joined together to match the source sentence. Thus, aligning a parallel corpus also entails its manual tagging, since translating is not a linear task. Translators can omit words, phrases or sentences from the source text, insert new ones as well as reorder segments or whole sentences in the translation. Here are some examples of the tagged *Turigal* parallel corpus.

This is an example of an omission in the translation, which means that a source segment or sentence has no correspondence in the translated text. The omitted segment is placed in bold, between the tags "<hi type="supr">" and "</hi>".

Table 3: Example of an omission in corpus Turigal.

<tu>

<tuv lang="PT-PT">

<seg>Se gosta de desportos radicais, nada como fazer uma descida no Rio Minho (Rafting), tendo já em Melgaço Associações que preparam tudo (profissionalmente) [[hi type="supr"]] **para que a descida seja um êxito** [[/hi]]. </seg>

</tuv>

<tuv lang="EN-GB">

<seg>If you are a radical Sports lover, try to the descending of the river Minho (rafting) in Melgaço, there you can also contact a Professional Association to organise all . </seg>

</tuv>

</tu>

This second example is an addition, in which the translator decides to add segments that do not exist in the source text. The added segments are placed in bold, between the tags "<hi type="incl">" and "</hi>".

Table 4: Example of an addition in corpus Turigal.

<tu>

<tuv lang="PT-PT">

<seg>São as cumeadas da serra do Gerês, as Terras de Bouro, as praias de riba Minho, as Terras Soajeiras, os contrafortes da Senhora da Peneda e da Senhora do Sameiro, Barcelos e as margens ridentes do Cávado.</seg>

</tuv>

<tuv lang="EN-GB">

<seg>[[hi type="incl"]] **Minho is there for you to discover it:** [[/hi]] the peaks of the Serra do Gerês [[hi type="incl"]] **(mountains)** [[/hi]], the municipality of Terras de Bouro, the beaches of Riba Minho, the territory around the Serra do Soajo, the spurs of Senhora [[hi type="incl"]] **(Lady)** [[/hi]] da Peneda and Senhora do Sameiro, Barcelos and the luxuriant banks of the Cávado river. </seg>

</tuv>

</tu>

Finally, a reordering, i.e., changing the position of segments in the translation, compared to their position in the source text. Since the alignment is always from source text to translation, the reordered segment always has to match the source sentence. The reordered segment is placed in bold, between the tags "<hi type="reord" x="1" >" and "</hi>". The tag "<ph x="1"/ >" indicates the original position of the reordered segment.

Table 5: Example of a reordering in corpus Turigal.

<tu>

<tuv lang="PT-PT">

<seg>- Azulejos da nave, historiados, Barrocos e monocromáticos, de fabrico Lisboeta, alusivos a Santa Cruz e à vida de Santo Agostinho - Púlpito da autoria de Nicolau de Chanterenne, sendo considerado uma obra prima do Renascimento. </seg>

</tuv>

<tuv lang="EN-GB">

<seg>- The nave with historiated and monochromatic baroque tiles made in Lisbon and representing Santa Cruz and Saint Augustin's life. - A Pulpit, made by Nicolau de Chanterenne, [[hi type="reord" x="1"]] **and considered a masterpiece of the Renaissance;** [[/hi]] </seg>

</tuv>

</tu>

<tu>

<tuv lang="PT-PT">

<seg>Data de 1521.</seg>

</tuv>

<tuv lang="EN-GB">

<seg>dating back to 1521 [[ph x="1"/]] </seg>

</tuv>

</tu>

Encoding omissions, additions and reorderings in the corpus allows the automatic search of these translation strategies and facilitates their analysis. However, the purpose of the present research is not the study of translation strategies, but the use of an aligned parallel corpus for term extraction.

3. CONCLUSIONS

The main objective of this paper was to provide insight into the compilation of parallel corpus *Turigal*. This corpus was built for terminological purposes: to enable us to retrieve terms, get examples of use and translation equivalents, and distinguish multiple meanings

of terms. Our ultimate goal was the creation of a descriptive terminological resource which is available for free consultation at <http://sli.uvigo.es/termoteca/> (Gómez Clemente & Gómez Guinovart, 2006).

References

Biber, D., Conrad, S., & Reppen, R. (2004). *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bowker, L., & Pearson, J. (2002). *Working with Specialized Language: A Practical Guide to Using Corpora*. London and New York: Routledge.

Buck, R. (1977). The ubiquitous tourist brochure: explorations in its intended and unintended use. *Annals of Tourism Research*, 4, 195-207.

Cypresoft (2000). *TRANS Suite 2000 Align*. Belgium.

Dann, G. (1996). *The Language of Tourism: a Sociolinguistic Perspective*. Oxon: Cab International.

Gómez Clemente, X. & Gómez Guinovart, X. (dir.) (2006-). *Termoteca - Banco de Datos Terminolóxico da Universidade de Vigo*. Vigo: Universidade de Vigo. <http://sli.uvigo.es/termoteca/>

Gómez Guinovart, X. (dir.) (2003-). *Corpus CLUVI - Corpus Lingüístico da Universidade de Vigo*. Vigo: Universidade de Vigo. <http://sli.uvigo.es/CLUVI/>

McEnery, T. & Wilson, A. (2004). *Corpus Linguistics: An Introduction*. 2nd ed.. Edinburgh: Edinburgh University Press.

Pearson, J. (1998). *Terms in Context*. Amsterdam – Philadelphia: John Benjamins Publishing Company.

Savourel, Y. (2005). *TMX 1.4b Specification*. Localisation Industry Standards Association. Retrieved from <http://www.lisa.org/standards/tmx/specification.html>.

Criterios específicos para la elaboración y diseño de los corpus especializados en Terminografía

Isabel Durán Muñoz

Universidad de Málaga

Resumen

La especificidad de la Terminografía basada en corpus (Meyer y Mackintosh, 1996: 258), en contraposición a la Lexicografía basada en corpus u otras aplicaciones de los corpus (traducción, enseñanza de segundas lenguas, etc.), obliga al establecimiento de una serie de requisitos o criterios específicos para el trabajo terminográfico. Algunos de ellos serán comunes a los criterios generales de la compilación y diseño de los corpus y otros, como veremos, presentarán algunas diferencias. En este trabajo, pretendemos ilustrar estas diferencias y determinar las características propias que debe presentar un corpus compilado con un objetivo terminográfico, con objeto de mejorar los resultados de cualquier trabajo terminográfico.

Palabras clave: terminografía, corpus especializado, terminografía basada en corpus, representatividad.

Abstract

Specifity in corpus-based Terminology (Meyer y Mackintosh, 1996: 258), in comparison to corpus-based Lexicography and other corpus-based studies (on translation, second-language acquisition, etc.), requires the establishment of a series of specific criteria to carry out a terminology work. Some of these criteria coincide with the criteria established to design and compile corpora in general but other are different and need to be taken into account. In this paper, we pretend to illustrate these differences and determine the own features that a corpus compiled in the framework of a terminology work should present, with the aim to obtain a corpus adequate to our terminological necessities.

Keywords: terminography, specialised corpus, corpus-based terminography, representativity.

1. La Terminografía basada en corpus

La Lingüística de Corpus ha sido una de las disciplinas lingüísticas que más ha influido en la Terminología y, por ende, en la *Terminografía*. Tanto es así que hoy en día la idea de la contextualización de los términos a través de los corpus textuales está totalmente aceptada en la comunidad de terminógrafos, y los corpus[2] son considerados recursos indispensables para cualquier trabajo de esta naturaleza. De forma general, podemos afirmar que en *Terminografía* el uso de los corpus textuales se ve motivado por dos motivos fundamentales:

Por un lado, el trabajo terminográfico no consiste en la invención de denominaciones para unos conceptos previamente establecidos como propugnaban la terminología tradicional, sino en «la identificación y recopilación de los términos que los especialistas utilizan en realidad» (Cabré Castellví, 1993: 113). Por este motivo, si un terminógrafo debe estudiar los términos que los especialistas utilizan en su trabajo diario, deberá consultar directamente con los especialistas del campo de especialidad en cuestión o realizar un estudio detallado de las producciones lingüísticas que estos especialistas crean para comunicarse entre ellos o con otros actores. De estas dos opciones, la primera no es siempre posible, ya que puede resultar complicado disponer de los especialistas adecuados y, cuando se puede acceder a ellos, a menudo encuentran dificultades a la hora de explicar el significado y el uso del lenguaje que emplean, al fin y al cabo, de forma intuitiva. En palabras de Meyer y Mackintosh (1996: 264):

> While experts obviously *know* their domains, they do not all *explain* their knowledge *clearly* (whether orally or in writing), *completely* (it is up to the knowledge acquirer to make sure that all important areas in the field are covered), or *consistently* (experts often disagree with each other or change their minds).

Por ello, la segunda opción, la consulta de documentación especializada en forma de corpus textual, es más accesible, rápida y directa y, por tanto, determinante en el trabajo terminológico.

Por otro lado, se hace imprescindible el uso del corpus en *Terminografía* en la dimensión conceptual. Para poder identificar y recopilar los términos que los especialistas emplean en la realidad, que se incluirán en el recurso terminológico, los terminógrafos necesitan estudiar las estructuras de conocimiento (conceptos y sus relaciones) y familiarizarse con el tema específico de su trabajo, lo que Cabré Castellví (1999: 144) denomina «competencia cognitiva». De la misma forma que en el caso anterior, los terminógrafos pueden dirigirse a especialistas en el ámbito de especialidad en cuestión o consultar documentación especializada y, de nuevo, en la mayoría de las ocasiones, será más fácil familiarizarse con el ámbito de especialidad, con sus conceptos y estructuras a través de la documentación especializada y consultar puntualmente a los especialistas.

2 En *terminografía*, el tipo de corpus utilizado se denomina *corpus especializado*, al tratarse de un tipo de corpus especial que ha sido diseñado con un propósito específico y que tiene la finalidad de ser representativo de un tipo particular de lengua, como por ejemplo un campo de especialidad o un grupo particular de hablantes.

La necesidad de la utilización de documentación especializada en el trabajo terminográfico sistemático queda patente con estas dos razones expuestas. Y, gracias a los avances que se han producido principalmente en el ámbito de la informática que han permitido tener a disposición gran cantidad de información en formato electrónico y poder procesarla de forma automática o semiautomática, el uso de los corpus electrónicos se ha extendido en la actualidad, convirtiéndose en la herramienta esencial para la mayoría de los trabajos terminográficos y dando lugar a lo que Leech (1992: 106) considera «a new way of thinking about language».

2. El corpus en las fases del trabajo terminográfico

Como hemos visto en el apartado anterior, el corpus se ha convertido en una herramienta esencial para el trabajo terminográfico, puesto que hace posible la consulta y el procesamiento de grandes cantidades de información en un tiempo muy reducido y, a menudo, de forma automática o semiautomática. Sin embargo, las ventajas que aporta el corpus en *terminografía* no se limita a la consulta de información en la fase de elaboración de la terminología, sino que se trata de una herramienta que acompaña al terminógrafo en todas las fases de su tarea, es decir, se utiliza desde la fase inicial de preparación del proyecto hasta la fase final de validación. Así pues, la documentación, y en consecuencia, los corpus y las herramientas de análisis de corpus permiten llevar a cabo:

La adquisición de conocimiento conceptual y la familiarización del terminógrafo no experto del dominio mediante la consulta de documentos, a fin de identificar la estructura interna del dominio, las relaciones con otros campos de especialidad y las fuentes de conocimientos adecuadas.

La identificación de unidades terminológicas dentro de un discurso de especialidad, es decir, la nomenclatura.

El análisis y la preparación de entradas mediante la adquisición de información lingüística, pragmática y semántica extraídas del corpus, como por ejemplo definiciones, contextos o colocaciones.

La detección y descripción de nuevos conceptos, así como la identificación de etiquetas léxicas que se están atribuyendo a dichos conceptos dentro del dominio. En algunos casos, también la propuesta de un neologismo adecuado (cuando todavía no se ha acuñado un término en una lengua).

La estandarización de términos sinónimos o cuasisinónimos que son utilizados por diferentes expertos.

Y la clarificación de dudas y preguntas sobre inconsistencias y otros asuntos que, de lo contrario, solo se podría llevar a cabo mediante consultas a expertos del dominio.

Llegados a este punto, podemos resumir los roles de la documentación en dos aspectos: en primer lugar, para el análisis lingüístico (extracción de términos, elaboración de definiciones, etc.), y, en segundo lugar, para la adquisición de conocimiento (conceptualización del dominio, relaciones semánticas, etc.). Para poder extraer adecuadamente la información lingüística, pragmática y conceptual de los corpus, será necesario que estos estén compilados de forma apropiada y según unos criterios generales y específicos establecidos, teniendo en cuenta que «the corpus needs to be as linguistically and conceptually rich as possible» (Meyer y Mackintosh, 1996: 266).

3. Criterios para la compilación de corpus especializados

En *Terminografía*, a pesar de las ventajas reconocidas que aporta el uso de corpus electrónicos en el estudio del lenguaje en uso, esta herramienta se ha incorporado muy recientemente (Meyer y Mackintosh, 1996: 257). Asimismo, la *Terminografía basada en corpus* sigue utilizando la metodología y herramientas de la Lexicografía basada en corpus, lo que no siempre es recomendable como nos indican las autoras (ibíd.: 258):

> Hence, it is essentials that terminographers begin to specify the types of corpora they need and the corpus-analysis tools that best suit their task. […] in broad terms, the *specificity* of corpus terminography. We shall argue that terminography and lexicography are in many ways substantially different, and consequently, that the tools and techniques developed in corpus lexicography cannot be applied intact to terminography. On the contrary, some can actually be *dangerous*.

En este contexto, es esencial que, para desarrollar una *terminografía basada en corpus* que mejore la calidad de la investigación terminológica, se concierten un conjunto de técnicas y herramientas propias que permitan a esta disciplina realizar sus estudios y trabajos sobre la Terminología de los lenguajes de especialidad y, así, abandonar las herramientas desarrolladas propiamente para la Lexicografía. Uno de los aspectos que debemos tener en cuenta en este sentido son los criterios de compilación de los corpus, de los cuales algunos coincidirán con los seguidos en lexicografía pero otros serán específicos de la *terminografía basada en corpus*.

3.1. Criterios generales

Dentro de los criterios generales de compilación, determinamos cuatro: cantidad, calidad, documentación y simplicidad. Estos criterios son comunes a cualquier rama lingüística que utilice corpus textuales, aunque presentarán unas características concretas en cada una. A continuación, indicamos sus características en el contexto terminográfico:

El criterio de la cantidad es un aspecto polémico para la compilación de *corpus especializado*s. Para algunos autores (Meyer y Mackintosh, 1996: 268), un *corpus especializado* puede ser mucho más pequeño que uno de propósito general, pero para

otros no sería necesario poner un límite a la cantidad de texto que se recopile (Pearson, 1998: 57), siempre y cuando siga unos criterios establecidos previamente. En definitiva, no existe un acuerdo acerca del volumen que debe tener un *corpus especializado* para que sea representativo. Desde la lexicografía basada en corpus se recomienda que el volumen del corpus sea muy elevado, pero en *terminografía* resulta más importante la densidad terminológica que el volumen propio del corpus utilizado, por lo que el número de palabras no significa que sea más o menos representativo. En este sentido, no es posible establecer a priori el número de textos o de palabras que necesita un corpus para ser representativo, aunque sí es posible medir su *representatividad* de forma objetiva y cualitativa posteriormente mediante el uso de la aplicación informática *ReCor*, que permite estimar la *representatividad* de los corpus en función de su densidad terminológica una vez compilados que presenta un corpus electrónico (cf. Seghiri Domínguez, 2006; Corpas Pastor y Seghiri Domínguez, 2007*a*, 2007*b*).

Con respecto al criterio de calidad, los textos deberán cumplir una serie de requisitos: primero, los textos deben ser recientes y actuales, es decir, se debe compilar un corpus sincrónico formado por textos que representen el estado actual del dominio, a fin de proporcionar los términos utilizados en el campo de especialidad en un momento concreto y la información conceptual actualizada; segundo, los textos deberán estar escritos por diferentes autores expertos (ya sean especialistas, instituciones, organizaciones, etc.) en la medida de lo posible para ofrecer fidelidad en el contenido y neutralizar cualquier tipo de idiosincrasia del autor; tercero, los textos deben ser preferentemente originales y no traducciones, ya que solo de esta manera se puede asegurar que la terminología empleada es la adecuada y la original en cada ámbito especializado,[3] así como es recomendable también que los textos estén escritos por hablantes nativos, para evitar cualquier tipo de influencia de la lengua materna del autor, errores (léxico, gramaticales, etc.) y cualquier giro o uso incorrecto; cuarto, los textos incluidos en los corpus deben ser textos completos, es decir, se debe evitar el uso de fragmentos con objeto de evitar la posible descontextualización o error en la información contenida; y, por último, según Pearson (1998: 60), los textos deben haber sido publicados previamente a su inclusión, con la idea de que adquieran un valor más formal, serio y respetable dentro del ámbito especializado. Asimismo, si el trabajo terminográfico está enfocado para unos países determinados, se deberá establecer unos límites geográficos para la búsqueda de los textos y, por tanto, el terminógrafo deberá comprobar que todos los textos pertenecen o han sido publicados dentro de esos límites geográficos. Con todos estos criterios de calidad se pretende conseguir la mayor fiabilidad posible de los textos incluidos en un corpus con un objetivo terminográfico.

El criterio de documentación ocupa un lugar muy relevante en la compilación, ya que permite llevar a cabo un trabajo sistemático y homogéneo. Consiste en registrar las referencias de los textos contenidos en el corpus (autor, lugar de publicación, fecha o actualización, etc.) con objeto de realizar un seguimiento de los textos seleccionados y de las fuentes de información utilizadas. También es recomendable utilizar un código

3 A pesar de que el corpus compilado para un trabajo terminográfico debe ser comparable (textos originales) para poder extraer la terminología adecuada del ámbito de especialidad en cada lengua de trabajo, también suele haber un subcorpus paralelo (textos originales y sus traducciones) que permite detectar y extraer los posibles equivalentes en otras lenguas.

unívoco entre el registro del texto y el propio texto para establecer una vinculación entre los datos registrados y el texto que nos permita acceder a su información de referencias.

Por último, el criterio de simplicidad hace referencia al tipo de información añadida al texto original. Esta información es principalmente morfológica, sintáctica, léxica y semántica y permite al terminógrafo utilizar el corpus de forma más eficiente y más precisa a la hora de realizar búsquedas, estudios concretos y clasificar la información contenida en los textos. En *terminografía*, las anotaciones más utilizadas son las semánticas y léxicas, que se utilizan para el estudio del discurso especializado, extracciones terminológicas o de patrones semánticos, aunque también se puede encontrar otro tipo de estudios lingüísticos.

3.2. Criterios específicos

Además de estos criterios generales expuestos anteriormente, consideramos que son útiles otros, de carácter más específico, aunque muy relacionados con los anteriores:

El criterio de delimitación de fronteras, que exige que, antes de comenzar con cualquier proyecto, los terminógrafos deben delimitar el campo de especialidad que tiene que representar el corpus de dos formas diferentes: «side boundaries», es decir, con respecto a los campos de especialidad más cercanos al objeto de estudio y «upper boudaries», con relación a los niveles de especialización que va a incluir el corpus (de más especializado a menos) (Meyer y Mackintosh, 1996). En relación con este criterio, los terminógrafos deberán compilar un corpus que represente lo mejor posible el campo de especialidad delimitado y, para ello, deberán seleccionar textos que permitan un equilibrio en todos los aspectos del dominio, cubriendo todos los subdominios y dominios relacionados de la forma más equitativa posible.

Otro criterio específico hace referencia a la apertura del corpus. En *Terminografía*, debido a la rapidez con la que se producen los cambios, el *corpus especializado* debe ser abierto, a fin de que se pueda ir actualizando, ya sea eliminando o incluyendo textos, con el paso del tiempo.

Por último, encontramos el criterio de pragmática, que se encuentra en la misma línea que el criterio de delimitación de fronteras, aunque en este caso hace referencia a la situación comunicativa para la que va destinado el producto final. Este criterio exige tener en cuenta la situación comunicativa, a saber: los receptores, el contexto, el tipo textual y el nivel de especialización que habrá en el momento en el que se emplee el recurso final. Así pues, dependiendo de quiénes sean los usuarios, para qué utilicen y cuándo utilicen el recurso final, el corpus estará formado por un tipo de textos u otros.

En nuestra opinión, estos serían los criterios que deberían tenerse en cuenta para compilar cualquier *corpus especializado* en un contexto terminográfico y, en definitiva, para garantizar la calidad del producto final.

4. Conclusiones

Como hemos visto, la *terminografía basada en corpus* se trata de la *terminografía* de actualidad, fruto del cambio de paradigma sufrido en la terminología que ha supuesto el paso de la terminología tradicional de Wüster a la terminología moderna, así como de la evolución de las herramientas informática que han facilitado el procesamiento de grandes cantidades de información en formato electrónico. Sin embargo, a pesar de lo extendido de su uso, aún seguimos muy vinculados a la metodología utilizada en la lexicografía basada en corpus, lo que limita a veces el trabajo terminográfico.

Por este motivo, necesitamos tener en cuenta las especificaciones de la *terminografía* para establecer una metodología propia que nos permita alcanzar los mejores resultados en nuestros proyectos y en las diferentes fases del trabajo terminográfico. En esta línea, hemos propuesto unos criterios generales y específicos dirigidos a la compilación de *corpus especializado*s para el trabajo terminográfico que nos permita extraer los mejores beneficios para el objetivo de nuestro proyecto, que a menudo difiere de los objetivos perseguidos en la lexicografía basada en corpus.

5. Bibliografía

Cabré Castellví, M. T. (1993). *La terminología. Teoría, metodología, aplicaciones*. Barcelona: Antártida/Empúries.

Cabré Castellví, M. T. (1999). Hacia una teoría comunicativa de la terminología: aspectos metodológicos. En M. T. Cabré. 2000 (Ed.). *La Terminología: Representación y Comunicación. Elementos para una teoría de base comunicativa y otros artículos* (pp. 129-150). Barcelona: IULA. Universidad Pompeu Fabra.

Corpas Pastor, G. y Seghiri Domínguez, M. (2007a). Determinación del umbral de representatividad de un corpus mediante el algoritmo N-Cor. *Procesamiento del lenguaje natural*, *39*, 165-172.

Corpas Pastor, G. y Seghiri Domínguez, M. (2007b). Specialized Corpora for Translators: A Quantitative Method to Determine Representativeness. *Translation Journal*, *11*(3). Disponible en http://www.translationjournal.net/journal/41corpus.htm

Leech, G. (1992). Corpora and Theories of Linguistic Performance. En J. Svartvik (Ed.) *Directions in Corpus Linguistics. Proceedings of Nobel Symposium* (pp. 105-134). Berlín/Nueva York: Mouton de Gruyter.

Meyer, I. y Mackintosh, K. (1996). The Corpus from a Terminographer's Viewpoint. *International Journal of Corpus Linguistics*, *1/2*, 257-285.

Seghiri Domínguez, M. (2006). *Compilación de un corpus trilingüe de seguros turísticos (español-inglés-italiano): aspectos de evaluación, catalogación, diseño*

y representatividad. Tesis Doctoral. Málaga: Servicio de Publicaciones de la Universidad de Málaga.

Herramientas y criterios para la creación de un banco de conocimiento sobre los usos del lenguaje en la red

Joseba Ezeiza Ramos y Agurtzane Elordui Urkiza.

Universidad del País Vasco/Euskal Herriko Unibertsitatea. Departamento de Filología Vasca.

Resumen

Las nuevas formas y modos de comunicación en la red han generado un gran interés de los lingüistas y otros profesionales por los usos del lenguaje en los medios digitales. Una de las vías de investigación que promete ser productiva en este ámbito es la basada en el estudio de corpus. Sin embargo, las particulares características de la comunicación digital y de los cibertextos exigen reconceptualizar algunas ideas, revisar algunos criterios y adaptar algunos instrumentos habitualmente aplicados en este tipo de trabajos. El proyecto que se presenta a continuación trata, precisamente, de aportar algunas herramientas y criterios que nos permita abordar dichos retos y dar respuesta a algunas de las necesidades que plantea el estudio de los usos lingüísticos en la comunicación digital.

Palabras Clave: corpus, cibermedios, cibeliteratura, análisis facetado

Abstract

The new ways of communication in the network have generated a great interest among linguists and other professionals on the language usages in Internet. One of the lines of investigation that promises to be particularly productive in this study is corpus linguistics. However, the specific characteristics of the cybertexts demand to re-conceptualize some criteria of that field and also to adapt the tools used till now to suit the needs of searching and classifying digital texts. That is the main goal of this project, to offer some criteria and tools that able us to study the linguistic usages in digital communication.

Key Words: Corpus, cybermedia, cyberliterature, faceted analysis

0. Introducción

En este trabajo se presentan las herramientas y los criterios de un proyecto basado en la metodología de corpus. El proyecto tiene como objetivo el estudio y la caracterización de los usos lingüísticos en la red. Se trata de un proyecto I+D+I[4] que comprende tres vertientes: Una *vertiente de investigación teórica,* interesada por estudiar las características lingüístico-dicursivas de la comunicación digital y la influencia de los medios digitales e Internet en los usos lingüísticos. Una *vertiente tecnológica*, que persigue desarrollar metodologías e instrumentos que permitan estudiar los usos lingüísticos y su evolución con un cierto nivel de refinamiento. Y una *vertiente aplicada* que pretende ofrecer recursos innovadores de información, formación y consulta al conjunto de la sociedad, pero, más en particular, a los estudiantes y profesionales del ámbito de la comunicación.

Las bases teóricas del proyecto se asientan sobre propuestas de análisis y clasificación textual que tienen especialmente en cuenta el contexto sociodiscursivo de los textos, y que permiten acercarse al estudio de los cibertextos desde persperctivas multiples. El esfuerzo realizado en el ámbito tecnológico se ha centrado en el desarrollo de herramientas que permitan clasificar, recuperar y analizar tanto textos periodísticos como literarios creados para su difusión a través de Internet. Por último, para poner en marcha el proceso de generación de materiales de consulta y formación con base empírica, se ha creado una estructura integrada de diversos recursos tecnológicos, pensados para que puedan servir de soporte para la elaboración de un banco de conocimiento u observatorio sobre los usos de la lengua en los medios digitales y, más concretamente, en Internet.

A continuación se desglosarán brevemente los aspectos medulares del proyecto. En primer lugar, explicaremos la perpectiva teórico-metodológica desde la que abordamos nuestra propuesta y algunos de los retos que se nos plantean en el estudio de los usos linguísticos en internet. En segundo lugar, se presentarán algunos de los resultados más relevantes del trabajo realizado hasta la fecha. Finalmente, se hará referencia a algunos trabajos de desarrollo actualmente en curso, y se apuntarán también algunas posibles líneas de continuidad a abordar en el futuro.

1. perspectiva teórico-metodológica

El proyecto se aborda desde un enfoque multidimensional de los usos lingüístico-discursivos (Biber 1988, 1995; Bathia 1993, 2004). Dentro de este enfoque adoptamos el modelo de análisis propuesto por Bathia (1993, 2004), *Multi-perspective model of discourse* que integra una perspectiva socio-critica y pedagógica del discurso. Así entendemos el discurso en la red como una práctica social, como una práctica profesional, y también como un género y como un texto.

Desde esta visión la taxonomía y las herramientas que proponemos para el estudio de los usos lingüísticos en la red nos deben permitir estudiar dichos usos tanto desde una

4 Este trabajo se adscribe a los proyectos DB (OTRI: 2007.0077), EHLB (OTRI: 2008.0368), HIZLAN (DIPE08/16), EBALUA (EHU08/53) y GARATERM2 (US10/01).

perspectiva que tenga en cuenta un contexto sociocultural amplio, como desde una perspectiva que se centre en los aspectos formales y funcionales de los textos digitales teniendo en cuenta sólo su cotexto.

Por otra parte, el proyecto se asientan en los estudios de tres líneas de investigación abiertas entorno a los cibertextos. Estos estudios nos han acercado a las particularidades de la práctica social y profesional del ciberperiodismo y la ciberliteratura y, sobre todo, a la caracterización lingüístico-discursiva de los genéros y textos de la red, así como a las metodologías para su recopilación y clasificación.

La primera de ellas es la línea de investigación que se centra en la caracterización lingüístico-discursiva de los textos de la red, en particular aquellas investigaciones que toman en cuenta las especificidades contextuales y cotextuales de estos textos (Landow 1997; Crystal 2001; Engebretsen 2001; Storrer 2002; Posteguillo 2003; Santini 2005 a, b; Biber & Kurjian 2007, entre otros). Estos estudios evidencian la necesidad de que una taxonomía de documentos web tenga en cuenta características contextuales específicas del medio digital, en particular características como la funcionalidad (incluyendo en ella la interacción y la interactividad) la hipertextualidad y la multimedialidad. Por ejemplo, al clasificar los ciberdocumentos de la web se debe tener en cuenta que los documentos digitales forman parte de una pieza o unidad informativa/comunicativa articulada de forma hipertextual o hipermedia. La taxonomía debe dar cuenta de los diferentes elementos de esa pieza comunicativa, y también del lugar que el documento ocupa en la estructura hipertextual o hipermedia. Debe recoger, por lo tanto, cómo se complementa el documento a clasificar con el resto de documentos de la estructura; si hay una relación de yustaposición o integración con el resto de los documentos de la pieza comunicativa, sean estos documentos textos, vídeos, audios, infografías, etc. No debemos olvidar que uno de los mayores retos lingüísticos del cibertexto consiste en alcanzar la unidad comunicativa en mensajes que contengan todos esos ingredientes lingüísticos y visuales.

La segunda línea de investigación en la que hemos ahondado a la hora de hacer nuestra propuesta taxonómica, han sido los estudios de género y más concretamente las propuestas de clasificación de los cibergéneros periodísticos (Díaz Noci & Salaverría 2003; Lamarca 2007; Larrondo 2008) y literarios (Borras 2005). Analizar el discurso por géneros nos ayuda a entender cómo los miembros de una comunidad de discurso construyen e interpretan esos géneros para lograr sus objetivos comunicativos, y también para saber por qué lo hacen de la manera que lo hacen. Además la clasificación por géneros responde a los objetivos aplicados para los que pretendemos la base documental, especialmente para aquellos de carácter didáctico. Sin embargo esta labor conlleva muchas dificultades en el caso de los cibertextos (Santini 2005 a, b). En particular debemos tener en cuenta que características como la hipertextualidad y la multimedialidad hacen que aumente aún más la tendencia a la hibridación de géneros, particularmente en el caso de textos periodisticos y literarios para los que la hibridación es ya un característica inherente.

Finalmente hemos tenido en cuenta las propuestas teóricas y metodológicas que desde la lingüística del corpus focalizan en el acceso, búsqueda y recopilación de textos en la

web (Fletcher 2007, Renouf& Kehoe & Banerjee 2007) o para corpus monitor *offline*. En especial hemos tenido en cuenta aquellas propuestas y experiencias que tratan la recopilación de estos textos para estudio de variación lingüística y cambio (Mair, 2007) y en particular las basadas en criterios de registro y género (Lim, Lee, Kim 2004; Santini 2005 a, b; Biber & Kurjian 2007). Estas últimas nos ofrecen resultados interpretables en términos sociolingüísticos y discursivos y, por lo tanto, responden de forma más efectiva a los objetivos teóricos y aplicados para los que va orientada la herramienta taxonómica que hemos creado.

En las mencionadas líneas de investigación se destacan las principales dificultades o retos que debe afrontar cualquier estudio de usos lingüísticos en la red y, sin duda, de un trabajo taxonómico como éste que tiene como objetivo el estudio de esos usos. El primer reto que se nos planteaba es cómo identificar y acotar los documentos a cargar en el corpus sin perder elementos con relevancia comunicativa vinculados con el mismo. Como ya hemos señalado que la hipertextualidad y la multimedialidad abren un amplio abanico de posibilidades para crear documentos en los que se vinculan, integran o yuxtaponen diversos textos y otro tipo de elementos que pueden ser interactivos y audiovisuales, y ello provoca dificultades para determinar los límites de un documento digital. Esta complejidad nos obliga, en cierto modo, a tratar de compilar junto al material de naturaleza textual del propio documento tanto el conjunto de elementos comunicativos que lo conforman (pieza textuales, imágenes, audios, etc.) como una cierta "imagen" que ponga de manifiesto el contexto digital en el que situamos dicho documento (por ejemplo, una captura de pantalla). En ocasiones hemos considerado necesario conservar la imagen de las diversas interfaces desde las que se puede acceder a dicho documento (webs, blogs, redes sociales…), ya que el entorno puede condicionar su función discursiva y su interpretación.

Por otra parte, dentro del paradigma digital se ha producido un rápido desarrollo del modelo social 2.0 que, entre otras contribuciones, ha creado nuevas formas de gestionar el conocimiento. Una de las aportaciones más relevantes de esta aproximación son las *folksonomías* o sistemas complejos de identificación y caracterización de documentos desarrollados por los usuarios de manera, en general, bastante intuitiva y asistemática. En este contexto, nacen nuevas "etiquetas" y "colecciones de etiquetas" que, en muchos casos, sustituyen las nomenclaturas y modelos clasificatorios tradicionales. Esto nos plantea buscar una vía de convergencia entre las tradicionales fórmulas jerárquicas y cerradas de clasificación habitualmente utilizadas para estructurar los corpus y los nuevos esquemas horizontales y abiertos derivados de la filosofía 2.0.

Dar respuesta a estos retos ha requerido trabajar en un triple frente. Por una parte, hemos tenido que establecer una definición, al menos tentativa, de la unidad nuclear de comunicación. Una definición que nos permita acotar de forma operativa los límites de un documento digital; esto es, que permita identificar dentro de la compleja maraña de elementos comunicativos interconectados esa unidad de comunicación. Por otra parte, hemos tenido que establecer qué elementos vinculados a dicha unidad nuclear conviene compilar en el corpus, junto al documento textual en cuestión. Finalmente, hemos tenido

que trazar los ejes principales de una taxonomía que nos permitiera clasificar de forma matizada las producciones textuales que deseamos considerar en nuestra tarea. Para ello hemos aplicado un sistema taxonómico de tipo facetado que, por un lado, reconoce la idiosincrasia de cada producción lingüística en particular, pero que, al mismo tiempo, permite evidenciar los nexos comunes entre unos documentos y otros, por muy puntuales y discretos que estos sean. En cierto modo, con ello se asume la idea de que cada producción particular se sitúa en un punto de un mapa multidimensional, representado por un vector de rasgos taxonómicos que lo vinculan de una forma específica y particular al sistema de comunicación lingüística en la red. Por otra parte, el modelo taxonómico facetado propuesto ofrece no solo una solución para estructurar el corpus a caballo entre las aproximaciones 1.0 y 2.0, además nos situaría en la dirección hacia la que apuntan algunos sistemas que tratan de reconducir las folksonomías hacia el aún emergente paradigma 3.0, aplicando ciertos criterios para estandarizar el etiquetado, asociando a las etiquetas determinados valores semánticos (Epelde, 2010; Payá, 2010).

2. Criterios y herramientas

Partiendo de los postulados apuntados en el apartado anterior, el año 2007 se abordó el desarrollo de una infraestructura para compilar, caracterizar, clasificar, recuperar y analizar un amplio espectro de documentos y materiales distribuidos por canales electrónicos: documentos tipo *pdf*, *dot* o *doc*, documentos *htm* o *html*, audios, vídeos, aportaciones e interacciones desarrolladas en redes sociales, etc. Con este fin, se estableció un convenio de colaboración entre el Departamento de Filología Vasca de la UPV/EHU[5] y el grupo Ametzagaiña A.I.E.[6], unidad de I+D empresarial de la Red Vasca de Tecnología.

Así pues, en sucesivos proyectos fueron desarrollándose los módulos necesarios para la generación y consulta de instrumentos de corpus, así y como su administración a través de la red. Dichos instrumentos han quedado integrados en HIZLAN (Ezeiza, 2009-b y 2010), una plataforma trabajo en línea[7], concebida para la gestión de la calidad lingüística de la comunicación y para la generación de recursos formativos a partir de corpus textuales. Uno de los módulos de esta plataforma ha sido específicamente adaptado para el estudio de los usos del lenguaje en los medios digitales (ciberperiodismo y ciberliteratura principalmente).

El diseño de este módulo requirió abordar tres tareas principales: a) desarrollo de un protocolo adecuado, para compilar, caracterizar y clasificar de forma matizada documentos electrónicos diversos atendiendo a sus rasgos comunicativos más relevantes; c) diseño e implementación una arquitectura informática que permita realizar en línea las operaciones de carga, indexación, procesamiento, recuperación y consulta de documentos, de forma ágil, segura y fiable; y, c) identificación y selección de aquellas metodologías de análisis y tratamiento de datos que pudieran responder adecuadamente a los objetivos trazados.

5 URL: http://www.ef.ehu.es/s0113-home1/es
6 URL: http://www.ametza.com/castellano/index.htm
7 URL: http://www.hizlan.org

En primer lugar, se estableció una definición operativa de unidad nuclear de comunicación con el objeto de determinar qué características deberían cumplir los documentos susceptibles de ser incorporados al corpus. Esta definición se adaptó de la propuesta por Salaverria (2001) para el concepto de "unidad mínima de comunicación", incorporando una serie de matices que hacen referencia a rasgos propios de la comunicación digital que no habían sido considerados en la formulación original.

Así pues, en una acepción general, la noción de unidad MM&D se refiere a la *unidad mínima estructurada de significación* constituída por uno o más *elementos de comunicación textual y/o multimedia* que presenta *autonomía funcional y unidad de contenido* dentro del contexto digital en el que se presenta, y que configura un *espacio de comunicación* en el que se optimizan *códigos, recursos y herramientas* para armonizar los procesos de *información, participación e interacción* que pretende facilitar. En general, una unidad MM&D estará formada por un *marco, un elemento nuclear de naturaleza textual y una serie de elementos secundarios* que pueden ir *integrados* en el cuerpo del texto principal o que pueden ser también *externos* o *periféricos*. En cualquier caso, formarán parte de la unidad MM&D solo aquellos *elementos consubstanciales a la finalidad comunicativa* para la que esta haya sido desarrollada, sean estos *atribuibles a los autores originales* o bien sean producto de las *aportaciones de usuarios distintos a ellos*.

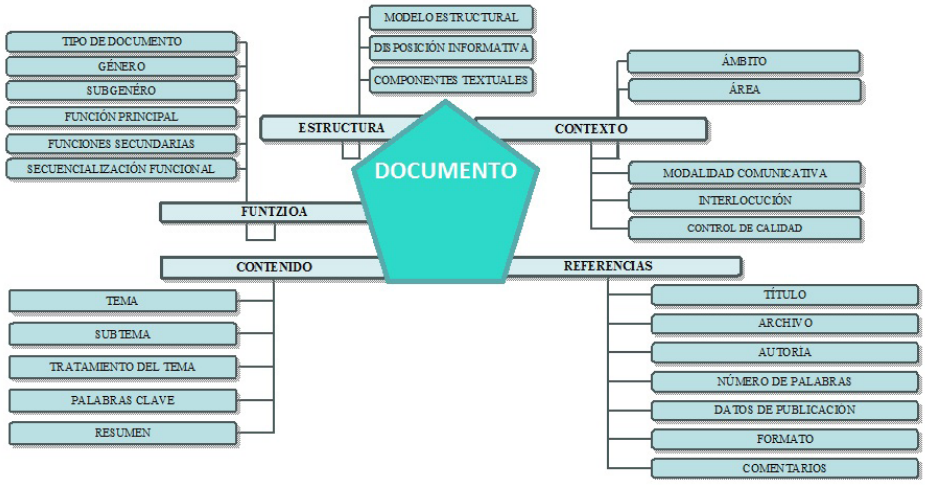
Una vez definido un criterio general para identificar y acotar los documentos a cargar en el corpus, se estableció un protocolo para introducir en la base de datos tanto el elemento nuclear del documento, como el resto de elementos de relevancia comunicativa vinculados con el mismo. Así, se determinó que el análisis documental (metadatos) se aplicará al elemento nuclear de la unidad comunicativa, y el resto de textos, imágenes y elementos asociados se cargaran como anexos. Por otra parte, siguiendo el criterio apuntado en el apartado anterior, también se determinó que los documentos del corpus se acompañaran de, al menos, una imagen (captura de pantalla o similar) que reflejaran con cierto nivel de detalle el entorno digital (interfaz) en el que fueron localizados el documento en cuestión y todos los elementos a él asociados. Para que esto fuera posible, hubo que realizar una serie de desarrollos técnicos que permitieran la carga y recuperación de todo este material así como el procesamiento selectivo a demanda del material textual que acompañe al documento principal, sin interferencia en los procesos primarios de indexado y consulta.

Una vez adoptados los criterios para afrontar los dos primeros problemas detectados, abordamos el desarrollo de la estructura taxonómica que nos permitiera caracterizar y clasificar los documentos digitales de una manera sistemática y altamente tipada, pero, al tiempo, flexible y abierta, siempre partiendo de la visión arriba apuntada de las producciones textuales como *acciones de comunicación* (Ciapuscio, 2003; Ezeiza, 2009; Ezeiza, Payá, Elordui y Epelde, 2011).

Así pues, tal y como se puede observar en la figura nº 1, en un primer nivel de concreción del esquema taxonómico se encontrarían las cuatro *dimensiones o facetas* principales contempladas en la definición de *unidad nuclear de comunicación* antes apuntada: contexto de producción, función comunicativa, contenido semántico y estructura interna.

A cada dimensión le correspondería dentro del módulo de carga del instrumento de corpus un *formulario* independiente. Cada formulario incorpora una serie de *campos* (segundo nivel de concreción del esquema taxonómico) que permiten atribuir a los documentos diversos tipos de (*meta)datos*. Cada uno de estos campos, a su vez, desplegará un listado variable (y abierto) de *etiquetas* que, en definitiva, reflejarán en un tercer nivel de concreción, los *rasgos* idiosincrásicos particulares del documento. En la figura nº 3 únicamente se recogen las categorías correspondientes a los dos primeros niveles de análisis. En el apéndice que se puede consultar al final del artículo, se puede consultar el esquema completo de análisis desarrollado para la elaboración del corpus ZIBERDOK, concebido para el estudio de los usos de la lengua en los medios de comunicación de masas en la red.[8]

Figura 1: Desarrollo de los dos primeros niveles del esquema taxonómico facetado



La arquitectura informática[9] que da soporte al sistema de gestión de corpus reproduce sobre una estructura de base de datos el esquema que hemos descrito. El módulo central lo constituye una base documental (*Dokumentu Biltegia*[10]) a la que se accede desde cualquier punto de la red. Dicha base documental cuenta con un módulo de carga que permite incorporar al corpus el documento nuclear y todos los elementos que lo acompañan, de forma ágil y sencilla. Cuenta con un segundo módulo para identificar y caracterizar en todas las facetas previstas en el esquema taxonómico tanto el documento principal como los anexos. Para analizar los usos lingüísticos en un documento concreto o en un grupo de documentos que comparten determinados rasgos, se cuenta con una serie de herramientas de análisis que operan sobre el material textual previamente indexado, lematizado y etiquetado.

8 Agradecemos la inestimable aportación de Xabier Payá e Imanol Epelde (investigadores contratados en el proyecto DIPE08/16) al desarrollo de ambos corpus.

9 ® SS-239-09

10 ® SS-236-09

Se trata de un paquete de utilidades que permite operar y obtener información lingüística en tres niveles: textual, morfosintáctico y léxico. Este módulo está pensado tanto para consultas puntuales, como para estudios sistemáticos. Y ofrece información tanto cualitativa (contextos de aparición) como cuantitativa (ocurrencias y frecuencias). El motor de búsqueda documental que opera sobre la estructura taxonómica arriba descrita facilita la discriminación selectiva de documentos y también la realización de comparativas y, en su caso, la exportación de textos y datos a otros programas y plataformas. Todo ello permite la obtención y edición de segmentos textuales representativos de los usos por los que podamos estar interesados, y también agiliza la realización de análisis estadísticos bien de tipo descriptivo, bien de tipo inferencial o incluso predictivo, exportando los datos ofrecidos por el sistema a un programa de tratamiento de datos adecuado.

3. Líneas futuras de trabajo

La fase de desarrollo técnico y pilotaje del proyecto ha finalizado en diciembre del 2010. Los resultados más relevantes de la fase de desarrollo son los que acabamos de señalar: a) un constructo teórico y metodológico para el análisis de los usos lingüísticos en los medios digitales; b) un esquema taxonómico para clasificar de forma matizada las producciones lingüísticas en la red; y c) una arquitectura informática que nos permite generar, administrar y consultar corpus de textos electrónicos de una manera ágil y accesible. Estos resultados han quedado recogidos en tres informes de investigación (Ezeiza, 2009c; Ezeiza, Payá, Epelde & Elordui, 2011 y Ezeiza, Epelde, Elordui & Payá, 2011) y cuatro herramientas informáticas originales integradas en la plataforma HIZLAN (http://www.hizlan.org)[11].

En esta primera fase nos hemos centrado en el desarrollo y la validación téorica, técnica y empírica de estos tres aspectos. Como resultado aplicado de este esfuerzo contamos con un pequeño corpus de textos electrónicos que contiene una amplia variedad de producciones de diversa naturaleza y complejidad, con sus correspondientes modelos de análisis. Desde el punto de vista cuantitativo es aún insuficiente para realizar estudios lingüísticos detallados, pero desde el punto de vista cualitativo creemos que ofrece soluciones válidas a las principales dificultades que nos plantean este tipo de textos.

De cara al futuro, nos enfrentamos ahora al reto de establecer los criterios para alimentar, estructurar y equilibrar adecuadamente el corpus, de manera que nos pueda aportar la información necesaria para estudiar la configuración de los usos en los nuevos canales y medios.

Hemos comenzado ya a caminar en esta dirección. Como primer paso, nos encontramos actualmente tratando de estructurar la red conceptual que nos permitiría describir dichos usos, siempre desde una aproximación a ellos que asume su inherente complejidad. Esperamos poder presentar próximamente el sistema de categorías discursivo-lingüísticas que pueden resultar relevantes para el análisis de las cibertextos de los medios de

11 ® SS-236-09, SS-27-11, SS-29-1 y SS-28-11

comunicación y la literatura. Para ello nos apoyamos en resultados empíricos basados en el estudio de los corpus que estamos desarrollando actualmente.

APÉNDICE: Desarrollo del tercer nivel del esquema taxonómico

REFERENCIAS: IDENTIFICACIÓN DEL DOCUMENTO (ficha de carga)

- **TÍTULO O CÓDIGO DE IDENTIFICACIÓN**
- **AUTOR(ES)**
- **DATOS DE PUBLICACIÓN**
- **FORMATO: pdf, mp3, HTML…**
- **Nº PALABRAS (automático)**
- **COMENTARIOS**

DIMENSIÓN-1: CONTEXTO

ÁMBITO: MEDIO

- Diario de actualidad general
- Revista de actualidad general
- Diario especializado
- Revista especializada
- Boletín informativo
- Web temática
- Web educativa
- Web corporativa o institucional
- Web comercial
- Blog de un medio de comunicación
- Blog corporativo o institucional
- Blog comercial
- Blog colectivo o social
- Blog personal
- Microblog
- Foro de discusión
- Red social
- Otros (carga manual)

ÁREA: SECCIÓN

- Portada
- Contraportada
- Internacional
- España
- Francia
- Euskal Herria
- Provincial o regional
- Local
- Sociedad
- Cultura
- Deportes
- Economía
- Ciencia y tecnología
- Literatura/Libros
- Cine y audiovisuales
- Radio y televisión
- Artes plásticas
- Opinión
- Entrevistas/Personajes
- Colaboraciones
- Agenda
- Pasatiempos
- Suplemento
- Periodismo ciudadano
- Videoteca
- Fonoteca
- Álbumes de fotos
- Publicidad
- Otros (carga manual)

MODALIDAD COMUNICATIVA

- Privada / Pública
- General / Especializada
- Síncrona / Asíncrona
- Estática / Dinámica
- Permanente / Temporal
- Iniciativa: Individual / Colaborativa / Social
- Interacción: Unidireccional / Bidireccional / Multidireccional
- Formato: Textual / Multimedia / Interactivo / Animado
- Soporte: Documento / Hiperdocumento / Hipermedia
- Canal original: Televisión / Radio / Papel / Internet
- Otros (carga manual)

INTERLOCUCIÓN

- **Emisor:** agencia, empresa, institución, equipo de redacción, periodista identificado, colaborador, colectivo social, experto o especialista, particular, otro (carga manual)
- **Receptor:** público en general, público especializado, colectivo social o profesional determinado, otro (carga manual)

CONTROL DE CALIDAD

- Espontáneo , autogestionado
- Interno: colegiado
- Interno: departamento de calidad lingüístico
- Externo (indicar empresa o institución)
- No consta
- Otros (carga manual)

DIMENSIÓN-2: CONTENIDO

- **ÁREA TEMÁTICA (criterios CDU)**
- **TEMA (criterios CDU)**
- **TRATAMIENTO DEL TEMA**

- General
- Teórico
- Aplicado
- Didáctico
- Divulgativo
- Crítico
- Creativo

- Objetivo
- Subjetivo

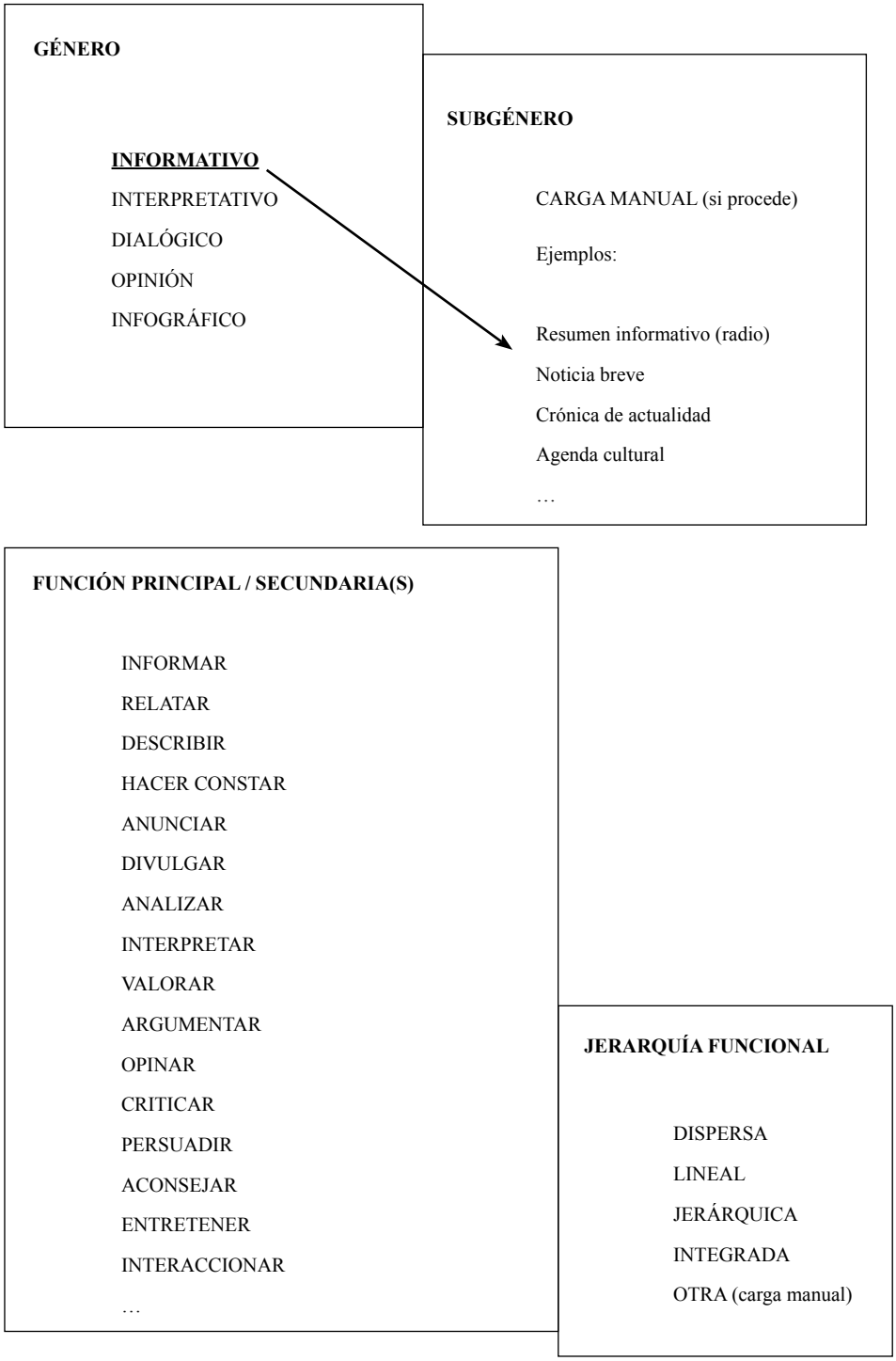
- Otro(s) (carga manual)

- **RESUMEN**
- **PALABRAS CLAVE**

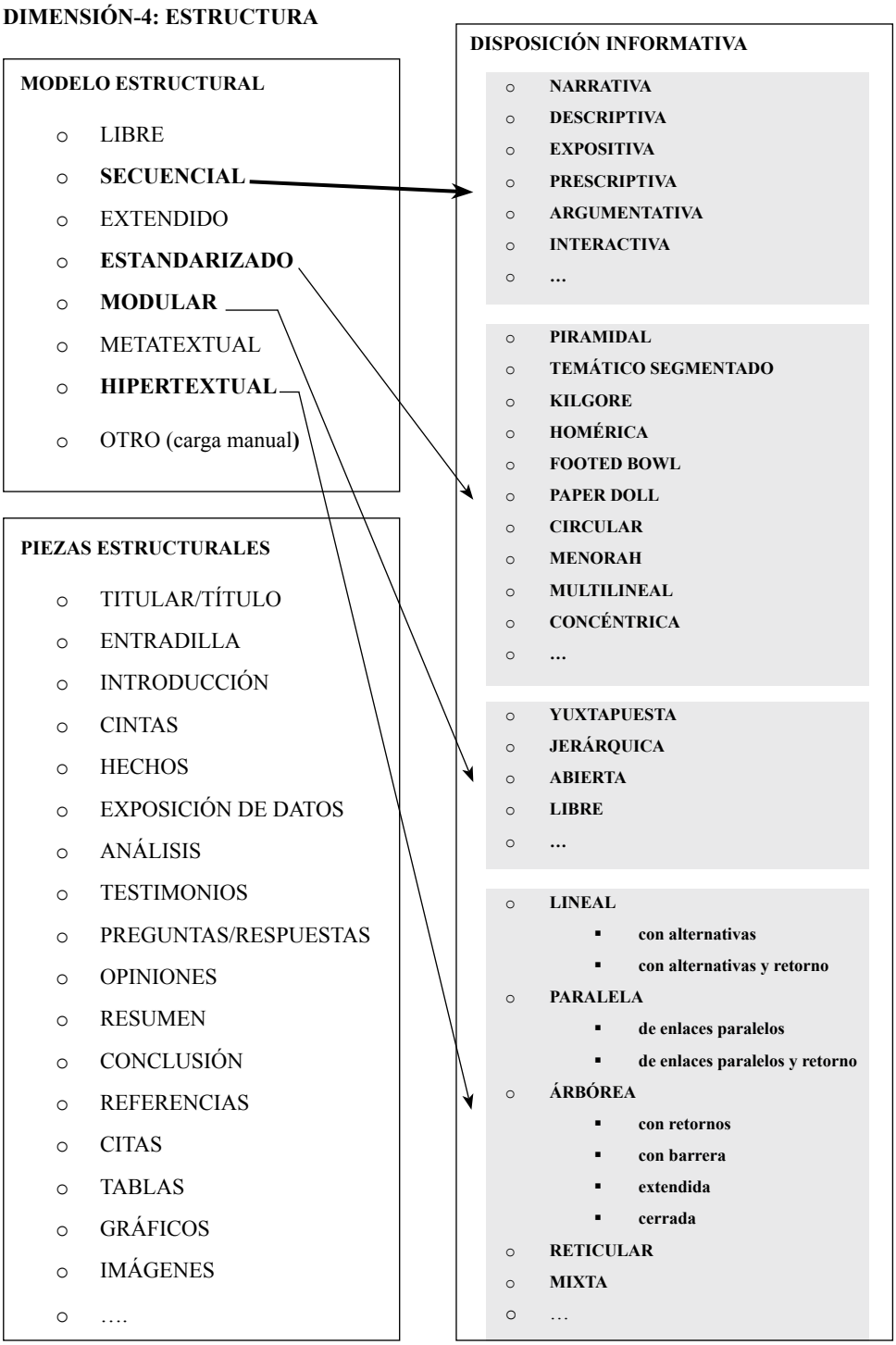
DIMENSIÓN-3: CARACTERIZACIÓN FUNCIONAL

TIPO DE DOCUMENTO

	ENTRADA DE BLOG
NOTICIA	COMENTARIO
RELATO	MICROTEXTO
TESTIMONIO	RECETA
ENTREVISTA	ANUNCIO
CRÓNICA	AVISO
REPORTAJE	CÓMIC
CARTA	HILO DE DISCUSIÓN
ARTÍCULO DE OPINIÓN	CANAL DE NOTICIAS
DEBATE	MURO DE FACEBOOK
CRÍTICA	CADENA DE MICROTEXTOS
EDITORIAL	TEXTO LITERARIO: relato
NOTA ACLARATORIA	TEXTO LITERARIO: poesía
GUÍA	TEXTO LITERARIO: otro (carga manual)
CATÁLOGO	DEBATE
PROGRAMA	NECROLÓGICA
MENSAJE ELECTRÓNICO	OTROS (carga manual)


GÉNERO
INFORMATIVO
INTERPRETATIVO
DIALÓGICO
OPINIÓN
INFOGRÁFICO
SUBGÉNERO
CARGA MANUAL (si procede)
Ejemplos:
Resumen informativo (radio)
Noticia breve
Crónica de actualidad
Agenda cultural
…
FUNCIÓN PRINCIPAL / SECUNDARIA(S)
INFORMAR
RELATAR
DESCRIBIR
HACER CONSTAR
ANUNCIAR
DIVULGAR
ANALIZAR
INTERPRETAR
VALORAR
ARGUMENTAR
OPINAR
CRITICAR
PERSUADIR
ACONSEJAR
ENTRETENER
INTERACCIONAR
…
JERARQUÍA FUNCIONAL
DISPERSA
LINEAL
JERÁRQUICA
INTEGRADA
OTRA (carga manual)

DIMENSIÓN-4: ESTRUCTURA



Referencias

Bathia, V. K. (1993). *Analysing Genre-Language Use in Professional Settings*. Londres: Longman.

Bathia, V. K. (2004). *Worlds of written discourse*. *A genre-based View*. Londres: Continuum.

Biber, D. & Kurjian J. (2007). Towards a taxonomy of web registers and text types: a multidimentional analysis. En Hunt, M., Nesselhauf , N. & Biewer, C. (eds.) *Corpus Linguistics and the web*, 109-132.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1995). *Dimenstions of Register Variation: a Cross-Linguistic Perspective*. Cambridge: Cambridge University Press.

Borras, L. (2005). Teorías literarias y retos digitales. En Borras, L. (Eds.) *Textualidades electrónicas*. Barcelona: EDIUOC.

Crystal, D. (2001). *Language and the internet*. Cambridge: Cambridge University Press

Díaz Noci, J. & Salaverría R. (coords) (2003). *Manual de Redacción Ciberperiodística*. Barcelona: Ariel.

Engebretsen, M. (2001). Hypernews and coherence. *Journal of Digital Information*, Vol 1, Nº 7.

Epelde, I. (2010). *Euskarazko literatur testuak sarean: folksonomia baterako proposamena.* Memoria del máster universitario en Comunicación Multimedia EITB-UPV/EHU (edición 2009/10). Disponible en http://www.27zapata.com/wp-content/uploads/2010/09/Euskarazko-literatur-testuak-sarean-folksonomia-baterako-proposamena.pdf

Ezeiza, J. (2009). Herramientas para la compilación, estudio y gestión de la producción lingüística en la universidad: una aproximación didáctica y social. En Caridad de Otto, E. & López de Vergara (comp.). *Las lenguas para fines específicos ante el reto de la Convergencia Europea.* La Laguna: Universidad de la Laguna, 553-567

Ezeiza, J. (2009b). *Criteris, metodologies i eines per a la gestió, l'estudi i la dinamització de la producció lingüística a la universitat: una aproximació socia*l. Barcelona: UAB. Disponible en http://www.slideshare.net/sdl_uabidiomes/josebaezeizaproducciolinguistica

Ezeiza, J. (2009c). *Dokumentu biltegia*: *Proiektuaren diseinua eta protokoloa jasotzen duen txostena*. Informe de investigación no publicado. ® SS-239-09.

Ezeiza, J. (2010). DB (*Dokumentu Biltegia*): corpus akademikoak sortzeko eta kudeatzeko azpiegitura teknologikoa. En Salaburu, P. eta Alberdi, X. (arg.), 2010. *Euskararen garapena esparru akademikoetan*. Leioa: UPV/EHU, 168-190.

Ezeiza, J., Epelde, I., Payá, X & Elordui, A. (2011). *Literaturaren alorreko dokumentu digitalak sailkatzeko taxonomía.* Informe de investigación no publicado. ® SS-26-11.

Ezeiza, J., Payá, X, Epelde, I. & Elordui, A. (2011). *Hedabideetako dokumentu digitalak sailkatzeko taxonomia*. Informe de investigación no publicado. ® SS-25-11.

Ezeiza, J., Payá, X., Elordui, A. y Epelde, I. (2011, en prensa). Towards a faceted taxonomy to structure web-genre corpora. *Revista de Lingüística y Lenguas Aplicadas*. (Aceptado el 19/04/2011).

Fletcher, W. H. (2007). Concordancing the web: promise and problems, tools and techniques. En Hunt, M., Nesselhauf , N. & Biewer, C. (Eds.) *Corpus Linguistics and the web*, 25-46.

Hunt M., Nesselhauf N. y Biewer C. (Eds.). (2007) *Corpus Linguistics and the web.* Amsterdam: Rodopi.

Lamarca, M. J. (2007). *Hipertexto: el nuevo concepto de documento en la cultura de la imagen*. Madrid: Universidad Complutense de Madrid.

Landow, G (1997). *Teoría del hipertexto*. Barcelona: Paidós

Larrondo, A. (2008). *Redacción Ciberperiodística. Contexto, teoría y práctica actual*. Leioa: Universidad de País Vasco.

Lim C.S., Lee, K.J. &. Kim, G.C. (2004). Multiple sets of features for automatic genre classification of web documents. *Elsevier Information Processing and Management*, 1263-1276.

Mair, C. (2007). Change and variation in present-day English: integrating the analysis of closed corpora and web-based monitoring. En Hunt, M., Nesselhauf , N. & Biewer, C. (Eds.) *Corpus Linguistics and the web*, 233-248.

Payá, X. (2010). *Ziberhedabideetako dokumentuen taxonomia sistema baterako proposamena*. Memoria del máster universitario en Comunicación Multimedia EITB-UPV/EHU (edición 2009/10). Disponible en http://issuu.com/xabipaya/docs/mkm_txostena

Posteguillo, S. (2003). *Netlinguistics. An analytical Framework to study Language, Discoourse and Ideology in Internet*. Castelló: Publicacions de la Universitat Jaume I. Universitas,10.

Renouf A., Kehoe, A. & Banerjee, J. (2007). Webcorp: an integrated system for web text search. En Hunt, M., Nesselhauf , N. & Biewer, C. (Eds.) *Corpus Linguistics and the web*, 47-68.

SALAVERRIA, R. (2001). Aproximación al concepto de multimedia desde los planos comunicativo e instrumental. *Estudios sobre el mensaje periodístico* 7. Disponible en http://www.ucm.es/info/emp/Numer_07/7-5-Inve/7-5-13.htm

SANTINI M. (2005a). *Linguistic Facets for Genre and Text Type Identification: A Description of Linguistically-Motivated Features*. Brighton: Brighton University.

SANTINI, M (2005b). Clustering Web Pages to identify emerging Textual Patterns. *RECITAL* 2005, Dourdan

STORRER, A. (2002). Coherence in text and hypertext. *Document Design* 3(2), 157-168.

Are you a man? On seeing gender in Shakespeare

Heather Froehlich

University of Strathclyde

Through a literary-linguistic approach I will present a new patterns of gender in Early Modern drama. I suggest that the use of gender-specific terms are not in proportion to the character population of a play. Using AlphaX and Excel, I assemble examples of grammatical gender within lines of a play. This study presents an overview of grammatical (subject/object) and thematic roles through a comparative study of third-person personal pronouns nouns in Macbeth and The Merry Wives of Windsor through the building of a pilot database of each word within the context of a sentence. The implications of proportional representation of a cast have been largely ignored in (feminist) stylistic studies of Shakespeare's texts - through the building of this database, I comment on the predictability of gender representation through semantic, rather than grammatical, relationships.

Keywords: corpus stylistics, gender, Shakespeare, Early Modern English

Usando un acercamiento literario-lingüístico, esta ponencia expone una pauta de género nunca estudiada en el drama de principios de la edad moderna inglesa. Sugiere que, en algunas obras de teatro, el uso de términos de género no está en proporción con el número y género de los personajes. Usando los programas AlphaX y Excel, he recopilado ejemplos de género gramatical dentro de los renglones de las obras. Este estudio presenta un panorama general de los papeles temáticos y gramaticales (sujeto/objeto) a través de un estudio comparativo de los pronombres personales de tercera persona en Macbeth y The Merry Wives of Windsor, aprovechando así la construcción de una base de datos de cada palabra dentro del contexto de la frase. Las implicaciones de la representación proporcional de los actores/personajes han sido, por lo general, ignoradas en los estudios estilísticos, y feministas, de los textos de Shakespeare. Por medio de la construcción de esta base de datos, esta ponencia ofrece comentarios sobre la previsibilidad de la representación de género a través de las relaciones semánticas, más que las relaciones gramaticales.

Palabras clave/Términos de búsqueda: estilística basada en corpus, género, Shakespeare, inglés moderno temprano.

Feminist stylistic and literary critics (Mills, 1995; McEachern, 1988; Parker, 1996; Howard & Rackin, 1997; Mendelson & Crawford, 1996; Levin, 1988; Black & Coward, 1988) often make sweeping generalizations about women without any concrete linguistic evidence of objecthood, such as argument structure or complements. These critics appear to "understand the meaning of a linguistic message solely on the basis of the words and structures of the sentences(s) used to convey that message" and note that "we certainly rely on syntactic structure and lexical items used in a linguistic message to arrive at an interpretation" (Brown & Yule, 1983: 223), though "there are discursive constraints on the roles that women characters are supposed to play in texts" (Mills, 1995: 170). Such readings of *Shakespearean* plays are wholly dependent upon the sociocultural context of Elizabethan England, rather than addressing linguistic issues ingrained in texts.

Attempting to fill a stylistic void grounded in linguistic theory, I construct two databases which procure grammatical structures of gender-specific terms from *Macbeth* and *The Merry Wives of Windsor*; commenting on the unreliability of grammatical structures to illustrate specific relationships using semantic relationships as more accurate criteria for determining object-relationships. In this paper, I address grammatical and semantic relationships of *gender-specific lexical items* within *Macbeth* and *Merry Wives of Windsor* to prove that the textual representation of *gender* is encoded by the language used: through the act of building these databases manually, I begin to address the predictability of proportional gender representation.

To observe gendered terms in the two plays, I adapt methodologies of *corpus stylistics* (Starcke, 2010: 55), analyzing each text for proportional gender representation. Allowing a specific play to be a sample population (Biber, 1993; Sinclair, 1991), and testing for patterns of gender representation in texts using anticipated collocates and grammatical structures while remaining conscious of priming for specific results using predetermined keywords, new patterns of gendering in Early Modern drama can be found. This study was done manually, using the text editor AlphaX and Microsoft Excel.

Unlike Culpeper 2002 and Scott 1999, who define keywords based on statistical frequency in a text using "evidence of the lexical organization of their language […] by studying patterns of significant collocation" (Sinclair, 1970: 77), which is strictly a corpus linguistics approach, I selected my search terms on the basis that they encompass grammatical gender through inclusion of conceptualized male and female features in English using the sex and gender hypothesis (Vigliocco, Vinson, Paganelli, & Dworzynski, 2005; Sinclair, 1991). While form does not necessarily correlate with gender, these will illustrate the differences between grammatical and natural gender. Although English is not inflected for gender, grammatical gender is still present.

This paper is part of a larger study, wherein I address many more gender-specific search terms from *Macbeth* and *The Merry Wives of Windsor*. Here I will be only addressing *he/she, him/her* and *his/hers*. These gender-specific terms are not consistently in proportion to character populations of a play, contrary to our expectations.

Macbeth has three and a half times more male than female characters.

MALE	29/39 characters (74%)
FEMALE	7/39 characters (18%)
NOT SPECIFIED	3/39 characters (8%)

Figure 1. Characters by Gender in Macbeth

With more male characters, masculine pronouns are expected to appear with high frequencies, whereas feminine pronouns appear less frequently. *The Merry Wives of Windsor* has a similar character distribution:

MALE	22/26 Characters (85%)
FEMALE	4/26 characters (15%)
NOT SPECIFIED	0/26 characters (0%)

Figure 2. Characters by Gender in Wives

Both plays have similar distributions: the majority of characters in each play are men, though there are fewer women in *Wives* than there are in *Macbeth*.

Dramatis personae show that male characters have significantly larger speaking roles. The distribution of characters is generally consistent across both plays:

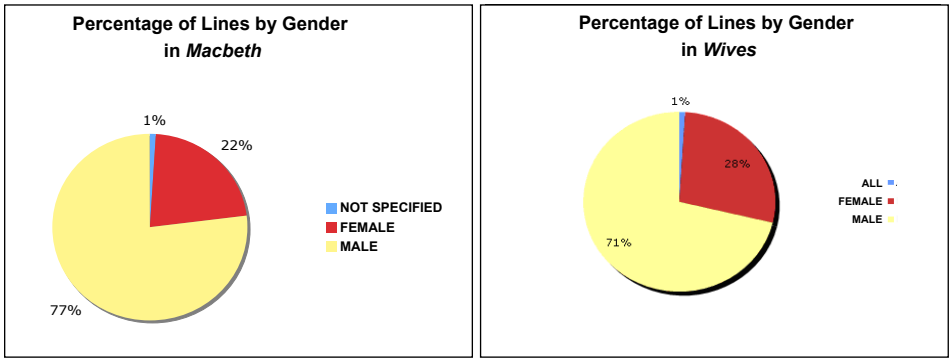


Figure 3. A comparison of lines spoken (percentage) by each gender in each play

A minority of female characters means that prototypically male pronouns will appear with a much higher overall token frequency, and are much more likely to be said by male characters. Gender-specific terms appear in a proportional relationship to the gender of the characters: the more male characters in a play, the more likely masculine pronouns are to appear with higher frequencies; similarly, fewer women in the dramatis personae means feminine pronouns are less likely to appear.[12] Although *Wives* is a play with proportionally few female characters, the action centers around women; they talk about each other much

12 However, this does not necessarily preclude that men will talk exclusively about each other; for example, see *Antony and Cleopatra*.

more than women of *Macbeth* will –women in *Macbeth* barely interact with each other. Linguistic representation of gender is dependent upon the roles of women and men in the play itself, thus making *Macbeth* and *The Merry Wives of Windsor* ideal candidates for such a study.

Through the compilation of two pilot databases, patterns of grammatical structures begin emerging. Because *Macbeth* is written almost entirely in verse and *Wives* is almost entirely in prose, working within clauses avoids some difficulties with line inconsistencies. These texts can be regularized by using criteria of subject/object relationships within a phrase structure, rather than within lines or sentences.[13] Literary language does not function in the same way as natural language does, which is a formal linguistic constraint of working with *Shakespearean* writing; both verse and prose appear as literary language, but neither are entirely reflective of natural Early Modern language.

A typical English sentence follows Subject/Verb/Object structure, but permutations of this structure are common in verse; regardless of grammatical organization of a sentence, subjects are obligatory in English, whereas predicates are not.[14] Third-person pronouns are organized by case and conceptualized gender. Deviation from grammatically assigned case is not allowed, retaining specific lexical properties: nominatives are subjects, whereas accusatives, datives, and genitives are objects. Syntactic relationships can change within poetic language, as the rules of syntax are often broken in poetic verse; regardless of syntactic organization, agent/patient relationships remain static.

Encoded semantic relationships are less predictable: syntactic relationships within poetic language are irregular, but agent/patient relationships remain static. Feminist literary critics often assume female characters will always appear as objects rather than subjects by conflating grammatical and literary content-domains into real-world implications. Understanding literature as reflective of realities of the period wrongly conflates literary and non-literary contexts as equivalent while presupposing specific structures. Semantic roles, following Carnie, 2007 and Dowty, 1991 are more nuanced: though subject/object relationships will not directly map onto social roles as these literary critics would like, semantic roles might.

Relationships between characters and gendered language is wholly dynamic, requiring pragmatic and conceptual processing of gender as a biological and grammatical feature, which are not necessarily connected in a logical way – an issue that Vigliocco et al address in their "sex and gender hypothesis" (2005: 502). "Gender-specific terms", as I call them, are conceptualized as gendered in English rather than being inherently male or female - we see that pronouns gain the binary conceptual feature [±male] or [±female] because we have indexed them with someone or something that is assigned a binary sex.

The conceptualization of gender is relevant to both literary and natural language: terms selected are pronouns that are specifically related to gender identity. The terms selected

13 Quotations included in this paper utilize a variation on the through-line numbering system. Please see Open Source Shakespeare (http://www.opensourceshakespeare.org/info/technicaldetails.php) for more about TL

14 For example: "I swam" is a grammatical sentence, whereas *"Swam the ocean" is not. Subjects can be null but implied, as seen in imperative sentences, such as "Go home".

of interest are therefore prototypically "male" or "female" in some sense (Murphy, 2002) – they are categorized as having the binary categorical feature [± male] or [± female]. The differentiation between sex and gender functions internally and externally within the text itself: literary language depends wholly upon the adoption of real-world implications into the "content domain" of a specific text (Vigliocco et al, 2005); thus the real-world implications of sex and gender are considered when discussing gender in plays.

Subjects and objects are predictable.

Nominative	Accusative	Dative	Genitive
He	Him	Him	His
She	Her	Her	Hers

Figure 4. The third-person gender-specific pronouns

Pronouns in the nominative case appear as the subject of a clause. There are two kinds of *his* and *her* which could be used in these examples: the determiner *his/her*, part of a D+N construction, and the independent pronoun *him/her*. The D+N construction has two features; representing possession and inanimate ownership. However, this pair does not map perfectly: there is also the relationship of *his*/*hers*, the grammatical possessive[15]. This claim seems to be contradicted by the presence of *her+NP;* however, these forms must be counted and addressed separately.

ANALYSIS

He in *Macbeth* f = 108 (86%)	*He* in *Wives* f = 184 (74%)
She in *Macbeth* f = 17 (14%)	*She* in *Wives* f = 64 (26%)
Where f = relative frequency of search term *X* in each text	

Figure 5. Nominative pronouns.

This chart appears to support feminist theory; men appear much more frequently as nominatives than women do. Character proportions of *Macbeth* match usage proportions for third-person nominative pronouns; though *Wives* produces many more examples of *she* than proportions would imply. Despite there being are only 4 female characters in *Wives*, feminine nominatives appear with a much higher frequency than in *Macbeth*. Nominative pronouns appear as subjects, but through thematic analysis, subject positions do not necessitate agency. These patterns are consistent with expectations – every example in the nominative case from both plays are subjects and appear to prime for agentivity. A nominative pronoun in the subject position can function as a patient rather than agent:

15 *Hers* does not appear in either text, therefore I cannot comment on the relationship between *his* and *hers*

(1) "What a taking was *he* in when your husband" (Wives 1650); *he* as subject and source

(2) "Heaven knows what *she* has known" (Macbeth 2385); *she* as subject and experiencer

Nominative pronouns are always subjects, but semantic relationships are much more dynamic.

There is an uncharacteristic shift towards *her* in *Wives*, which does not follow our expectations.

His in *Macbeth* f = 141 (54% of examples)	*His* in *Wives* f = 117 (29%)
Him in *Macbeth* f = 85 (33% of examples)	*Him* in *Wives* f = 134 (34%)
Her in *Macbeth* f = 35 (13% of examples)	*Her* in *Wives* f = 146 (37%)
Where f = relative frequency of search term *X* in each text	

Figure 6. Non-nominative pronouns.

Although there are only four women in *Wives*, they speak substantially more than the women of *Macbeth*. *Her* has a very high frequency in *Wives*, despite evidence implying that *him* and *his* would have higher frequencies. *Wives* dispels the theory that women are insignificant in the Early Modern period: the disproportionate representation of *her* to the number of women does not account for this skewing. These proportions should follow lines by gender, as they do in *Macbeth*: if women are main characters of a play, they have more opportunities for prototypically female lexical items to appear - perhaps if women in *Macbeth* were given more lines, they would have the opportunity to use feminine lexical items more frequently.

Despite expectations of subject-as-agent and object-as-theme, this does not appear consistently. In (4) and (5), *his* functions as the agent, but in (6) *his* is the theme. Some difficulties in analysis of *his/ her* appear when D+N makes a nominative phrase, representing a single semantic unit, and *his*/*her* also function as a single semantic unit:

(3) "*Her mother* hath commanded *her* to slip" (*Wives*, 2583); *her mother* as the agent and *her* as experiencer

(4) "*His secret murderers* sticking *on his hands*" (*Macbeth*, 2446); *his secret murderers* are both subject and agent; *on his hands* is the object (part of a PP) but semantically the location

(5) "All *his successors* gone before him hath" (*Wives*, 19); *his successors* are the experiencers

(6) "Who wear our health but sickly in *his life*" (*Macbeth*, 1240); *his life* is the theme

While *his*/*hers* have predetermined grammatical roles, semantic roles are derived independently. They are proforms, functionally different from D+N constructions; here are some proforms carrying their own semantic roles within their own phrases and within the larger context of a clause:

(5) "All *his successors* gone before him hath" (*Wives*, 19); *his successors* are the experiencer and the subject

(6) "Who wear our health but sickly *in his life*" (*Macbeth*, 1240); *his life* is the theme and a PP

(7) "I come to speak *to her*" (*Wives*, 2436), *to her* is a PP and the experiencer

(8) "Why it stood *by her*" (*Macbeth*, 2363); *by her* is a PP and the experiencer

(9) "How might we disguise *him*?" (*Wives*, 2141); *him* is the experiencer and DO

(10) "Let *him* not strike the old woman" (*Wives* 2251); *him* is the agent and DO

(11) "I will drain *him* dry as hay" (*Macbeth*, 142); *him* is the goal and DO

(12) "I grant *him* bloody" (*Macbeth*, 2103); *him* is the theme and DO

Men might be more likely to appear in Early Modern plays as characters and have significantly larger speaking roles; the number of male lexical items will be higher, as characters have to refer to each other with more frequency. Syntax appears to privilege masculine lexical items over feminine lexical items; a sense of patriarchic entitlement towards the male characters as subjects of both plays is expected, but a semantic study illustrates that grammatical and thematic relationships are independent of each other: male lexical items in subject positions do not necessitate semantic agency; feminine lexical items in grammatical object positions do not necessarily have to be the patient of the clause. Gender does not implicate a subject or object position, as feminist readings often attempt to prove.

Feminist readings stress the role of women as objects; syntactically, this seems true, especially in *Macbeth*, but this is an unreliable method for observing gender. Women, socially disadvantaged in the Early Modern period (Mendelson & Crawford, 1988) would be unlikely to have ownership of objects, yet examples in the texts disprove this statement. Relationships in *Wives* contradict women's insignificance in this period: disproportionate representations of *her* in *Wives* do not account for this. If women are main characters of a play, there will be more opportunities for prototypically female lexical items to appear – the larger a character's speaking role, the more likely their gender-specific lexical items will appear at a higher frequency.

Patterns of subject/object relationships are identifiable, but do not necessarily entail the rampant sexism implied by feminist critics. The relationship between grammatical and semantic roles are encoded and manifest themselves into a literary representation of gender: the textual representation of gender is encoded by the language used, but do not necessarily entail the objectification of women as previously identified by literary critics.

Bibliography

AlphaX, freeware available at http://alphatcl.sourceforge.net/wiki/pmwiki.php/Software/AlphaX

Bakhtin, M. M. (2008) *The Dialogic Imagination: Four Essays*. Trans. Michael Holquist. Austin: University of Texas.

Biber, Douglas (1993). "Co-occurrence Patterns among Collocations: A Tool for Corpus-Based Lexical Knowledge Acquisition." Computational Linguistics 19: 531–538.

Black, M., & Coward, R. (1998). "Linguistic, Social, and Sexual Relations: A Review of Dale Spender's *Man Made Language*." In Deborah Cameron, *The Feminist Critique of Language: a Reader*. 100-18. London: Routledge.

Brown, G., and Yule, G. (1983) *Discourse Analysis*. Cambridge: Cambridge UP, 1983.

Cameron, Deborah. (1998). *The Feminist Critique of Language: a Reader*. (2nd ed). Cambridge: Cambridge UP

Carnie, Andrew (2007). *Syntax: a Generative Introduction*. (2nd ed). Oxford: Blackwell Publishing.

Culpeper, Jonathan (2001) *Language and Characterisation: People in Plays and Other Texts*.

Culpeper, Jonathan (2002). "Computers, Language and Characterisation: An Analysis of Six Characters in *Romeo and Juliet*". In Ulla Melander- Marttala, Carin Östman and Merja Kytö. (Eds), *Conversation in Life and in Literature: Papers from the ASLA Symposium,*

Association Suedoise de Linguistique Appliquee (ASLA), 15. Universitetstryckeriet: Uppsala. 11- 30.

Dowty, David. (1991) "Thematic Proto-Roles and Argument Selection." *Language* 67(3). 547-619.

Fabb, Nigel. (2010) "Is Literary Language a Development of Ordinary Language?" *Lingua* 120(5): 1219-232.

FISCHER-STARCKE, BETTINA. (2010). *Corpus Linguistics and the Study of Literature: Jane Austen and Her Contemporaries*. London: Continuum.

GIVON, T., & GIVON, R. (2002). *Syntax: A Functional-Typological Introduction*. Amsterdam: John Benjamins.

PARKER, PATRICIA. (1996). *Shakespeare from the Margins: Language, Culture, Context*. Chicago: University of Chicago

LEVIN, RICHARD. (1998) "Feminist Thematics and *Shakespearean* Tragedy." *PMLA* 103(2): 125-38.

HOWARD, J. E., AND RACKIN, P. (1997) *Engendering a Nation: a Feminist Account of Shakespeare's English Histories*. London: Routledge.

HUDDLESTON, RODNEY D. (1984). *Introduction to the Grammar of English*. Cambridge: Cambridge UP

HUNSTON, S., & FRANCIS, G. (2000) *Pattern grammar: A corpus-driven approach to the lexical grammar of English.* Amsterdam: John Benjamins.

IRIGARAY, LUCE. (1998) "Linguistic Sexes and Genders." *The Feminist Critique of Language: a Reader*. By Deborah Cameron. 119-123. London: Routledge

MCEACHERN, CLAIRE. (1988) "Fathering Herself: A Source Study of Shakespeare's Feminism." *Shakespeare Quarterly* 39(3), 269-90.

MENDELSON, S. H., & CRAWFORD, P. (1998) *Women in Early Modern England, 1550-1720*. Oxford: Clarendon

MILLS, SARA. (1995) *Feminist Stylistics*. London: Routledge.

MURPHY, GREGORY L. (2004) *The Big Book of Concepts*. Cambridge, MA: MIT.

NEVALAINEN, TERTTU. (2006) *An Introduction to Early Modern English*. New York: Oxford UP

SINCLAIR, JOHN M. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford UP.

SCOTT, M. (2008) WordSmith Tools version 5, Liverpool: Lexical Analysis Software.

THOMS, GARY STEWART. (2010). "Poetic Language: A Minimalist Theory". PhD Dissertation. University of Strathclyde, Glasgow.

VIGLIOCCO, G., VINSON, D., PAGANELLI, F., & DWORZYNSKI, K. (2005). "Grammatical gender effects on cognition: Implications for language learning and language use". *Journal of Experimental Psychology: General 134*(4), 501-520.

The Corpus of Greek Aphasic Speech: Design and compilation

Dionysis Goutsos[1], Constantin Potagas[2], Dimitris Kasselimis[2 & 3],

Maria Varkanitsa[1] & Ioannis Evdokimidis[2]

[1]*Department of Linguistics, School of Philosophy, University of Athens*

[2]*Department of Neurology, Medical School, University of Athens, Aeginition Hospital*

[3]*Psychology Department, School of Social Sciences, University of Crete*

The paper presents the design and compilation of the Corpus of Greek Aphasic Speech, a new resource for the study of aphasia in Greek, and discusses its possible applications. The aims and design of the corpus and the methods followed for its compilation are presented. A pilot corpus, including two texts (spontaneous speech and picture description) from the spoken output of 20 patients was first created. On the basis of this, a classification of paraphasias or speech errors has been attempted and some preliminary findings have been gathered, while the target corpus is planned to include texts from 120 patients of about 50.000 words. It is argued that computer language corpora can offer a new perspective to the linguistic study of aphasia, especially in Greek, by providing systematic evidence for patients' linguistic profiles, drawn from language use rather than from isolated examples of lack of competence.

Keywords: annotation, aphasia, speech errors; Greek

En este artículo se presenta el diseño y la compilación del Corpus del Discurso Afásico Griego, un nuevo recurso para el estudio de la afasia en griego, y se analizan sus posibles aplicaciones. Los objetivos y el diseño del corpus y los métodos seguidos para su elaboración se presentan. Un corpus piloto, incluyendo dos textos (el habla espontánea y la descripción de imagen) del discurso de 20 pacientes fue creada. Sobre la base de esto, una clasificación de las parafasias o errores del habla se ha intentado y algunos resultados preliminares han sido recogidos, mientras que el corpus de destino está previsto incluir textos de 120 pacientes alrededor de 50.000 palabras. Se argumenta que un corpus puede ofrecer una nueva perspectiva para el estudio lingüístico de la afasia, especialmente en griego, proporcionando evidencia sistemática de los perfiles lingüísticos de los pacientes, basado en el uso del lenguaje en lugar de ejemplos aislados de falta de competencia.

Palabras clave: anotación, afasia, los errores del habla; griego

1. Corpora and aphasia

As late as in 1990 Menn & Obler remark that "there is absolutely no tradition of publication or even archiving of aphasic texts" (1990: 12). Although the collection of data from the speech of aphasic patients has had a long history in the study of aphasia, it is true that even today only a few studies of aphasia have taken full advantage of corpus methodology. Thus, most studies use databases of individual words or sentences rather than extended talk or use corpora in order to simply draw illustrative examples rather than as data to be exhaustively described. In addition, as Perkins has rightly pointed out, "most corpora of disordered language remain very small […] hard to access […] and – above all – not in machine-readable format" (1995: 129). Notable exceptions for adult aphasic discourse include Perkins & Varley (1996) for English, Gallardo & Moreno (2005) for Spanish and Westerhout & Monachesi (2007) for Dutch. Still, wide availability to data is only found in projects like the Aphasia TalkBank, which aims at the creation of an enormous multimedia database of aphasic speech in English and other languages (MacWhinney, 2007).

There are several reasons for this rather curious absence of corpus methodology from the study of aphasia. As Edwards finds, "most models applied to aphasia have been concerned with single-word processing" (2005: 23) rather than the analysis of extended talk. In fact, single-case studies predominate in the field, something which concurs with an almost exclusive focus on competence aspects of aphasic speech production and comprehension, tested through completion tasks, grammaticality judgements etc. In other words, the interest has been on what the aphasic patients are able (or, mainly, not able) to say rather than on what they have actually said in specific contexts. Without doubt, the predominance for several decades of competence-based models in linguistics, following the dominant generative research paradigm, has significantly weighted the scales against empirical investigations of aphasic speech performance.

This bias has particularly influenced the study of aphasia in Greek, which, as a highly inflected language, offers itself for the study of grammatical phenomena. This probably explains why most of the studies on Greek aphasia concern aspects of verb or noun morphology at the expense of the analysis of other linguistic levels. Moreover, these studies are exclusively based on competence testing rather than other evidence. Most importantly, there has been no attempt for an overall description of aphasia in Greek; instead, existing studies fall in the mainstream tradition of testing isolated phenomena, aiming at validating or disqualifying theoretical points.[16]

The advantages of using electronic corpora in the study of aphasia are quite obvious. First of all, corpora allow researchers to study large amounts of occurring data, by focusing on actual language use rather than linguistic competence. This is necessary if our goal is a comprehensive typology of aphasia based on linguistic principles (Crystal 2002). The collection of the spoken output of aphasic patients into a corpus enables us to treat linguistic deviations as instances of aphasic discourse, rather than as accidental, isolated

16 See Goutsos et al. (2011) for a detailed discussion of the literature on Greek.

errors and thus to account for their function in the patient's linguistic system. In addition, as both Perkins (1995) and Crystal (2002) emphasize, the occurrence of linguistic patterns may enable us to group patients into a number of linguistically defined diagnostic types, with a view to helping diagnosis, assessment and intervention. Furthermore, data from corpora of aphasic discourse can be used to compare with linguistic patterns in general reference corpora. In this way, we can assess the degree to which aphasic data diverge from other kinds of data on a firm empirical basis, rather than by relying on intuition. Finally, corpora enable easy accessibility to the original data and their transcription and/or annotation. They thus enhance verifiability of research, make possible cross-linguistic investigations and allow different kinds of researchers to study the same set of data. As hinted at above, there is an additional advantage from the use of aphasic corpora in the case of Greek, in which large-scale empirical findings on aphasia are still missing.

2. The Corpus of Greek Aphasic Speech

The Corpus of Greek Aphasic Speech (CGAS) is a development of a common project of the Department of Linguistics and the Department of Neurology (Aeginition Hospital) of the University of Athens. In the following sections we present the project's aims and describe the corpus composition and the stages of its compilation.

2.1. Corpus aims and composition

In designing the CGAS we have taken into consideration the relative absence of specialized corpora of aphasic discourse both in Greek and other languages, as well as the problems pointed out above the study of Greek aphasia. In addition, general principles of corpus design such as Sinclair's (2005) have been followed. In particular, the Corpus of Greek Aphasic Speech (CGAS) was compiled with the following aims:

a) to contain whole texts produced by a significant number of Greek aphasic patients,

b) to include texts from a variety of contexts,

c) to relate linguistic data to extra-linguistic metadata in a systematic way, and

d) to make data on Greek aphasic discourse available to the research community.

In particular, CGAS includes in its pilot phase data from 20 patients, who were treated at the Aeginition Hospital between 2006 and 2008. Two type texts from each patient's spoken output are included, namely spontaneous speech and picture description. In other words, the corpus includes 40 texts, two from each participant. Both text types were produced in the situation of doctor-patient interviews and, in this sense, they are not, strictly speaking, conversational, but rather guided monologues. In total, 12,663 words are included in the Corpus at present, of which 10,332 come from the patients' discourse.

Our plans for the development of CGAS aim at 120 patients, that is 240 texts from the same two text types. This will constitute a corpus of approximately 50,000 words, of which roughly 41,000 will come from the patients' speech. Table 1 below summarizes this information.

Table 1. Structure of the CGAS.

	Number of patients	*Text types*	*Number of texts*	*Word number*
Pilot corpus	20	2	40	12,663 (10,332)
Target corpus	120	2	240	~50.000 (41.000)

2.2. Corpus compilation

The stages of corpus compilation involved data and metadata collection, transcription and annotation, including marking for speech errors. The speakers were recruited from a large pool of patients with stroke treated at the Aeginition Hospital. Aphasics were identified through language assessment with the Boston Diagnostic Aphasia Examination–Short Form (BDAE-SF), adapted for Greek (Tsapkini et al. 2009). CT and/or MRI scans were obtained for each patient and lesion sites were identified by two independent neuro-radiologists.

The data were collected in typical doctor and patient interactions, in which psychologists interviewed patients with regard to what happened to them ('stroke stories') and, on another occasion, administered the description of the Cookie Theft picture from the Boston Diagnostic Aphasia Examination (Goodglass & Kaplan, 1983). Sessions were audio-recorded with either a tape-recorder or a digital voice recorder in a quiet setting. We have excluded data from patients who did not perform in either the spontaneous speech or the picture description task and from those who suffered severe fluency impairment, so that they could not produce any recognizable words.

All collected material was orthographically transcribed in a first transcript and then checked for accuracy by two different transcribers. (For reasons of transparency, an orthographic rather than phonetic transcription was adopted). However, fluency problems, voiced and unvoiced starters and fillers, repetitions and other phenomena of spoken interaction such as noise from the outside, coughing etc. were carefully noted, following conventions for spoken data transcription (Georgakopoulou & Goutsos, 2004: vii).

2.3. Corpus annotation

In the first phase of annotation, already transcribed data have been tagged for speech errors or paraphasias. These errors have also been tagged for part of speech. The second phase of annotation includes the extension of the part of speech tagging to the entire pilot

corpus, whereas at a later stage intonation contours will also be marked throughout. In parallel, the spoken data transcription is linked to the sound archives of the recordings, with a view to developing a multi-modal corpus that will allow flexibility in the analysis, as well as access to the representation of several layers of data.

Because of the gap in the comprehensive description of aphasia in Greek, it has been necessary to identify paraphasic errors at many levels of linguistic analysis, as has already been the case in the relevant bibliography (e.g. Nespoulous & Roch-Lecours, 1984; Ahlsén, 2006: 56-57; Ingram, 2007: 23; Turgeon & Macoir, 2008). Thus, the following categories and sub-categories of errors were identified: a) phonological errors, b) morphosyntactic errors, c) lexical errors, d) neologisms and e) periphrasis. Appendix 1 presents our detailed classification, along with the criteria used for each category and an indicative example from the data.

It must be noted here that there may be some overlap between the categories identified, due to the inevitable process of interpretation involved in the tagging for speech errors. For cases in which two categories could be attributed to the same error, it was decided to mark both, while a similar practice was followed in cases where two different errors were found to occur.

2.4. Corpus availability

At present, the annotated version of the pilot CGAS is available through the Aphasia Talkbank project.[17] For this release, data have been transcribed according to CLAN conventions (MacWhinney, 1996). Since this is a work in progress, we are currently using our experience from the pilot corpus to build the target corpus. Our plan is to make the target CGAS available to the research community via a dedicated webpage interface. Depending on ethical considerations (personal data restrictions), recordings will also be made available along with their transcription, so as to allow access to the primary material.

3. Preliminary findings and implications

Our preliminary analysis of the pilot CGAS suggests that the corpus can be immensely helpful in the study of Greek aphasia. First of all, a new series of questions hitherto unexplored in the literature can now be raised with reference to authentic data of aphasic speech. In particular, information can be adduced on the frequency and types of phonological and lexical errors in Greek, including neologisms and other semantically-related errors. Linguistic phenomena such as periphrasis, which is indicative of the speaker's linguistic strategies, can now be placed alongside traditional types of paraphasias. In addition, the corpus can offer specific details about the particular errors occurring in Greek aphasic discourse. For instance, phonological omission errors seem to mainly concern consonant clusters in nouns, involving phonemes such as /r/, /s/, /a/ and /n/. These details are crucial,

17 For more details, see the project's site: http://talkbank.org/AphasiaBank

not only for a fully-fledged analysis of Greek aphasia, but also for orienting clinical and post-clinical intervention towards concrete findings.

Furthermore, a simple comparison of the ten most frequent words in the CGAS text types and related text types in a reference corpus of Greek such as the Corpus of Greek Texts (CGT, see Goutsos, 2010) can also be illuminating.

Table 2. Most frequent tokens in CGAS and CGT text types.

CGAS: interviews	**CGAS: picture description**	**CGT: spoken data**	**CGT: interviews**
και 'and'	*το* 'the'-NEUT	*και* 'and'	*και* 'and'
το 'the'-NEUT	*να* 'to'	*το* 'the'-NEUT	*να* 'to'
ε 'eh'	*είναι* 'is'	*να* 'to'	*το* 'the'-NEUT
να 'to'	*εδώ* 'here'	*ναι* 'yes'	*την* 'the'-FEM
μου 'my'	*ε* 'eh'	*είναι* 'is'	*είναι* 'is'
δεν 'not'	*αυτό* 'this'	*δεν* 'not'	*που* COMP
στο 'at'	*δεν* 'not'	*που* COMP	*ότι* COMP
με 'me'	*και* 'and'	*θα* 'will'	*η* 'the'-FEM
αυτό 'this'	*τα* 'the'-NEUT-PL	*τα* 'the'-NEUT-PL	*του* 'the'-GEN
ναι 'yes'	*τι* 'what'	*μου* 'my'	*της* 'the'-FEM

As can be seen in Table 2, aphasic data diverge from the reference corpus in interesting ways: hesitation ('eh') and vague words ('this', 'what') are more frequent, whereas the conjunction *και* ('and'), which is systematically the most frequent item in most text types in Greek, is very low in frequency in the picture description data. Furthermore, complementizers like *που* and *ότι* seem to be also much less frequent in CGAS.

A final remarkable aspect of aphasic speech concerns the use of word clusters or lexical bundles (Biber et al., 1999). In CGAS the most frequent clusters include phrases such as *δεν μπορώ/μπορούσα να το πω/να καταλάβω* 'I cannot/could not say/understand it', *πώς να το πω/τι να πω* 'how to say it/what can I say', *πρέπει να είναι* 'it must be', *αυτά εδώ πέρα/αυτό το πράγμα* 'these things/this thing over here'. These clusters are indicative of the discourse strategies followed by aphasic speakers (e.g. avoidance, modality, periphrasis) and can offer a first glimpse at formulaic language, which may be processed in different ways than the rest of the vocabulary in aphasia (Wray, 2002).

In conclusion, the development of the Corpus of Greek Aphasic Speech puts a much needed emphasis on spontaneously produced data and the analysis of speech errors in their discourse context. It thus allows assessing paraphasic errors as the product of situated language use by specific speakers rather than as isolated examples of lack of competence. Ease of access to both the original data and the levels of transcription, annotation etc., are also expected to significantly contribute to the understanding of aphasia in Greek and to enhance our knowledge of the field.

Appendix 1

The following categories of speech errors have been distinguished in the annotation of the pilot corpus:

1. Phonological paraphasias: errors affecting isolated phonemes or syllables.

   - PH1: phoneme deletion/omission: *άντας* [‘adas], instead of *άντρας* [‘adras] ‘man’

   - PH2: phoneme addition: *αχαρτί* [aχa’rti], instead of *χαρτί* [χa’rti] ‘paper’

   - PH3: phoneme substitution: *γρυκά* [γri’ka], instead of *γλυκά* [γli’ka] ‘sweets’

   - PH4: syllabic: *σκαμπόβο* [ska’bovo], instead of *σκαμπό* [ska’bo] ‘stool’

2. Morphosyntactic paraphasias: errors affecting grammatical morphemes.

   - MS1: morpheme deletion/omission: *αυτό άντρα* ‘this man’, instead of *αυτός είναι άντρας* ‘this is a man’

   - MS2: morpheme addition: not found in the data

   - MS3: morpheme substitution: substitution (general): *για να πέσει κάτω*, instead of *θα πέσει κάτω* ‘he will fall down’ [the complementizer *για να* substitutes *θα*]

   - MS4: morpheme substitution: aspect: *δεν είδε καλά το μάτι*, instead of *δεν έβλεπε καλά το μάτι* ‘the eye could not see’ [synoptic/perfect in the place of continuous/ imperfect stem]

   - MS5: morpheme substitution: tense: *το λόγο που έχω πριν*, instead of *το λόγο που είχα πριν* ‘the speech I had before’ [present in the place of past stem]

   - MS6: morpheme substitution: agreement: *ένα κυρία* ‘a.NEUT lady’, instead of *μια κυρία* ‘a.FEM lady’

   - MS7: other: *δουλεύω κάτι* ‘I work something’, instead of *δουλεύω σε κάτι* ‘I work in something’

3. Lexical paraphasias: errors affecting whole words, particularly, substitution of a word by another pre-existing similar or non-similar word.

   - L1: formal: words related by formal similarity: *πλακάκι* [pla'kaci] 'tile', instead of *νεράκι* [ne'raci] 'some water'

   - L2: verbal: meaning similarity: *άνθρωπος* 'person', instead of *παιδί* 'child'

   - L3: unrelated: no similarity: *νούμερα* ['numera] 'numbers', instead of *μπισκότα* [bi'skota] 'biscuits'

4. Neologisms: errors affecting whole words (more than 50% of the word form): substitution of a word by another similar or non-similar word, not occurring in Greek.

   - N1: possible but non-existent words of Greek, classifiable to a part of speech: *γερεβύτης* [jere'vitis], instead of *νεροχύτης* [nero'çitis] 'basin'

   - N2: non-recognizable words, non-classifiable according to grammatical category: *πενιχθεσινίδις* [peniχθesin'iδis]

5. Periphrasis: errors affecting whole words: substitution of a word by an extended phrase.

   - P1: circumlocution: the extended phrase refers periphrastically to a word: *αυτό που έχει το νερό* 'this which has the water', instead of *βρύση* 'tap'

   - P2: vagueness: the extended phrase avoids specific reference to a word: *έπαθα μια αυτή* 'I had a this'

References

Ahlsén, E. (2006). *Introduction to Neurolinguistics*. Amsterdam/Philadelphia: John Benjamins.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

Crystal, D. (2002). Clinical linguistics. In M. Aronoff & J. Rees-Miller (Eds), *The Handbook of Linguistics* (pp. 673-682). Oxford: Blackwell.

Edwards, S. (2005). *Fluent Aphasia*. Cambridge: Cambridge University Press.

Gallardo, B. & Moreno, V. (2005). *Afasia no fluente*. Valencia: Guada Impresores.

Georgakopoulou, A. & Goutsos, D. (2004). *Discourse Analysis. An Introduction*. (2nd ed.). Edinburgh: Edinburgh University Press.

GOODGLASS, H. & KAPLAN, E. (1983). *The Assessment of Aphasia and Related Disorders*. (2nd ed.). Philadelphia: Lea & Febiger.

GOUTSOS, D. 2010. The Corpus of Greek Texts: A reference corpus for Modern Greek. *Corpora, 5* (1), 29-44.

GOUTSOS, D., POTAGAS, C., KASSELIMIS, D., VARKANITSA, M. & EVDOKIMIDIS, I. (2011). Studying paraphasias in the Corpus of Greek Aphasic Speech. In C. Potagas & I. Evdokimidis (Eds.), *Discourse and Memory* (pp. 23-47). Athens: Synapses. [In Greek]

INGRAM, J. C. L. (2007). *Neurolinguistics. An Introduction to Spoken Language Processing and its Disorders*. Cambridge: Cambridge University Press.

MACWHINNEY, B. (1996). The CHILDES system. *American Journal of Speech Language Pathology, 5*, 5-14.

MACWHINNEY, B. (2007). The TalkBank project. In J. C. Beal, K. P. Corrigan & H. L. Moisl (Eds.), *Creating and Digitizing Language Corpora* (pp. 163-180). London: Palgrave Macmillan.

MENN, L. & OBLER, L. K. (1990). Theoretical motivations for the cross-language study of agrammatism. In L. Menn & L. K. Obler (Eds.), *Agrammatic Aphasia: A Cross-language Narrative Sourcebook* (pp. 3-12). Amsterdam: John Benjamins. Vol. 1.

NESPOULOUS, J.-L. & ROCH LECOURS, A. (1984). Clinical descriptions of aphasia: Linguistic aspects. In D. Caplan, A. Roch Lecours & A. Smith (Eds.), *Biological Perspectives on Language* (pp. 141-157). Cambridge, Mass: MIT Press.

PERKINS, M. (1995). Corpora of disordered spoken language. In G. Leech, G. Myers & J. Thomas (Eds.), *Spoken English on Computer. Transcription, Mark-up and Application* (pp. 128-134). London: Longman.

PERKINS, M. R. & VARLEY, R. (1996). *A Machine-Readable Corpus of Aphasic Discourse*. University of Sheffield: Department of Human Communication Sciences/Institute for Language, Speech and Hearing.

SINCLAIR, J. (2005). Corpus and text-basic principles. In M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 1–16). Oxford: Oxbow Books.

TSAPKINI, K., VLAHOU, C. H. & POTAGAS, C. (2009). Adaptation and validation of standardized aphasia tests in different languages: Lessons from the Boston Diagnostic Aphasia Examination. *Behavioural Neurology*, *22* (3), 111-119.

TURGEON, Y. & MACOIR, J. (2008). Classical and contemporary assessment of aphasia and acquired disorders of language. In B. Stemmer & H. A. Whitaker (Eds.), *Handbook of the Neuroscience of Language* (pp. 3-11). Amsterdam: Elsevier.

WESTERHOUT, E. & MONACHESI, P. (2007). A pilot study for a Corpus of Dutch Aphasic Speech (CoDAS). Retrieved from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.109.1882&rep=rep1&type=pdf.

WRAY, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Interaction of technology and methodology in building and sharing an annotated learner corpus of spoken German

Hanna Hedeland

This paper discusses the technological and methodological challenges in creating and sharing HAMATAC, the Hamburg Map Task Corpus. The first version of the corpus, consisting of 24 recordings with orthographic transcriptions and metadata, is publicly available. A second version featuring different types of linguistic annotation is in progress. I will describe how the various software tools and data formats of the EXMARaLDA system were used for transcription and multi-level annotation, to compile recordings and transcriptions into a corpus and manage metadata, to publish the corpus, and how they can be used for carrying out corpus queries (KWIC) and analyses. Some recurrent issues in corpus building and sharing and the interaction of technological and methodological aspects will be illustrated using HAMATAC.

Keywords: spoken language, learner corpus, transcription, multi-level annotation

Este artículo trata los retos tecnológicos y metodológicos de la creación y publicación de HAMATAC, el Hamburg Map Task Corpus. La primera versión del corpus, que consiste en 24 grabaciones con transcripción ortográfica y metadatos, está disponible públicamente. Está en desarrollo una segunda versión que incluye distintos tipos de anotación lingüística. Voy a describir cómo las diversas herramientas de software y formatos de datos del sistema EXMARaLDA se utilizaron para la transcripción y la anotación multinivel, para reunir grabaciones y transcripciones en un corpus, para administrar los metadatos y para publicar el corpus, y cómo pueden ser usados para realizar consultas en el corpus (KWIC) y análisis. Se ilustrarán usando HAMATAC algunos de los cuestiones recurrentes de la creación y publicación de un corpus y la interacción de los aspectos tecnológicos y metodológicos.

Palabras clave: lenguaje oral, corpus de aprendices, transcripción, anotación multinivel

1. Introduction

Over the last decades, technological advances have contributed to an increased use of spoken language corpora in linguistic research, mostly because recording of digital audio and even video using standardized formats has become much easier for non-specialists. These technological advances open up interesting opportunities when it comes to creating, managing, analyzing and sharing linguistic data, especially when considering recent global standards for structured digital multilingual textual data. In response to these opportunities, various data formats and software tools making use of these technological advances have been developed for corpus linguistic research.

This paper discusses the technology and methodology of corpus building and sharing. It is structured as follows: In section 2., the main characteristics of the HAMATAC corpus will be described; in section 3., the EXMARaLDA system used to create the corpus will be introduced; section 4. discusses some recurrent issues in spoken language corpus development and section 5. illustrates some cases of interaction of technology and methodology in a process such as the development of HAMATAC and provides some concluding remarks.

2. The Hamburg Map Task Corpus[18]

HAMATAC (Schmidt et al., 2010) consists of 24 recordings of advanced learners of German solving a map task in twelve pairs. A map task is a classic information gap task where one participant has a map with a path from start to goal and instructs the other participant, who has no path on his map, how to draw this path. The map task we used was originally designed to elicit native varieties of German for the project *Deutsch Heute* (German Today) (Brinckmann et al., 2008). HAMATAC can therefore

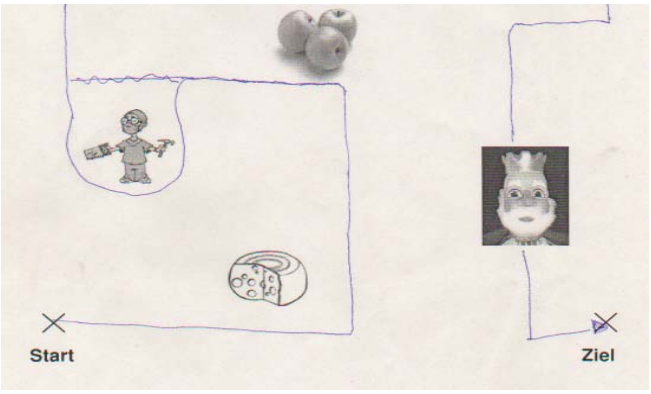


Figure 1. Part of a map with a path drawn from Start to Goal.

18 We are grateful to Kim-Chi Hamze, Seçil Yusun, Yael Dilger and Fideniz Ercan who supported us as student assistants in different stages of the corpus creation.

be said to complement the existing *Deutsch Heute* corpus with corresponding learner varieties. The speakers' mother tongues include Romance, Slavic and Persian languages, as well as languages from Non-Indo-European families. Such information on the speakers' language biographies is provided along with other relevant metadata on speakers and communications. The first corpus version, consisting of original recordings, orthographic transcriptions and metadata, is publicly available via a web interface with password-protected access. In total, the recordings amount to 3:17 hours and there are 21433 transcribed words.

The next version of HAMATAC will include annotations describing various linguistic levels and phenomena such as part of speech, lemmas, phonetic form and disfluencies. Whereas automatic methods can be applied to the first three, the disfluency annotation serves as an example of subjective phenomena that require annotation schemes with necessarily interpretative categories.

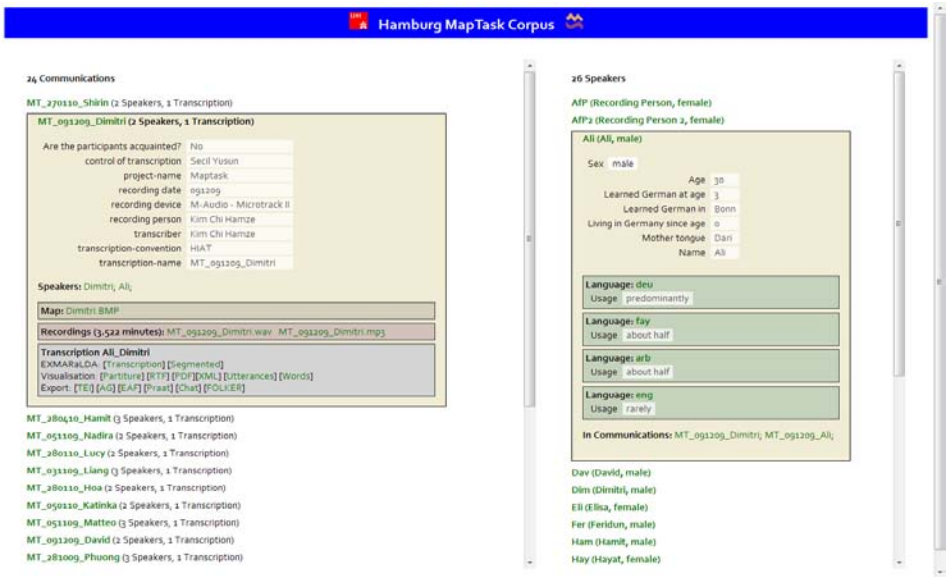


Figure 2. The corpus overview of the 24 communications and 26 speakers.

3. The EXMARaLDA System

The HAMATAC corpus was created with the EXMARaLDA system (Schmidt & Wörner, 2009) developed at the Research Centre on Multilingualism in Hamburg. The EXMARaLDA system consists of data models, formats and software tools for transcribing, annotating, managing and analyzing spoken language corpora. Below, I will demonstrate how the software components—transcription editor, stand-off annotation tool, corpus manager and search and analysis tool—were used in the corpus building process and how they can be used to analyze corpus data.

3.1. Transcription and annotation

The Partitur-Editor is a tool for transcription and multi-level annotation of digital audio or video recordings. It implements the abstract model of a single timeline with multiple tiers for transcription and annotation (cf. Schmidt, 2005). The timeline aligns segments within the transcription file as well as the transcription file with the recording. The Partitur-Editor has built-in support for some major transcription systems (e.g. CHAT, HIAT), but can, in principle, be used with any convention.

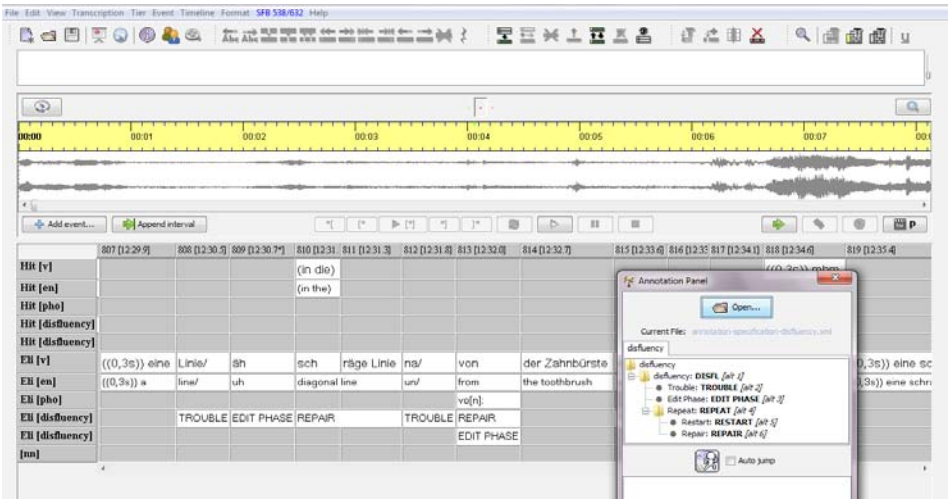


Figure 3. The Partiture-Editor used for transcription and multi-tier annotation. The Annotation Panel (bottom right corner) provides the categories of the annotation scheme and corresponding keyboard shortcuts.

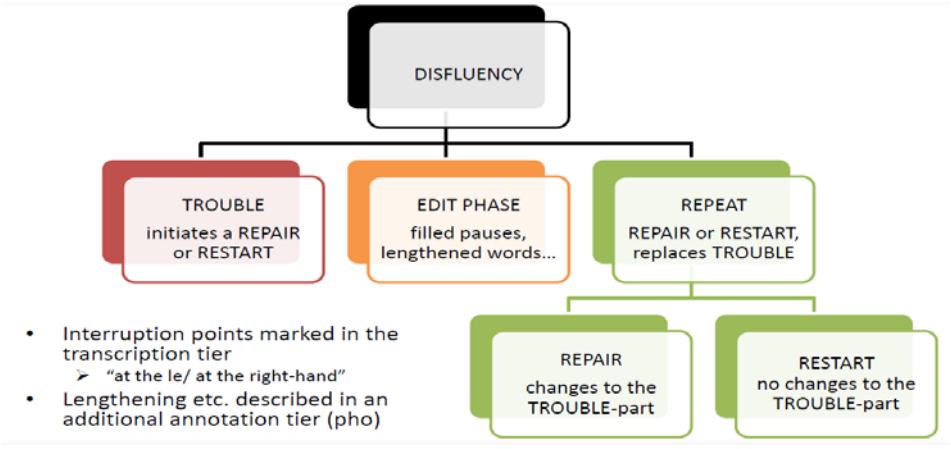


Figure 4. The disfluency annotation scheme

We annotated disfluencies (figure 4) such as hesitations, repetitions and self-repairs. For the self-repairs we used the idea of 'repair replaces reparandum' (cf. McKelvie, 1998). For PoS-annotation of the transcriptions the TreeTagger (Schmid, 1994) was integrated into the EXMARaLDA system and used to create stand-off annotation in the EXMARaLDA Sextant (Wörner, 2010) format. The stand-off annotation files can then be reviewed with the Sextant Tagger tool and manually corrected.

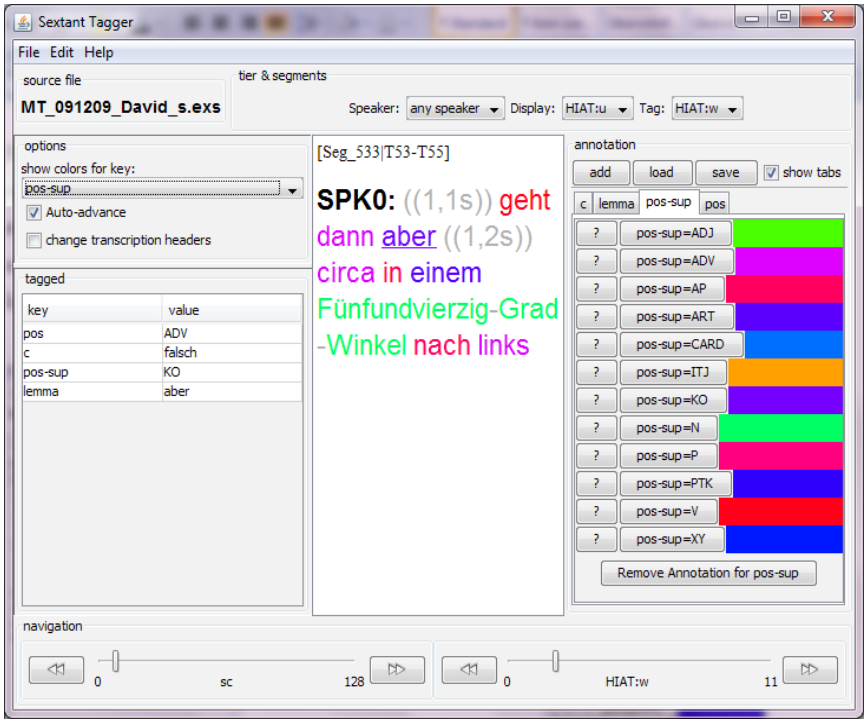


Figure 5. Reviewing TreeTagger annotations with the Sextant stand-off annotation tool.

3.2. Corpus compilation and metadata

The Corpus Manager (Coma) is a tool for managing corpus data and metadata. With Coma, recordings and transcripts were compiled into a corpus of communications with participating speakers. Figure 6 shows one of the communications. Participating speakers are marked by the paper clip-symbol in the speaker list. For the various abstract (communications, speakers) and concrete (transcription files, audio files) components of the corpus, we added the metadata collected at the recording sessions. Figure 7 shows the representation of the speaker Katinka and the metadata on her language knowledge and usage. By adding metadata in this structured format, consistency within and across projects can be achieved and the metadata also become available for corpus queries and analyses.

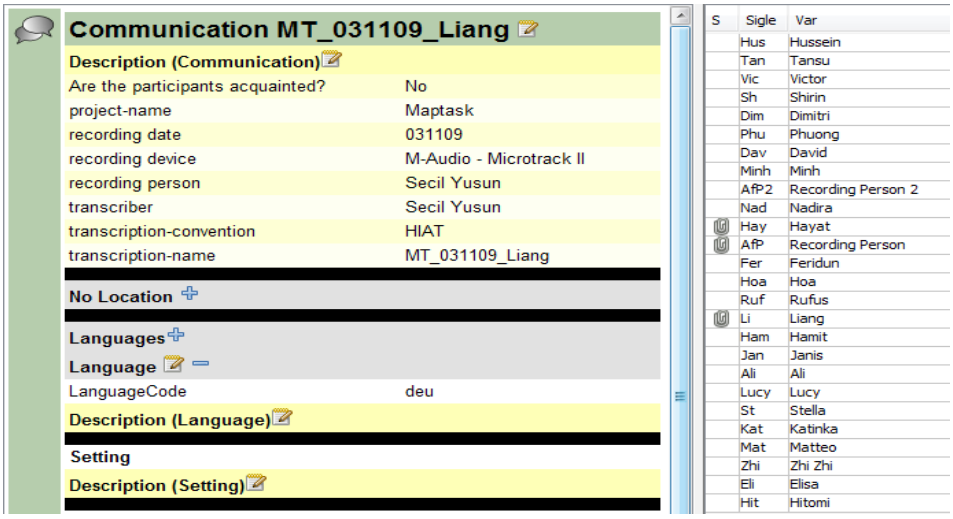


Figure 6. Communication with metadata and the participating speakers marked in the speaker list.

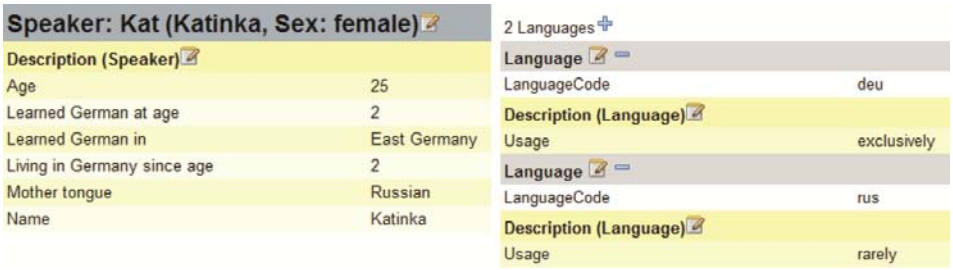


Figure 7. Example of metadata on speakers and their language knowledge and usage

3.3. Querying and analyzing corpus data

EXAKT is a tool for carrying out queries and analyses using both transcription data, i.e. the transcribed words and annotations, and the metadata on speakers and communications added in Coma. By correlating transcription data with information on age or mother tongue of the speaker (figure 8), we can filter the search results of the KWIC (KeyWord In Context) search accordingly. For each result, the transcription context is displayed as a musical score or an utterance list beneath the results list. The corresponding part of the aligned audio file can be played back. With the analysis function of EXAKT, it is also possible to further annotate such search results, as shown for the analysis of 'repair type' in figure 9.

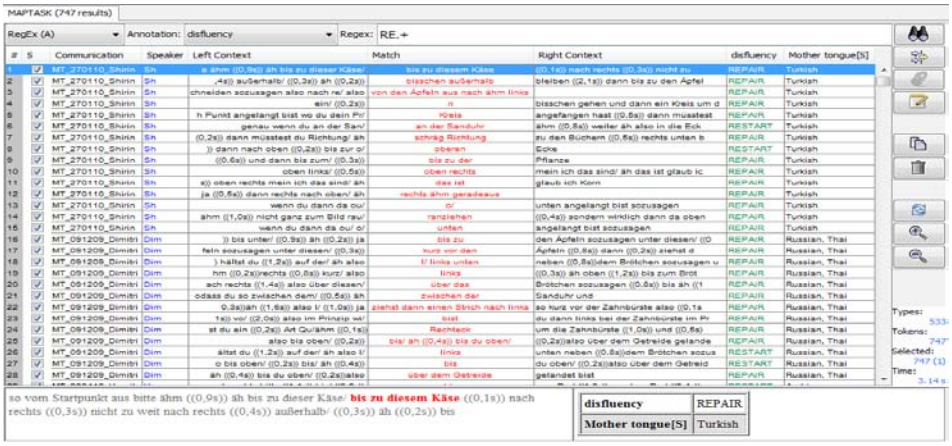


Figure 8. Search results for segments annotated as disfluencies starting with 'RE', correlated with speakers' mother tongue. The first search result with the match text bis zu diesem Käse is selected.

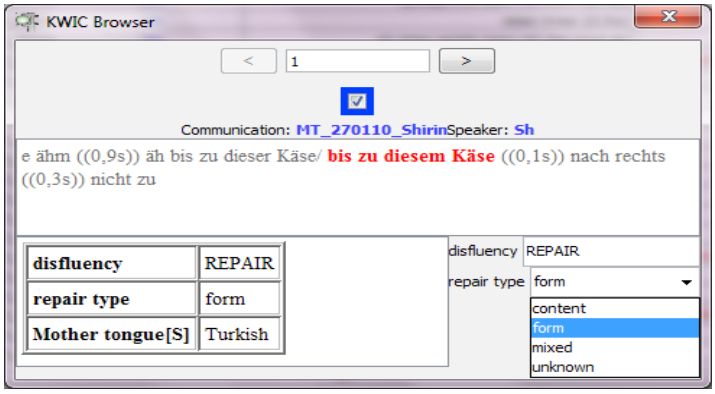


Figure 9. Here, the KWIC browser is used to add an analysis of the 'repair type' to the first search results

3.4. Corpus dissemination

Data sharing is a major desideratum, since "many researchers would agree that it is a basic scientific responsibility to make data collected in a research project available to the research community, especially when the research was supported by public funds" (The LIPPS Group, 2000: 134). A prerequisite is that all recorded speakers give their agreement to the publication and use of the data under specified conditions, e.g. anonymization/ pseudonymzation and only non-commercial use. We obtained the speakers' agreement

alongside metadata collection immediately before the recording session. Providing metadata on speakers and recording situations is another important aspect, since reusability of linguistic resources is often restricted without thorough documentation.

Standards such as XML and Unicode play an important role in creating sustainable linguistic resources. All EXMARaLDA file formats are plain XML. For publication, they can therefore be transformed into various XHTML formats with standard technologies like XSLT. For example, the overview page of the web interface (figure 2) is generated directly from the Coma file—the information used in figure 10 and 6 is thus entered once.

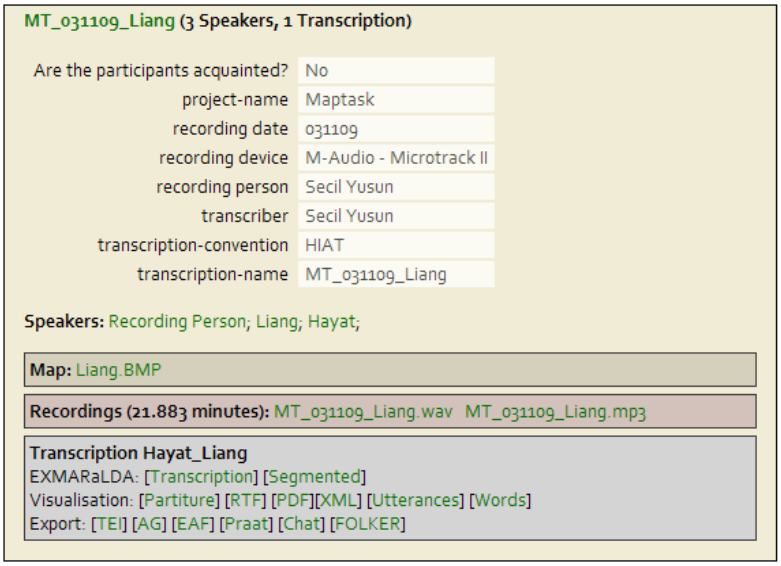


Figure 10. Web interface version of the MT_031109_Liang communication with metadata, speakers and associated files.

Apart from the EXMARaLDA transcription formats, there are automatically generated visualizations (figure 11) and export formats such as TEI or for tools like ELAN, Praat or CLAN. Since transcription is theory dependent, standardization to achieve comparability of digital corpora cannot apply to conventions of content and form across research fields. Providing different visualization or export formats is one solution to this problem.

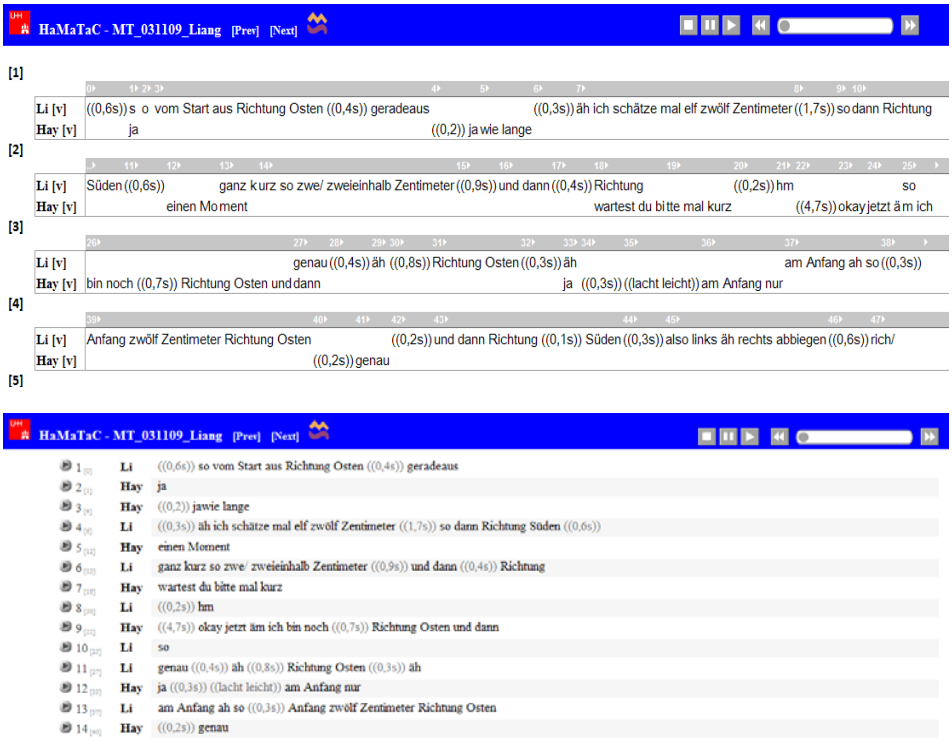


Figure 11. Visualizations generated from one and the same EXMARaLDA transcription in musical score format (above) and as utterance list (below).

4. Discussion

In this section I will address some recurrent methodological and technological issues in corpus building and sharing and try to show how technological and methodological aspects can be said to interact.

4.1. Transcription

One of the most fundamental questions arises from the non-trivial problems inherent in transcribing spoken language in general and learner language in particular: the representation of non-standard characteristics of the data. These are, especially for learner data, highly relevant. Though the HIAT transcription conventions (Rehbein et al., 2004) advocate the use of 'literary transcription' (*literarische Umschrift*) to encode pronunciation information by using nonstandard word forms, we decided to use standard orthography wherever possible and annotate such characteristics in a separate tier of the category 'pho'. There are two main reasons for this: Firstly, non-predictable literate transcription

impairs searchability, is imprecise and subjective. For example, the German word *rechts* (right) was annotated as either *reks*, *rek*, *lechts*, *lechtse* or *lechtsel* in 43 cases out of 297. Apart from being a search related problem, the initial consonant and the epenthetic vowel in the last three examples are probably not represented accurately enough by orthography for linguistic research questions. Secondly, as Gumperz and Berenz (1993) point out, existing stereotypes are an important aspect when creating and analyzing transcriptions of nonstandard varieties for e.g. conversational analysis. As the recording is aligned with the transcription, the pronunciation information can easily be retrieved from there at any time for a more thorough analysis.

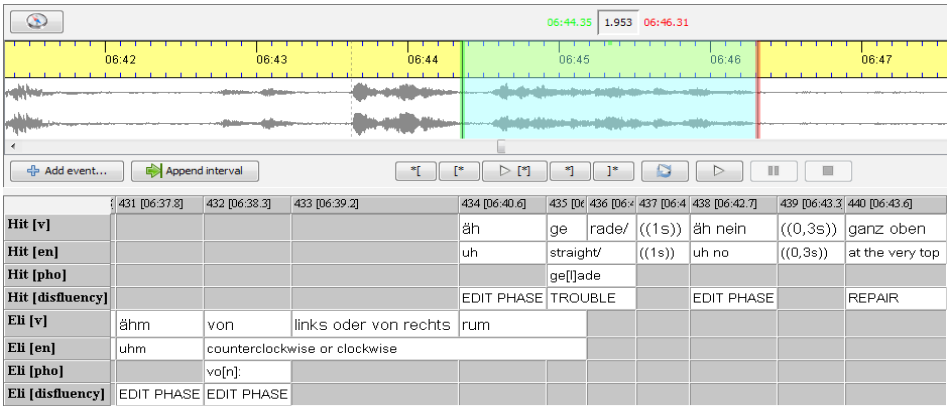


Figure 12. Annotations in pho-tiers for lengthening: vo[n]: (von) and learner pronunciation: ge[l]ade (gerade).

4.2. Annotation

When the complexity of the annotation scheme exceeds simple comments of one level of word-based annotation like POS, the technological solution becomes crucial. As figure 12 shows, the disfluency annotation system requires annotation tiers of different types ('disfluency' and 'pho'), and even multiple disfluency tiers for nested or overlapping structured disfluencies as is illustrated by the self-repairs in figure 13. The EXMARaLDA data model and the tools can handle independent multi-tier annotations: All disfluencies will be found in an EXAKT searches and the structure is visible in the transcription.

		48 [00:54.3*]	49 [00:55.4]	50 [00:55.6]	51 [00:56.0]	52 [00:56.6
Li [v]	veiter bisschen	da gibt es noch ne	Kro/	äh	Kro/	Kreise
Li [en]	a bit further	there, there's another	cirp/	äh	cirp/	circles
Li [disfluency]			TROUBLE	EDIT PHASE	RESTART	
Li [disfluency]					TROUBLE	REPAIR

	47 [00:46.1]	48 [00:46.8]	49 [00:47.2]	50 [00:47.3]	51 [00:47.4	52 [00:47.7]
Nad [v]	t da wo du die Runde/	bevor du	die/	die	Runde	gemacht hast
Nad [en]	there where you turned/	before you	tur/	turned		around
Nad [disfluency]	TROUBLE	REPAIR				
Nad [disfluency]			TROUBLE	RESTART		

Figure 13. Complex cases require multiple disfluency annotation tiers.

4.3. Quality

Another issue is the quality of manually created linguistic resources. Whereas e.g. part of speech most often can be corrected manually to achieve high quality, there is often no single correct answer for manual annotation tasks with interpretative categories. Technological solutions such as the annotation panel of the Partitur-Editor can help to increase consistency, but it will not solve the problem of disagreement in subjective annotation tasks.

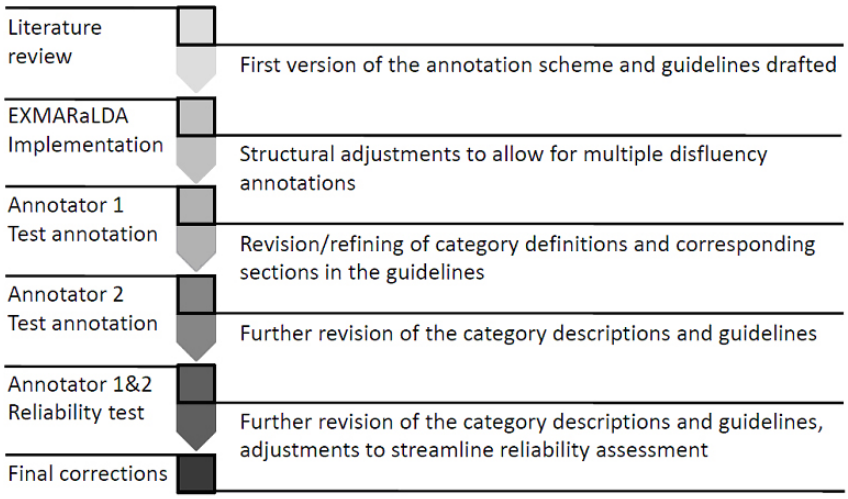


Figure 14. The disfluency annotation system development process

During the development of the annotation system and guidelines, we discussed the definition and examples of categories with our annotating student assistants and performed a reliability test with five transcriptions to gain insight into the difficulties of reliably performing manual annotation tasks. We found disagreement of various types and on different levels.

Hit [v]	((1,4s))	m	((0,9s))	wie n/	((0,4s))	wie letzte	((0,5s))	unten	((1s))	also	Rad
Hit [pho]											[L]ad
Hit [en]	((1,4s))	m	((0,9s))	as n/	((0,4s))	as last	((0,5s))	below	((1s))	well	Wheel
Hit [disfluency]		EDIT PHASE		TROUBLE		REPEAT				EDIT PHASE	

Hit [v]	((1,4s))	m	((0,9s))	(wie)/	((0,4s))	(wie)	letzte	((0,4s))	unten	((1s))	also	Rad
Hit [pho]				wie[n]:		wie[n]						[L]ad
Hit [en]	((1,4s))	m	((0,9s))	(as)/	((0,4s))	(as)	last	((0,4s))	below	((1s))	well	wheel
Hit [disfluency]				TROUBLE		RESTART						

Figure 15. Disagreement on the applicability of the EDIT PHASE category and on repair/ repetition type caused by differences in the transcription – wie vs. wie n/.

As discussed in Artstein and Poesio (2008) there are (inter-rater) reliability measures that work well for some annotation tasks. But it is not clear how to apply them in each case. How do we weigh disagreement on subcategories, extension of annotated segments or the structure of repair sequences? And how do we treat different interpretations of the recording? Without agreement on reliability metrics, thorough documentation and information on the annotators and their qualifications are indispensible to describe the quality of resources.

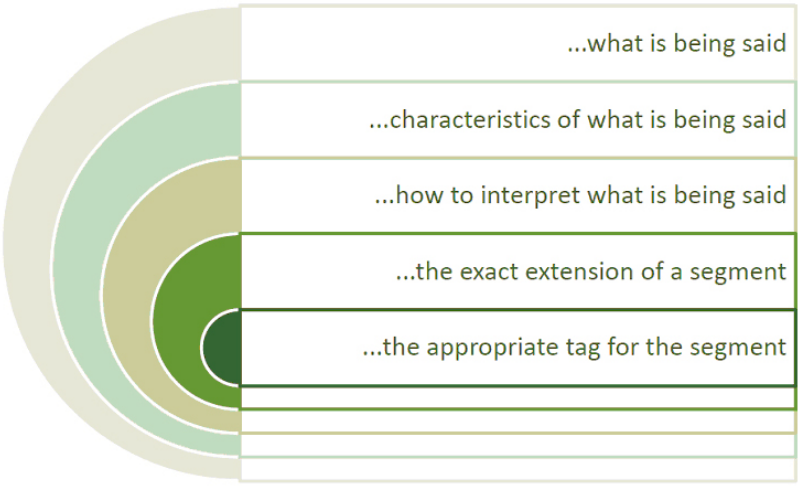


Figure 16. Agreement is required on several levels, on…

5. Outlook

As the discussion has shown, development through technological advances requires methodological reflection. The integration of multimedia into transcripts changes the conditions for methodological decisions regarding visual representation, especially when considering searchability. Searchability for large digital corpora requires the use of transcription system with systematic representation of lexical items, consistency in annotation and structured encoding of metadata, and these are methodological decisions. Multi-level (stand-off) annotation opens up new opportunities for annotation, though the design and validity of annotation systems remain methodological questions. And though any reliability measure could be automatically calculated, the precise application and interpretation of such metrics is not a technological question either. Consequently, the interaction with technological aspects plays an important role in further developing the methodology of linguistic corpus building and sharing.

References

Artstein, R. & Poesio, M. (2008). Inter-coder agreement for computational linguistics. Computational Linguistics 34(4), 555–596.

Brinckmann, C., Kleiner, S., Knöbl, R. & Berend, N. (2008). German Today: an areally extensive corpus of spoken Standard German. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Marokko.* Retrieved from http://www.lrec-conf.org/proceedings/lrec2008/pdf/806_paper.pdf

Gumperz, J. J. & Berenz, N. (1993). Transcribing Conversational Exchanges. In Edwards, J. A. & Lampert, M. D. (Eds.), *Talking Data: Transcription and Coding in Discourse Research* (pp. 91–121). Hillsdale, NJ: Erlbaum.

The LIPPS Group (2000). The LIDES Coding Manual. A document for preparing and analyzing language interaction data. Version 1.1 – July, 1999. *International Journal of Bilingualism, 4*(2) 131–270.

McKelvie, D. (1998). *The syntax of disfluency in spontaneous spoken language* (HCRC Research Paper HCRC/RP-95). Edinburgh: Human Communication Research Centre.

Rehbein, J., Schmidt, T., Meyer, B., Watzke, F. & Herkenrath, A. (2004). *Handbuch für das computergestützte Transkribieren nach HIAT* (Arbeiten zur Mehrsprachigkeit, Folge B, 56). Hamburg: Universität Hamburg (SFB Mehrsprachigkeit).

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging using Decision Trees. *Proceedings of the 1st International Conference on New Methods in Language Processing, Manchester, England.* Retrieved from http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf

SCHMIDT, T. (2005). *Time-Based Data Models and the Text Encoding Initiative's Guidelines for Transcription of Speech* (Arbeiten zur Mehrsprachigkeit, Folge B, 62), Hamburg: Universität Hamburg (SFB Mehrsprachigkeit).

SCHMIDT, T. & WÖRNER, K. (2009). EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics, 19*(4), 565-582.

SCHMIDT, T., HEDELAND, H., LEHMBERG, T. & WÖRNER, K. (2010). *HAMATAC – The Hamburg MapTask Corpus*. Retrieved from http://www.exmaralda.org/files/HAMATAC.pdf

WÖRNER, K. (2010). *Werkzeuge zur flachen und hierarchischen Annotation von Transkriptionen gesprochener Sprache.* PhD Thesis, Bielefeld University, Bielefeld. Retrieved from http://bieson.ub.uni-bielefeld.de/volltexte/2010/1669/pdf/diss_gold.pdf

Design and compilation of a legal English corpus based on UK law reports: the process of making decisions

María José Marín Pérez
Universidad de Murcia

Camino Rea Rizzo
Universidad de Murcia

Abstract

The scarceness of reliable specific teaching materials and corpora within the field of legal English has led us, as lecturers of this ESP variety, to engage into corpus design. The available corpora existing do not satisfy our needs as we intend to establish the core vocabulary of this ESP branch, so we have opted for creating our own, BLaRC: British Law Report Corpus. The selected genre are law reports (judicial decisions) as they stand at the very basis of common law systems (uncodified) where the existing jurisprudence plays a determining role. The purpose of this paper is thus to show and justify the decisions we have made in its process of compilation and design.

Keywords: *legal corpus, ESP, common law, representativeness, corpus size, word target.*

Resumen

La escasez de materiales y corpus específicos fiables dentro del área del inglés jurídico nos ha llevado, como profesoras de esta variedad de inglés específico, a involucrarnos en el diseño de un corpus. Los corpus específicos existentes no satisfacen nuestras necesidades pues nuestro objetivo es establecer el vocabulario básico de esta variedad, por este motivo hemos optado por crear nuestro propio corpus, BLaRC: British Law Report Corpus*. El género seleccionado son los* law reports *(decisiones judiciales) ya que son una pieza fundamental en los sistemas legales* common law *(no codificados) en los que la jurisprudencia existente juega un papel esencial. El propósito de este artículo es por tanto mostrar y justificar las decisiones que se han tomado en su proceso de diseño y compilación.*

Palabras clave: *corpus legal, ESP,* common law*, representatividad, tamaño del corpus, word target*.

1. Introduction

It is commonly agreed that the amount of available teaching materials in English for Specific Purposes is considerably scarce in most fields (Rea, 2010). This derives into a clear methodological void which must be filled, thus resorting to specific corpora becomes a valuable method for ESP professionals. As McEnery and Wilson affirm: "[...] such corpora can be used to provide many kinds of domain-specific material for language learning" (1996: 121).

One of the main obstacles we encountered as lecturers of legal English was deciding on what method and particularly what materials to select in order to teach this specially obscure variety of English. As stated above, resorting to specific corpora was an interesting option, however, to our knowledge, the amount of written legal corpora is also reduced and designing our own became an urge. We thus engaged into ESP corpus design and decided to create the British Law Report Corpus (BLaRC): a legal English corpus of law reports (judicial decisions) that could act as a reliable source of specific vocabulary which could be employed to create new materials, as well as information for further linguistic analysis.

The purpose of this paper is thus to present the process of design and compilation of BLaRC according to Corpus Linguistics standards as stated in Wynne (2005) for general corpora and its adaptation to specific corpora (Rea, 2010). First, we will focus on the state of the art by looking into the legal corpora available; next, we will justify the reasons that lead to the selection of this legal genre. The mode of the texts and the organization of the corpus into different categories will also be explained to finish with some final remarks on further corpus applications and future research.

2. State of the art

By using the term *state of the art* we are actually referring to the amount of available legal corpora we have found and their main goals and characteristics. As stated above, the amount of such corpora is scarce and the purpose they were created with, in most cases, did not satisfy our need for a specific legal corpus we could employ to identify the core vocabulary of law reports (a key genre within legal English) as well as to carry out further linguistic analyses.

The first corpus worth mentioning is the BoLC since this is probably the most comprehensive legal corpus existing due to its selection of texts from varied genres and topics. It is a multilingual comparable Italian-English corpus which aims at "representing the two different legal systems, in particular the differences between the civil law and the common law systems"[19].

However, the rest of corpora we found were either too small to act as a normative reference for us, or inaccessible. They either focused on aspects of the language we are

19 This quotation has been taken from the corpus website.

not interested in, or were conceived as parallel corpora with translational or comparative purpose.

The JRC-Acquis Corpus is one of them. It is a multilingual parallel corpus which includes European Union legislative texts affecting all Members States in twenty-two different languages.

The CorTec corpus is a scientific-technical parallel one divided into four sections, one of them deals with commercial law and includes agreements and contracts in English and Brazilian Portuguese.

As for the HOLJ corpus, it is a monolingual synchronic one comprising 188 judgments of the House of Lords from 2001 to 2003, its aim is to define a set of rhetorical role labels.

To finish, the Cambridge International Corpus, owned by Cambridge University Press, has a legal corpus section of twenty million words. It is not accessible or commercialised.

There also exist legal sections or materials within some of the best known general British English corpora like BNC or COBUILD, but they could not serve our purpose either as they are non-specific.

3. Law reports: a legal genre at the core of common law systems

Establishing the *sampling frame,* that is, "the entire population of texts from which we [would] take our samples" (McEnery and Wilson, 1996: 78), was our first objective, and law reports were selected due to the pivotal role they play in the UK judicial system as well as in any other common law countries. Following Sinclair: "the contents of the corpus should be selected (...) according to their communicative function in the community in which they arise" (Wynne, 2005: 5).

If representativeness is crucial for the design of any corpus (Biber, 1993; McEnery and Wilson, 1996; Sánchez, Cantos, Sarmiento and Simón, 1995; Sinclair, 1991; Wynne, 2005), narrowing the boundaries of our object of study became a must as we soon realised how legal language is intertwined with everyday language, how it is present both in the public and private fields, and consequently how the vastness of this ESP branch could not be covered or managed as a whole in a project of this nature. Therefore, we decided to focus on one of the most relevant legal genres in this ESP variety: law reports.

The United Kingdom belongs to the realm of common law, as opposed to civil or continental law which is the judicial system working in most Western European countries. In purely common law systems, the acts passed at their parliaments have gained greater importance being most often cited in case decisions. However, case law stands at the very basis of common law systems which rely on the principle of binding precedent to work, that is to say, a case judged at a higher court must be cited and applied whenever it is similar to the one being heard in its essence (the *ratio dicendi*). Another fact that makes

law reports an outstanding genre in common law legal systems is that they not only cover all the branches of law, but also touch upon other public and private law genres.

Due to the widespread use of information technologies, there is a tendency towards digitalising these texts and storing them in online databases, so case citation has recently become an easier task that used to take ages for legal practitioners to become fully informed about the existing jurisprudence in the past.

There are different ways of accessing cases online, most of them are restricted. However, the British and Irish Legal Information Institute (BAILII.org) has created a completely free and comprehensive online database where more than 200,000 authentic texts are available. It is supported by a number of sponsors like the Inns of Court (barristers' professional associations), law faculties like Cambridge, Oxford, Glasgow, law firms and other prestigious institutions, hence its importance and recognition by professionals. This is precisely why we have decided to use it as our main source to obtain the legal texts that form BLaRC.

4. Lexical variation and corpus structure

As well as abiding by hierarchical criteria when organizing the corpus, one of the first elements that conditioned our choice was the way that legal vocabulary varies according to the system where it is used. This is so because of the laws and regulations that organise the countries which the UK is divided into. The judicial systems of Northern Ireland, Scotland, England and Wales do not solely depend on UK institutions, but rather have their own autonomous systems and structure. But for the Supreme Court (in general terms) and the UK Tribunal Service (except for some cases), each country is fully independent as regards its judicial system.

This being so, we decided to structure BLaRC into five main categories depending on the jurisdictions of their judicial systems, that is, the geographical scope of their courts and tribunals:

1. Commonwealth countries
2. United Kingdom
3. England and Wales
4. Northern Ireland
5. Scotland

5. Time span covered by BLaRC

BLaRC is a specific synchronic monolingual corpus of judicial decisions made in the UK. Following Pearson: "(…) a specific corpus compiled for terminological studies, [should

include texts] (…) delivered in the last 10 years prior to the date of compilation" (1998: 51), this is why we decided to compile the texts produced at UK courts and tribunals from 2008 to 2010 as we expect to finish compiling them before the end of 2011.

Moreover, due to the changes that the structure of these courts has experienced due to the recent modifications of the law that regulates it, we considered that, if the structure of the corpus responds to the structure of UK courts and tribunals because of thematic and hierarchical reasons -as it will be shown below-, it should adjust to the latest modifications it has experimented following the new regulations, hence the time span covered. We are specifically referring to the *Constitutional Reform Act, 2005* and the *Tribunals, Courts and Enforcement Act, 2007*.

6. Text mode

Oral texts were disregarded given that obtaining oral samples of legal language that reflected the arguments, facts, and decisions dealt with at court, would have implied having access to courtrooms and permission to record the trial sessions, a certainly complicated objective for Spanish researchers merely interested in linguistic data. Therefore, BLaRC only covers the written mode and in raw text format, as the corpus is not intended to be tagged.

Regarding the texts themselves, they are authentic transcriptions of judicial decisions whose structure may vary depending on the nature of the case and the hierarchical position of the court where it was heard. They are full texts in digital format from BAILII gathered randomly within the time span established.

The average size of the texts is 2,000 to 2,500 words, although there is great variation. They have all been produced by British judges and reflect their decisions about the cases in question as well as the facts, arguments, prior decisions made at other courts, and any other kind of information relevant to the case.

7. Size and representativeness: key elements in corpus design

Representativeness is central to corpus design, as shown above, and the size of a corpus may determine whether it is representative of the variety of language it aims at covering, or simply an illustrative sample of it with no predictive value.

Authors do not agree regarding the recommended size for a specific corpus. Whereas Pearson (1998) proposes a million words as a reasonable number, Sinclair (1991) believes that corpora must be as large as possible. On the other hand, Kennedy (1998) does not think that a big corpus necessarily represents the language better than a small one.

Taking these arguments into consideration as well as the availability of legal texts and their high numbers (16,612 texts in total from 2008 to 2010), we established our word target. Also the relevance of law reports in the judicial system coupled with the great

amount of topics covered by these texts, was determining when we had to make the first decisions on the size of our corpus.

As a consequence, we established that, although this is a specific corpus based just on one legal genre, the target should be 6,000,000 words (approximately), six times as big as Pearson proposes, essentially because of the easy access to already digitalized texts in either .rtf or .pdf format and, naturally, all the principles behind corpus compilation.

8. The lexical comprehensiveness of law reports

Law reports should not only be paid special attention within ESP because of their essential function in common law systems, but also because of their vast topic coverage. This corpus has been organised according to the source where the texts originated, that is, what court or tribunal cases were heard at and decided on.

Tribunals and courts are specialized in a given branch of law: criminal law, family law, commercial law, intellectual law, etc., and law reports touch upon one and every branch of both the private and public fields. Judges are in charge of judging cases by both interpreting the law itself (the statutes passed at the parliament), and fundamentally taking into consideration the existing precedents. Therefore their judgments, as reflected on law reports, pertain to all the fields of law.

9. BLaRC structure and distributional criteria

McEnery and Wilson highlight the importance of justifying the categorisation of any corpus when citing Biber: "Biber ... emphasises the advantage of determining beforehand the hierarchical structure (or strata) of the population, that is, defining what different genres, channels and so on it is made up of" (1996: 79), this is why we believed it was essential to do so.

To begin with, our corpus retains the current UK tribunal and court structure as reflected on BAILII due to several reasons, the first one being the relevance of the hierarchy of courts and tribunals in the UK legal system. The principle of binding precedent, which the British judicial system revolves around, establishes that any decision made at a higher court or tribunal will set binding precedent as long as the case is similar to the one under examination, as stated above.

Secondly, if we maintain this structure, the texts will be grouped according to the field of law they belong to, so they will be similar in lexical terms, and comparing results by studying the categories separately will be easier and respond to a thematic criterion we consider fundamental as far as our further objective is concerned, that of establishing the core vocabulary of the genre.

To finish with the enumeration of the criteria that have conditioned the organization of the corpus, we would like to refer to the distribution of the population in the UK. As it

is shown in the UK official census 2011, elaborated by the Office of National Statistics, it appears that almost 90% of the population of the whole territory is concentrated in England and Wales while Northern Ireland only has about 3 % and Scotland 9%. Although we have not mathematically distributed the number of texts and word targets per category and subcategory depending on these figures, we did take them into account in order to reinforce the representativeness of the texts obtained from English and Welsh sources that amounted to approximately 55% of the total[20].

10. Final remarks

This paper has aimed at presenting all the stages followed in the design and ongoing compilation of a new corpus of legal English which may satisfy the linguistic needs of ESP students. As we firmly believe that the quality of the results deriving from corpus analysis depends crucially on the rigorous establishment of the corpus, we have closely observed the principles governing corpus compilation and tried to apply them to the design, collection and projection of BLaRC.

Taking advantage of the law reports made available on digital databases on the internet, we have access to a vast amount of naturally occurring samples of the language used by the judges that explain an order in any type of cases, and therefore, covering all possible issues reaching the court. Corpus Linguistics' techniques permit dealing with such amount of authentic samples and process them in such a way that we could obtain worthy and reliable results from the analysis of the language from several approaches.

An essential tool for corpus analysis is the computing programme selected for its processing. WordSmith.5 will be used to look into the samples through quantitative and qualitative analyses, first by assessing BLaRC's basic computational characteristics (types, tokens, type-token ratio, frequency lists, etc.) and second, by adopting a corpus comparison approach which enables to gain a deeper insight into legal English.

Even though BLaRC has been envisaged to serve multiple purposes in the long term, since the potential applications of a corpus are manifold, our overriding objective consists in identifying the essential vocabulary in legal English for the ease of teaching and learning. Moreover, we aim at filling in a gap for discipline-based lexical repertoires which may guide materials writers, assist ESP practitioners and notably meet students' specific needs (Nation, 2001; Hyland and Tse, 2007; Read, 2007; Rea, 2008). The framework of our future research is set by the long tradition of developing word lists (Coxhead, 2000; Nation, 1990; Thorndike and Lorge, 1944; West, 1953) for teaching and learning English as a second language.

20 The structure and distribution of the word targets per section and subsection will be exemplified in Appendix 1.

Appendix 1

This table exemplifies the structure of BLaRC divided into five main categories which are subdivided according the court and tribunal structure in each of them respectively. The UK court and tribunal section (number two in the general structure) comprises twenty-two subcategories, the distribution of the word targets in each of them has been made according to the number of texts available with respect to the total and also with respect to the 6m overall word target of the corpus. We have kept the numeric order main categories have been assigned in the general structure of the corpus.

Table A1

2. UK courts and tribunals

Court/ Tribunal	Available Texts	% Of Total	Word Target
2.1.Supreme Court	117	0,71%	42,600
2.2. House of Lords	74	0,45%	27,000
2.3. Upper Tribunal (Administrative Appeals Chamber)	550	3,31%	198,600
2.4.Upper Tribunal (Tax and Chancery)	44	0,27%	16,200
2.5.Upper Tribunal (Immigration and Asylum Chamber)	59	0,36%	21,600
2.6. Upper Tribunal (Lands Chamber)	135	0,82%	49,200
2.7. First Tier General Regulatory Chamber	124	0,75%	45,000
2.8. First-tier Tribunal (Health Education and Social Care Chamber)	139	0,84%	50,400
2.9. First-tier Tribunal (Tax)	865	5,21%	312,600
2.10. Competition Appeals Tribunal	100	0,61%	36,600
2.11. Nominet UK Dispute Resolution Service	370	2,23%	133,800
2.12. Special Immigrations Appeals Commission	24	0,15%	9,000
2.13. Employment Appeal Tribunal	971	5,85%	315,000
2.14. Financial Services and Markets Tribunal	16	0,1%	6,000
2.15. Asylum and Immigration Tribunal	141	0,85%	51,000
2.16. Information Tribunal including the National Security Appeals Panel	130	0,79%	47,400
2.17. Special Commissioners of Income Tax	80	0,49%	29,400
2.18. Social Security and Child Support Commissioners	219	1,32%	79,200
2.19. VAT & Duties Tribunals (Customs)	20	0,12%	7,200
2.20. VAT & Duties Tribunals (Excise)	92	0,56%	33,600
2.21. VAT & Duties Tribunals (Insurance Premium Tax)	1	0,01%	600
2.22. VAT & Duties Tribunals (Landfill Tax)	2	0,02%	1200

References

Alcaraz Varó, E. (1994). *El inglés jurídico: textos y documentos.* Madrid: Ariel Derecho.

Alcaraz Varó, E. (2000). *El inglés profesional y académico*. Madrid: Alianza Editorial.

Bhatia, V. (1993). *Analysing Genre: Language Use in Professional Settings*. London: Longman.

Bhatia, V. (2004). Applied genre analysis: a multi-perspective model. *Iberica 4,* pp 3-19.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing,* 8 (4).

Biber, Conrad and Reppen. (1998). *Corpus Linguistics. Investigating Language Structure and Use.* Cambridge: C.U.P.

BoLC (Bononia Legal Corpus). Disponible en http://corpora.dslo.unibo.it/bolc_eng.html

Consitutional Reform Act, 2005. Disponible en:
http://www.legislation.gov.uk/ukpga/2005/4/contents

Dudley-Evans, T. and St John, M. (1998). *Developments in English for Specific Purposes*. Cambridge: Cambridge University Press.

Hutchison, T. and Waters, A. (1998). *English for Specific Purposes*. Cambridge University Press.

Hyland, K. & P. Tse (2007). Is there an "Academic Vocabulary"? *TESOL Quarterly, 41*(2), 235-253.

Kennedy, G. (1998). *An introduction to corpus linguistics*. New York: Longman.

Kennedy, G. y Bolitho, R. (1984). *English for specific purposes*. London: Mcmillan.

Maley, Y. (1987). The Language of Legislation. *Language and Society, 16.*

McEnery, T. and Wilson, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Mellinkoff, D. (1963). *The Language of the Law.* Boston: Little, Brown & Co.

Nation, P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Orts Llopis, M.A. (2006). *Aproximación al discurso jurídico en inglés: las pólizas de seguro marítimo de Lloyd's.* Madrid: Edisofer.

PEARSON, J. (1998). *Terms in Context*. Amsterdam: John Benjamins Publishing Company.

REA, C. (2008). *El inglés de las telecomunicaciones: estudio léxico basado en un corpus específico.* Tesis doctoral. Disponible en http://www.tesisenred.net/TDR-0611109-134048/index_cs.html

REA, C. (2010). Getting on with Corpus Compilation: from Theory to Practice. *ESP World, 1* (27), Volume 9.

READ, J. (2007). Second Language Vocabulary Assessment: Current Practices and New Directions. *International Journal of English Studies*, *7* (2). Universidad de Murcia.

ROSSINI, R ET AL. (2001). Words from the Bononia Legal Corpus. *International Journal of Corpus Linguistics,* Vol. 6 (special issue), 13-34

SÁNCHEZ, A., CANTOS, P., SARMIENTO R., SIMÓN, J. (1995). *Cumbre. Corpus lingüístico del español contemporáneo. Fundamentos, metodología y análisis*. Madrid: SGEL.

SINCLAIR, J. (1991). *Corpus, Concordance and Collocation*. Oxford: Oxford University.

THORNDIKE, E.L. AND LORGE, I. (1944). *The teacher's Word Book of 30,000 Words*. New York: Teachers College, Columbia University.

TIERSMA, P. (1999). *Legal Language.* Chicago: The University of Chicago Press.

TRIBUNALS, COURTS AND ENFORCEMENT ACT, 2007. Disponible en http://www.legislation.gov.uk/ukpga/2007/15/contents

WYNNE, M. (Ed.) (2005).*Developing Linguistic Corpora: a Guide to Good Practice.* Oxford: ASDS Literature, Languages and Linguistics.

Gleaning micro-corpora from the internet: integrating heterogeneous data into existing corpus infrastructures

Karlheinz Mörth

Institute for Corpus Linguistics and Text Technology (Austrian Academy of Sciences)

Niku Dorostkar

Institute of Linguistics (University of Vienna)

Alexander Preisinger

Institute of Culture Studies and History of Theatre (Austrian Academy of Sciences)

Web as Corpus usually implies the wholesale and rather indiscriminate download of large amounts of data on the basis of so-called seeds, keywords used to create lists of relevant URLs. Corpus technology offers tools to exploit corpora created in such a manner for a great number of NLP purposes, in particular in the field of lexicographic research. However, problems often arise when scholars need access to the archived texts. Our paper touches on methodologies to create smaller corpora of internet data tailored to particular needs that go beyond access to the data on the level of words or sentences, that also allow researchers to perform more text oriented studies. We introduce a newly developed piece of software used to create such corpora and describe one particular use case, a project conducting research into racist language in online discussion forums applying methods of critical discourse analysis.

Corpus typology, corpus creation, corpus tools, corpus design

Generalmente el "Web as Corpus" implica la descarga indiscriminada de grandes cantidades de datos sobre la base de "seeds" (semillas), palabras clave utilizadas para crear listas de URLs. Hoy en día, la tecnología ofrece herramientas para explotar corpus creados de tal manera por un gran número de efectos de PLN, en particular en investigaciónes lexicográficas. Sin embargo, los problemas surgen cuando los investigadores necesitan acceso a los textos archivados. Nuestra ponencia trata de metodologías para crear pequeñas corpus de datos de internet adaptados a necesidades particulares que van mas allá del acceso a los datos sobre el nivel de palabras o frases, permitiendo a los investigadores llevar a cabo estudios más orientados a texto. Presentamos un nuevo desarrollo de software utilizado para crear corpus tales y describimos un caso de uso, una investigación sobre el lenguaje racista en foros de discusión en línea aplicando métodos de análisis crítico del discurso.

Tipología de corpus, creación de corpus, herramientas de corpus, diseño de corpus

Preliminaries

Over the past decade, the issue of *Web as Corpus* has been discussed and studied extensively (Grefenstette, 1999; Jones & Ghani, 2000; Duclaye, Yvon, & Collin, 2003; Fairon, Naets, Kilgarriff, & Schryver, 2007). Meanwhile, the existence of a number of very successful projects and the ever growing number of new corpora created from sources on the internet bears advocates of this new brand of NLP resources out. Corpora created from internet sources have been used in various fields. The most obvious area of application that comes to mind are of course the various linguistic disciplines (Hundt, Nesselhauf, & Biewer, 2007). In this field, studies based on data from the internet have a long-standing tradition (v. for instance Volk, 2000). A special type are parallel corpora created from internet sources which have been put to use in translation studies (Grefenstette, 1999; Resnik & Smith, 2003). The most prominent field is probably lexicography. Most software developments that have been presented in recent years have been geared towards the needs of researchers looking for words, less to the reading and interpreting kind of scholars, such as linguists interested in the text(s) or literary scholars.

The number of tools that serve the purpose of creating and accessing this type of corpus has steadily grown. It will suffice to mention two important examples: the *BootCaT Utilities* (Baroni & Bernardini, 2004) or *CLEANEVAL* (Baroni, Chantree, Kilgarriff, & Sharoff, 2008). Some of these tools also provide web-based interfaces. The meanwhile well-established methodology of creating corpora from the Web has produced applications that allow the wholesale creation of very large corpora (Kilgarriff & Grefenstette, 2003). However, there are no out of the box solutions for individually working researchers.

Use case: migration.macht.schule (MIMAS)

The investigations described in this paper were carried out in the framework of *Sparkling Science,* a funding programme of the Austrian Federal Ministry of Science and Research (BMWF) which started in 2007. The programme focuses on promoting young people's interest in science and research in an unconventional manner, encouraging scientists and scholars to work side by side with young people in real-world scientific research projects.

The objective of the MIMAS project was to conduct research into discourses of migration and education. The material originated from online discussion forums published in the online newspaper *derStandard.at*. Applying methods of critical discourse analysis in accordance with the discourse-historical approach and qualitative corpus-based methods, linguists collaborated in their research with students of a public high school. While racist comments in online discussion forums have repeatedly attracted public attention, more detailed research on the particular phenomenon needs yet to be conducted. Filling this obvious research gap was one goal of the project. In addition to that, the project was also supposed to give high school students a basic understanding of the notions of critical discourse analysis and to provide them with a set of tools to analyse discourses they are confronted with in their everyday life whether on the internet or in other media. This was meant to promote the students' language awareness as well as to improve their capability

to challenge racist discursive strategies and consequently to make an essential contribution to their civic education. Students aged 16-17 years were introduced to computer-assisted methods being applied to the ends of discourse analysis, learned how to annotate and analyse digital textual corpora.

Apart from the linguistic aspects, the project had also a strong technological focus. The performance requirements of the project included the development of tools to find and archive web pages for subsequent analysis. One of the research perspectives was the creation of a workable workflow for comparable endeavours.

The MIMAS Corpus

At this point, it is important to emphasize the difference from other projects working on data from the Internet. What was needed for the particular purpose, was a clearly defined scope of web pages, not a random collection as is usually the case with corpora created from Internet sources. When talking about *Web as corpus*, one usually thinks of very large amounts of texts and some of the largest text collections created for linguistic purposes have been collected from the internet. By contrast, the *Mimas Corpus* is very small, it amounts to under 4 million tokens. The tools and workflows described here are supposed to have a wider range of applicability as they can be applied to very different use cases. Actually, a number of such smaller experimentally motivated corpora was set up. Most of these experiments had a domain/genre specific orientation with a strong lexicographic/ terminological perspective.

The texts investigated are public internet discussion forums, communications taken from one particular online discussion site where people can communicate via posted messages, and the original journalistic reports to which they refer, on which they comment. The discussion forums under investigation are generally regarded as the oldest, largest, and positively most lively such medium in the country. Messages are posted under pseudonyms and organised in threads. The publishers of the online newspaper exercise a certain degree of supervision attempting to filter out explicit racist statements up front.

The language represented in these web pages represents–for the most part–a variety of Austrian German, actually a very particular hybrid sort of language in the sense that although purely written in form, it displays a large amount of features usually encountered in spoken language. It is characterised by irregular orthography, idiosyncrasies, often rather unorthodox wording. The texts abound in creative wordplay und puns. By contrast to the ICLTT's previously created corpora–for the greater part historical data from the 19th and 20th centuries–this data documents contemporary language.

In quest of tools: the downloadBrowser

Most of the existing tools used to create corpora from the web were not usable for what we intended to do in our project, which is due to a number of reasons. Software used in

Web as corpus projects usually proceeds from so-called seeds, keywords used to collect URLs from search engines. These URLs are subsequently downloaded from the web for future reference (Baroni & Bernardini, 2004). While this method allows a certain degree of control over the contents of the data, creating sufficiently granular metadata to conduct research on specific types of texts or within a particular genre usually turns out to be difficult.

While creating ever larger corpora has become a comparatively easy task for computational linguists (Pomikalek, Rychly, & Kilgarriff, 2009), other groups of researchers who might also be interested in archiving and exploiting such data still come up against a number of difficulties that often impede smooth access to data. Among the design objectives of our development project was also to enable non-technical users to archive data from the internet, to organise this data into reusable micro-corpora, to enhance data with more fine-grained metadata and to integrate them into an existing corpus infrastructure. What we had in mind was an easy to handle, integrated tool to take care of the downloading process. First of all, it had to be capable of browsing the internet, not in a random but in a purposeful manner. Secondly, it should download everything needed to conserve the data simply at the push of a button whenever users come across a site that matched their needs. Ease of handling was one of the basic prerequisites. While one might argue, that most of this functionality is basically provided by all up-to-date web browsers, these regrettably do not give the users much control about what, where and how to store. A third issue were metadata. The tool had to be capable of storing relevant information concerning the texts being archived in concurrence with the process of downloading.

The tool which was developed at our department as a response to this short wish list is called *downloadBrowser*. It is a standalone Windows application that–making use of existing libraries and software components–could be developed with very little programming overhead. Basically, it consists in a web browser component which allows users to surf the internet making use of any search engine, then they decide and download the required data. To perform the downloading process, the tool makes use of *GNU wGet*, a very useful free tool that was developed as part of the GNU project quite some time ago. *GNU wGet* is a command line tool which takes charge of the whole archiving process. It should be mentioned here that there exist a number of other graphical wrappers for *GNU wGet* (a list can be found on http://de.wikipedia.org/wiki/Wget).

Downloading is not an automatic process, although *downloadBrowser* offers bulk-downloading of lists of URLs as an option. What makes *downloadBrowser* as a corpus creation tool different is that it allows users to perform their collection tasks manually. It has been optimized for simplicity and is provided with some additional built-in functionalities such as automatic batch-download (a feature particularly needed to download the pages investigated in the *MIMAS* project which are usually split across several batches). The tool creates metadata while downloading and helps users to keep track of URLs in order to avoid repeated archiving of one and the same document.

Defining a practicable workflow

Having collected and archived whatever required for the particular purpose, the data have to be made available, have to be prepared in a manner that they become searchable. To achieve this end, one might conceive of a number of workflows targeting various applications. A prototypical workflow to achieve the task can be described in the following manner:

a) Conversion of texts

b) Metadata creation/refinement

c) Application of linguistic mark-up

d) Indexing

Most *Web as corpus* projects proceed by sifting out repetitive data, eliminating much of what is regarded as superfluous (boilerplate removal) (e.g. Sharoff, 2006). In our project all original data was conserved, nothing removed from the original. However, to smoothen processing procedures, a second data set was created from the original documents on which a certain degree of clean-up was performed. During this step, character encoding was normalized to UTF8 and the texts converted to XHTML resp. XML.

Although most basic metadata were collected during the download process, a number of secondary data were extracted from the downloaded material in a second step. This includes marking-up of nicknames used in the online forums, position within the particular threads and date and time of posting.

The application of a modicum of linguistic data was the next and final step of data refinement. The tag set we made use of was the Stuttgart-Tübingen Tagset (STTS) which is the most widely used system in German NLP applications.

Integrating the data into existing infrastructures

To realise the above basic workflow in the framework of our existing corpus infrastructure, we made use of several tools: tools developed at the department and other tools available for research purposes. The ICLTT's tools are all results of research endeavours of the past years and form part of a more general corpus toolbox.

The tool that was used in steps (a) and (b) is called *corpedUni,* an XML editor designed to support authoring of small to medium-sized XML documents. *corpedUni* was designed to perform a wide range of text encoding and corpus management tasks: it is used to create texts and to furnish these with mark-up. It is designed with a special focus on XML documents (but is also capable of processing HTML, XHTML, RTF, CSS, JS etc.) and has a number of built-in features catering in particular to the needs of Humanities text

encoders. *corpedUni* can also be used to define workflows and to protocol processing steps. The tool has been used by partner projects and is freely available for academic purposes.

In the third step (c) of the workflow, the *Treetagger* was used, a widely used product of the Institute for Computational Linguistics of the University of Stuttgart. *Treetagger* is a well-tested, well documented and very robust probabilistic part-of-speech tagger which also performs lemmatisation (Schmidt, 1994).

Indexing (d) is accomplished by means of the linguistic search engine *ddc-concordance*, an open source (http://sourceforge.net/projects/ddc-concordance) tool which is capable of performing general purpose word form searches. It offers neatly definable range operators and is somehow comparable to *SARA*, the search engine designed to access the British National Corpus. *ddc-concordance* was primarily designed as a linguistic search engine, as such it is very versatile and robust.

To access the corpora created in the above described manner, an interface was needed that would be capable of launching queries, visualizing the results and navigating the texts, both the converted and original texts. The ICLTT's *corpusBrowser* tool proved to be well suited to serve the purpose. Actually, it combines several integrated interfaces which makes it a multifunctional corpus access tool. The most basic functionality is provided by the query control which comes with query builder and query editor components.

All development activities have been carried out with a strong emphasis on standards. Technologically, they build on W3C recommendations (XML and cognate technologies and Unicode). For text encoding tasks TEI (Text Encoding Initiative, P5) as de-facto standard is applied, others such as XCES (Corpus encoding standard for XML) are also taken into consideration. In addition, ISO TC 37's *Data Category Registry* application (ISOCat) has been used extensively.

Perspectives

The joined efforts of departments of the Austrian Academy of Sciences and the University of Vienna have also to been seen as part of the ICLTT's commitment in setting up efficient corpus infrastructures. As the coordinator of CLARIN-AT, the ICLTT is highly interested in both a more distributed and yet more integrated corpus landscape in which interoperability of methods and standards play an important role.

Among the remaining tasks of the described project is the creation of more detailed technical documentation, such as for instance a step by step guide which will allow also technically less experienced users to set up their own corpora. The software developed and applied in the project will be maintained and further developed in upcoming projects. Current versions are available from the ICLTT's website.

References

Baroni, M., & Bernardini, S. (2004). BootCaT. Bootstrapping Corpora and Terms from the Web. *Proceedings of LREC 2004*. Lisbon, 1313-1316.

Baroni, M., Chantree, F., Kilgarriff, A., & Sharoff, S. (2008). CleanEval: A competition for cleaning webpages. *Proceedings of LREC 2008*. Marrakech. Retrieved from http://www.lrec-conf.org/proceedings/lrec2008.

Duclaye, F., Yvon, F., & Collin, O. (2003). Unsupervised incremental acquisition of a thematic corpus from the web. *Proceedings of Natural Language Processing and Knowledge Engineering*. International Conference on Natural Language Processing and Knowledge Engineering. 752-757.

Fairon, C., H. Naets, A., Kilgarriff, A., & de Schryver, GM. (Eds.) (2007). Building and Exploring Web Corpora. *Proceedings of the third WebasCorpus workshop, incorporating Cleaneval.* Presses Universitaires de Louvain, Louvain la Neuve, Belgium.

Grefenstette, G. (1999). The WWW as a resource for example-based MT tasks. *ASLIB Translating and the Computer Conference*. London.

Hundt, M., Nesselhauf, N., & Biewer, C. (Eds.) (2007). Corpus Linguistics and the Web. *Language and computers studies in practical linguistics*, 59. Amsterdam – New York.

Jones, R., & Ghani, R. (2000). Automatically building a corpus for a minority language from the web. *38th Meeting of the ACL, Proceedings of the Student Research Workshop*. Hong Kong, 29-36.

Kilgarriff, A., & Grefenstette, G. (2003). Web as Corpus: Introduction to the special issue. *Computational Linguistics*, 29 (3), 333–347.

Pomikálek, J., Rychly, P., & Kilgarriff, A. (2009). Scaling to Billion-plus Word Corpora. *Advances in Computational Linguistics. Special Issue of Research in Computing Science*, 41. Mexico City, 3-13.

Resnik, P., & Smith N. (2003). The Web as a parallel corpus. *Computational Linguistics* 29. 349-80.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*. 44-49.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. *Working papers on the Web as Corpus*. Bologna. Retrieved from wackybook.sslmit.unibo.it/pdfs/sharoff.pdf.

Volk, M. (2000). Scaling up. Using the WWW to resolve PP attachment ambiguities. *Proceedings of Konvens 2000*. Sprachkommunikation, Ilmenau, VDE Verlag, 151-156.

Towards a Latvian Treebank

Lauma Pretkalniņa*, Gunta Nešpore*, Kristīne Levane-Petrova, Baiba Saulīte

Institute of Mathematics and Computer Science

University of Latvia

Abstract

In this paper we describe preparatory work for constructing a Treebank for Latvian as no such resource currently exists. Previously elaborated SemTi-Kamols *hybrid dependency based grammar model has been extended to make it appropriate for broad coverage text annotation. We also have integrated extended* SemTi-Kamols *model with graphical tree editor* TrEd *and complementary toolkit, which originally was developed for Prague Dependency Treebank. Using the obtained environment we have annotated small amount of Latvian text.*

Keywords: Treebank, Latvian, dependency grammar, hybrid grammar, SemTi-Kamols

Introduction

Treebanks are among the crucial resources for the development of NLP tools. For Latvian no such resource currently exists. To address this deficiency the development of Latvian Treebank is ongoing.

As a grammatical framework for the Latvian Treebank, the *SemTi-Kamols* grammar model (Bārzdiņš, Grūzītis, Nešpore, & Saulīte, 2007; Nešpore, Saulīte, Bārzdiņš, & Grūzītis, 2010) is used. It is a hybrid grammar in relation to dependency and phrase structure grammars. This model covers both synthetic and analytical forms of Latvian in a linguistically adequate way. It is not a simple task as Latvian is a highly synthetic language with relatively free word order and rich morphology.

SemTi-Kamols model is strongly based on the pure dependency parsing mechanism described by Covington (2001). Meanwhile it is fundamentally extended with a constituency mechanism to handle analytical multi-word forms consisting of fixed order mandatory words. This enables us to elegantly overcome the limitation of the pure dependency grammars, where all dependants are optional and totally free-order. In *SemTi-Kamols* approach a head and a dependant don't have to be single orthographic words anymore (Bārzdiņš et al., 2007).

Apart from dependency links, the *SemTi-Kamols* model is based on a concept of *x-word*: a syntactic unit describing analytical word forms and relations other than subordination. The concept of *x-word* is analogous in some extent to the Tesnière's *nucleus* — the primitive element of syntactic description introduced by (Tesnière, 1988). From the phrase structure perspective, *x-words* can be viewed as non-terminal symbols, and as such substitute (during the parsing process) all entities forming respective constituents. From the dependency perspective, *x-words* are treated as regular words, i. e., an *x-word* can act as a head for depending words and/or as a dependent of another head word. Similarly as "ordinary" words *x-words* also have rich morpho-syntactic annotation. It is mostly inherited from their constituents, but additional information that specifies the kind of an *x-word* can be included as well, allowing to check for additional agreement restrictions while applying the dependency functions (Grūzītis, 2010).

When integration of *SemTi-Kamols* with *TrEd toolkit* (Hajič, Vidová Hladká, & Pajas, 2001) was started (Pretkalniņa, Nešpore, Levāne-Petrova, Saulīte, 2011), we saw that *SemTi-Kamols* model needs to be extended and clarified to cover texts of different domains and genres. *SemTi-Kamols* in its initial version covers only simple sentences, so the support for composite sentences has to be developed. Also the concept of *x-word* needed to be clarified and developed further.

Extended SemTi-Kamols grammar model

The key question for extending the *SemTi-Kamols* model was the following: what kind of relations do we need to model apart from dependency? The dependency relations in the extended *SemTi-Kamols* model are treated the same way as before. Dependency pairs are

the basic relation in the model — they cover subordination by attaching the subordinate element by its governor regardless the position (Nešpore et al., 2010).

The scope of *x-word* was narrowed down by excluding coordination from the *x-word* scope, and one additional construction — punctuation mark construct (*PMC*) — was introduced in the extended *SemTi-Kamols* model. The constructions dealing with other relations than subordination all can be treated similarly as the *x-words* in the initial model: from the dependency view it acts as the regular word, but from the phrase view it act as non-terminal symbol combining its components in the single unit. The distinction among these three constructions is their inner structure — which elements are mandatory, which elements are optional, which elements can act as dependency head and the syntactic relations (or absence of syntactic relations) between the elements.

Thus we arrive at four relation types: dependency, *x-word*, coordination, and punctuation mark construct. Each of these constructions (except coordination, but this may change in future) is divided further in subtypes to give more information about their inner structure and/or functions. *X-words* and coordinated parts of sentence use the rich morpho-syntactic tags developed in the initial grammar model.

Punctuation Mark Construct

The first relation type introduced anew is punctuation mark construct. The motivation behind this concept is the fact that punctuation in Latvian reflects its grammatical structure. This makes punctuation an essential component to determine the syntactic structure. For example, let us look at two sentences "Sodīt nedrīkst, apžēlot!" and „Sodīt, nedrīkst apžēlot!" („sodīt" — „to punish", „nedrīkst" — „is not permitted", „apžēlot" — „to amnesty"). The only difference between these two sentences is the comma, but the first sentence translates as 'It is not permitted to punish [somebody], [you] must amnesty [him]!' while the second sentence translates as 'It is not permitted to amnesty [somebody], [you] must punish [him]!'.

What distinct *PMC* from the phrase-like relations mentioned above (*x-word* and coordination) is its inner structure. *PMC* consists of one mandatory core element, some (usually one or two) optional punctuation mark elements and optional elements which bare no syntactic role in sentence (like addresses, insertions etc.). The mandatory element is the syntactic unit evoking the use of punctuation marks represented by the optional elements. The mandatory element usually is the only *PMC* element which can directly participate in the dependency relation. Elements with no syntactic role usually are *PMC* themselves (see Figure 1) and can have elements participating in dependency relations.

Owing to *PMC* we can handle most of the punctuation usage cases. The most important thing — the clauses of the compound sentence are represented by *PMC* with the predicate as core element (see Figure 3).

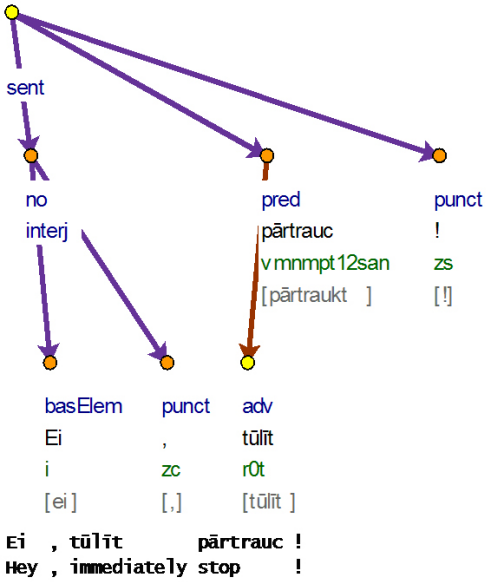


Figure 1. Sentence 'Hey, stop it immediately!' demonstrates two *PMC* — overall sentence is a *PMC* consisting from core element *'pred'* and optional elements *'no'* and *'punct'*; and element *'no'* also is a *PMC* consisting of an interjection and punctuation

In the initial model only some punctuation was covered and it was done with *x-words*.

Coordination

In the initial *SemTi-Kamols* approach the coordination relation was one of the *x-words*, as coordinated parts of sentence has the same syntactic function in a sentence (Nešpore et al., 2010).

However, the relation between coordinated parts of the sentence is fundamentally different from the relations between the constituents of the analytical forms or multi-word units, therefore in the extended model the coordination was distinguished as a separate relation. This brings the *SemTi-Kamols* model even closer to the Tesnière's structural syntax, where coordination (*jonction*) is one of the basic concepts. Coordination (horizontal) relationship differs from a subordination (vertical) relationship, it is formed by two or more homogenous nodes that have the same function but these nodes are not constituents of one nucleus like multiword units (Tesnière, 1988).

The coordination relation can link different types of syntactic units, therefore in the extended model the same relation is used to represent both coordinated parts of sentence (see Figure 2) and coordinated clauses (see Figure 3). If it links coordinated parts of sentence, it is annotated with morpho-syntactic tag inherited from those coordinated parts.

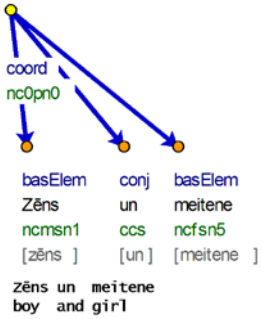


Figure 2. Fragment 'The boy and the girl' demonstrates the coordinated parts of sentence

Elements composing coordination structure can be divided in two types — elements representing coordinated parts and supporting elements (conjunctions and punctuation marks). Coordination structure must consist of at least two coordinated parts and usually at least one supporting element between each two coordinated parts. Only coordinated parts can act as heads of dependency.

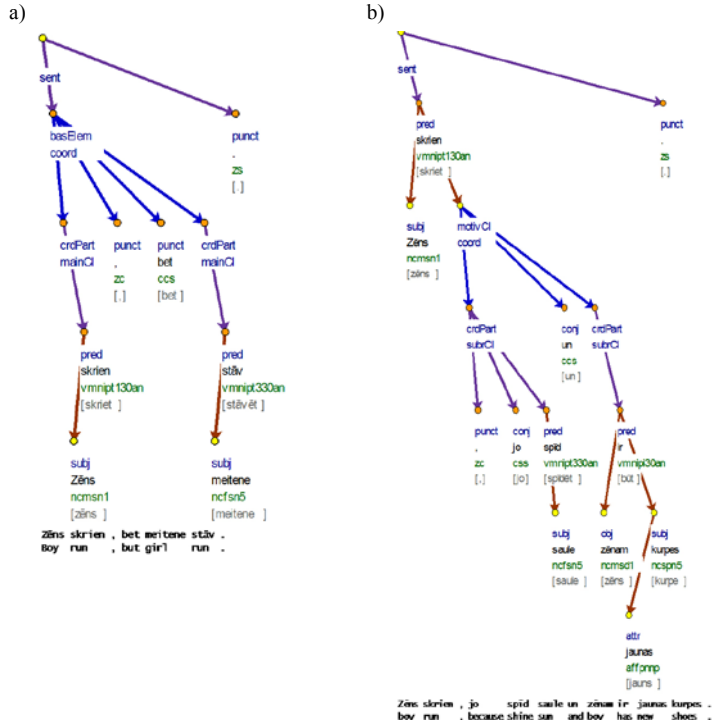


Figure 3. Sentence 'The boy is running, but the girl is standing.' (a) demonstrates coordinated main clauses. Sentence 'The boy is running because sun is shining and the boy has new shoes.' (b) demonstrates coordinated subordinate clauses.

X-word

X-word in the extended *SemTi-Kamols* somehow comes back to its original concept — being a multiword unit where every element is mandatory.

X-words are used to describe various syntactic constructions, though relations between elements in the inner structure of *x-words* are different. This information is reflected indirectly by the type of the particular *x-word* (*x-Verb*, *x-Preposition*, *x-Apposition*, etc.). This type also determines how the morpho-syntactic tag for the *x-word* is obtained and which *x-word* elements can act as dependency heads.

We have following types of *x-words* for Latvian. First are *analytical forms*: perfect tenses of verb (*x-Verb*, see Figure 4 a) and prepositional phrases (*x-Preposition,* see Figure 4 b). *X-Verbs* and *x-Prepositions* are formed by one content word and one or several function words. *X-Preposition* combines a preposition (rarely postposition) and a noun (or a pronoun), *x-Verb* combines at least one auxiliary verb and one content word (participle, noun, adjective, adverb or pronoun) (Nešpore et al., 2010). In these constructions usually only content word can act as dependency head.

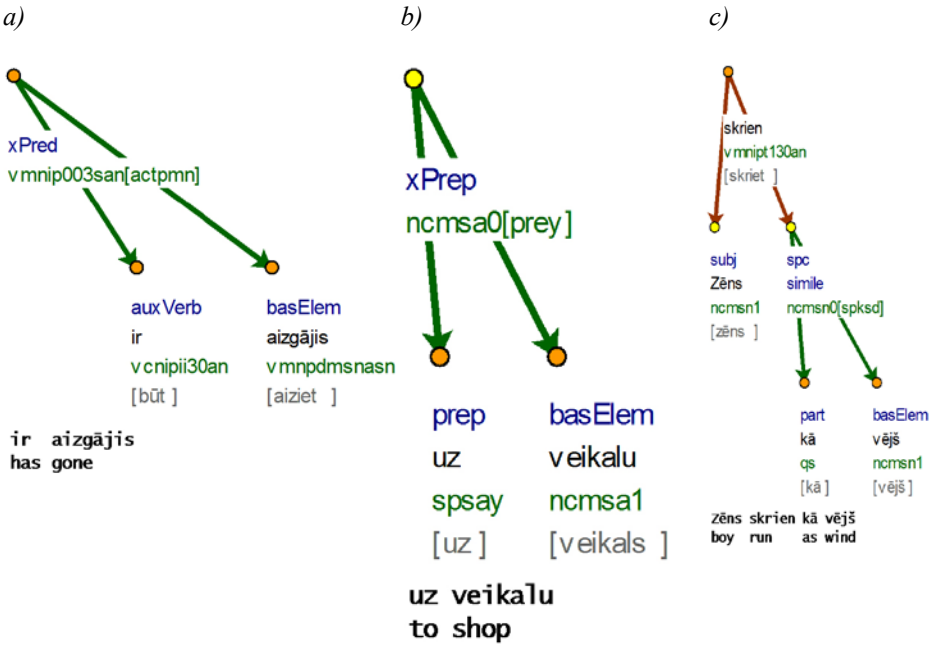


Figure 4. Fragment '[he] has gone' (a) demonstrates *x-Predicate*; fragment 'to shop' (b) demonstrates *x-Preposition*; fragment 'The boy runs like a wind' demonstrates simile and the way how *x-words* are incorporated in the syntax tree.

Second type of *x-word* is *simile* (see Figure 4 c). It is formed by one content word and one function word. In this case also only content word can act as a dependency head.

Third type of *x-word* is multiword units (*named entities*, *analogues of wordgroup*, *idioms* and *multiword numerals* and *appositions*). The distinctive feature of this type of *x-words* is that no element of these *x-words* can be used as dependency head, thus all the elements of these *x-words* will occur in the text one right after another.

Annotating *named entities* and *idioms* is one of easiest sources to the ambiguous annotation of the Treebank — distinguishing whether the given fragment of a text is an idiom or not often relays on an annotator's previous experience and subjective interpretation. When the annotation is done by multiple annotators, it is easy to obtain different annotations to the same text strings. This was the main concern why we decided to annotate inner syntactic structure of the idioms and the named entities that have clear tree representation (see Figure 5). In this way the representation of a string as an idiom or named entity becomes more similar to the case when the same string is not recognised as idiom or named entity, thus making post-processing of such potentially ambiguous mark-up easier.

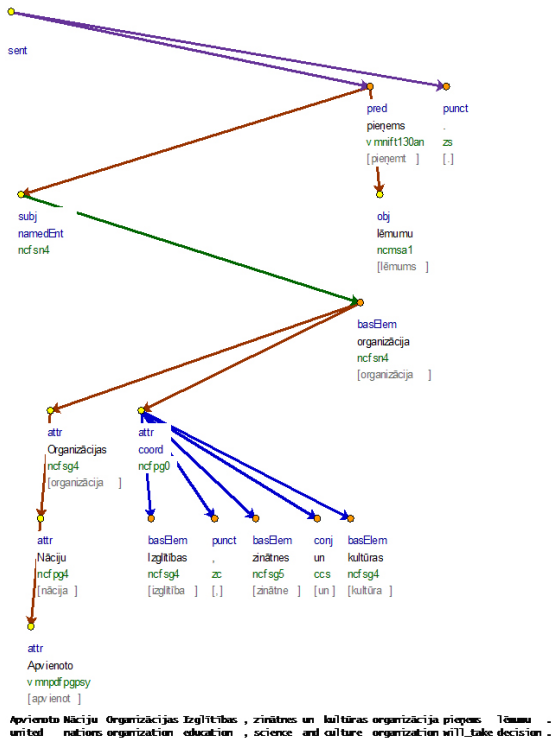


Figure 5. Sentence 'United Nations Educational, Scientific and Cultural Organisation will make the decision.' demonstrates how *named entity* is used; everything in the subtree below '*namedEnt*' belongs to this *named entity*

Integration with TrEd toolkit

We have integrated the extended *SemTi-Kamols* model with *TrEd toolkit* — tools developed for Prague Dependency Treebank (Hajič, Böhmová, Hajičová, & Vidová Hladká, 2000). The central tool in this toolkit is *TrEd* — customisable graphical editor for tree-like structures. The default data format for *TrEd* toolkit is Prague Markup Language (*PML*) (Pajas, & Štěpánek, 2006). It is *XML* based mark-up language developed to suit the needs of the linguistic annotations. It is independent from annotation scheme; it supports multilayer annotations and offers verification by the *PML schema*. *TrEd toolkit* also includes tools for querying treebanks and tools for batch processing trees (Hajič et al., 2001).

We have developed *PML* profile for extended *SemTi-Kamols* model annotations, thus obtaining *XML* based data format for Latvian Treebank (Pretkalniņa et al., 2011). Also we have developed an extension module for *TrEd* to enable full *TrEd* support for our format (Pretkalniņa et al., 2011). The extension we developed contains stylesheets, *PML* schemas for our data format and macros to automate common annotation tasks.

Using all the above mentioned *TrEd* can be used as an environment for manual creating/editing Latvian Treebank.

Summary

Preparatory work for Latvian Treebank development is successfully ongoing. We have extended *SemTi-Kamols* dependency based hybrid grammar model to fit most syntax constructions of Latvian by additional relations — like *punctuation mark construct* — and clarifying the existing relations — like *x-words* and coordination.

We have developed extension module enabling us to use graphical tree editor *TrEd* as an annotation environment.

Using the obtained results we have created small Treebank as a proof of concept. We have annotated first 100 sentences of J. Gaarder's "Sophie's World" (Pretkalniņa et al., 2011) and ~100 sentences of Latvian fiction text.

Even the annotated text amount is still small, it contains the broad coverage of syntax constructions of Latvian, and thus we estimate that *SemTi-Kamols* model is very close to cover all Latvian.

For creating bigger Treebank we are working on integrating the obtained environment with *SemTi-Kamols* rule-based partial parser (Bārzdiņš et al., 2007).

Rerferences

Bārzdiņš, G., Grūzītis, N., Nešpore, G., & Saulīte, .B. (2007). Dependency-Based Hybrid Model of Syntactic Analysis for the Languages with a Rather Free Word Order. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA)*, (pp. 13–20).

Covington, M.A. (2001). A Fundamental Algorithm for Dependency Parsing. In *Proceedings of the 39th Annual ACM Southeast Conference*, (pp. 95–102).

Grūzītis, N. (2010). *Formal Grammar and Semantics of Controlled Latvian Language*. Summary of Doctoral Thesis in Computer Science. Riga, University of Latvia.

Hajič, J., Böhmová, A., Hajičová, E., &Vidová Hladká, B. (2000). The Prague Dependency Treebank: A Three-Level Annotation Scenario. In A. Abeillé (ed.), *Treebanks: Building and Using Parsed Corpora*, Amsterdam: Kluwer, (pp. 103–127).

Hajič, J., Vidová Hladká, B., & Pajas, P. (2001). The Prague Dependency Treebank: Annotation Structure and Support. In *Proceedings of the IRCS Workshop on Linguistic Databases*, Philadelphia, USA, (pp. 105–114).

Nešpore, G., Saulīte, B., Bārzdiņš, G., & Grūzītis, N. (2010). Comparison of the SemTi-Kamols and Tesnière's Dependency Grammars. In *Proceedings of the 4th International Conference on Human Language Technologies — the Baltic Perspective*, Frontiers in Artificial Intelligence and Applications, Vol. 219, IOS Press, (pp. 233–240).

Pajas, P., & Štěpánek, J. (2006). XML-Based Represen-tation of Multi-Layered Annotation in the PDT 2.0. In *Proceedings of the LREC Workshop on Merging and Layering Linguistic Information (LREC 2006)*, (pp. 40–47).

Pretkalniņa, L., Nešpore, G., Levāne-Petrova, K., & Saulīte, B. (2011). A Prague Markup Language Profile for the SemTi-Kamols Grammar Model. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA),* (pp. 303–306).

Tesnière, L. (1988). *Основы структурного синтаксиса.* (Trans.) Ред. В.Г. Гак. Москва: Прогресс (Original work published 1959).

MATVA: a database of English television commercials for the study of pragmatic-cognitive effects of paralinguistic and extralinguistic elements on the audience of English TV ads

Laura Ramírez-Polo

Universidad de Valencia

Abstract

The elements of television commercials are chosen with a purpose in mind, that of maintaining the product in the public eye or persuading the audience to buy the product. Attesting the complexity of this type of texts, the MATVA research group (Multimedia Analysis of TV Ads) devised the creation of a database with commercials from the UK as a special speech corpus with an analysable textual component made up of the transcriptions of the voice-overs, on-screen text and dialogues. This paper addresses the difficulties encountered and the decisions made in the design and construction of the database, which constitutes a valuable resource for the study and analysis of linguistic, paralinguistic and extralinguistic elements of TV ads and well as their pragmatic-cognitive effects on the audience.

Keywords

speech corpus, TV advertisements, paralinguistic elements, extralinguistic elements, pragmatic-cognitive effects

Resumen

Los diferentes elementos que conforman un anuncio de televisión se seleccionan con un propósito en mente, a saber, mantener la atención del espectador o persuadirlo para comprar un producto. Teniendo en cuenta la complejidad de este tipo de textos, el grupo de investigación MATVA (Multimedia Analysis of TV Ads) propone la creación de una base de datos con anuncios del Reino Unido en forma de corpus oral especial con un componente textual analizable formado por las transcripciones. El presente artículo aborda las dificultades y las decisiones tomadas en el diseño y construcción de dicha base de datos que constituye un valioso recurso para el estudio y análisis de los elementos paralingüísticos y extralingüísticos de los anuncios de televisión así como sus efectos pragmático-cognitivos en la audiencia.

Palabras clave

corpus oral, anuncios de televisión, elementos paralingüísticos, elementos extralingüísticos, efectos pragmáticos-cognitivos

Introduction

The structure of television commercials, the soundtrack, voice-overs, actors' accents, etc. are not the result of random decisions. Rather they are chosen with a purpose in mind, that of maintaining the product in the public eye or persuading the audience to buy the product.

Indeed, there are a number of elements that can influence consumer behaviour. Studies such as that by Luna & Gupta (2001) stress the importance of cultural values in advertising, whereas Kassarjian (1977) addresses content analysis in consumer research in order to analyse how content and, especially, how the language that conforms that content, can influence consumer behaviour.

In the specific case of advertisement, Pennock-Speck (2005) studies how different voices correlate with certain products. Another example of the use of para- and extralinguistic elements as well as the verbal message is presented by Pennock-Speck & Del Saz Rubio (2009) who state that a wide range of strategies is used in menstrual-related products to create a positive image of their potential female customers.

Attesting the complexity of this type of documents, the MATVA research group (Multimedia Analysis of TV Ads) devised the creation of a database with commercials from the UK conceived as a spoken corpus as described by Sinclair (1996) but taking into account the multimodality of this genre. The particularity of this corpus resides in the fact that it is made up of written-to-be-spoken texts. It includes the sound and visual elements of the ads as well as an analysable textual component made up of the transcriptions of the dialogs, the voice over texts and the lyrics and the on-screen texts, resulting in a valuable parallel corpus for the study and analysis of both para- and extralinguistic elements and the pragmatic-cognitive effects these have on the audience.

This paper addresses the difficulties and decisions made in the design and construction of the database. In the first place, I address some of the theoretical questions we have faced in the conceptual design. First of all, I tackle the criteria established by the Eagles Spoken Language Working Group for the acquisition of data. Further, I discuss the criteria defined by Sinclair (1996) within the EAGLES initiative to create corpora: quantity, quality, simplicity and documentation. Besides, I consider the main aspects for designing a corpus dealt with by Torruella & Llisterri (1999): its goal(s), limits and the type of corpus.

I then discuss some practical issues regarding the metadata that accompanies each advertisement including the classification schema used to organize the commercials as well as the different variables that constitute the database: product types, ad duration, song lyrics, etc. I also explain the annotation schema developed in order to transcribe the commercials. Finally, I mention some technical factors such as the platform used to store the data as well as the structure of the data, and I end with the conclusions about the construction of the database and give future prospects for further research.

The matva database - Acquisition of data and design

Llisterri (1996) established for the EAGLES Spoking Language Working Group two main criteria for the acquisition of data[21]. These criteria have been overcome since then

[21] These criteria are: 1. If acceptable in the recording environment and for optimal acoustical quality use headset microphones an; 2.Use digital recording devices

as the technology available nowadays is much more sophisticated than that of 1996, when these recommendations were issued. Our video files were recorded digitally and converted to a video format, where both visual and audio quality were optimal. Besides, the ability to work with the files directly on the computer facilitated the transcriptions and the construction of the database.

Following the EAGLES guidelines, documentation concerning the recording session date and time was also compiled. The recordings were carried out during two consecutive days in March and two in June 2009 on ITV1 and Channel 4 from approximately 8 am to 6 pm.

The resulting video files were edited, the programs were cropped and files in mpg format containing around 2 hours of advertising for each day were created. There were 437 ads for the 19th of March, 219 ads for the 20th of March, 330 ads for the 24th of June and 306 ads for the 25th of June, though many of them were repeated as is to be expected. To date, only the ads from the June sub-corpus have been fully processed and the rest are still being transcribed.

The first step consisted in elaborating a list of all the ads, defining the type of information to be recorded and organizing them in a spreadsheet. We designed a schema containing twenty three fields (see appendix I). Though the relational model was first considered, we decided to start working in a spreadsheet for convenience reasons, being aware that later it would be possible to export the data into a structured database management system such as Microsoft Access or FileMaker.

The fields of the MATVA database comprise administrative fields such as the ID, the date and time in which the advertisement was broadcasted, the duration, previous and following programs and the name of the featured product. All of these fields represent extralinguistic information regarding the main fields, namely the transcriptions, which constitute the corpus to be analysed.

I now deal with the fields that were conceptually more challenging to implement: the advertising types and the product classification.

Advertising types

Fictional works such as advertisements have been categorised within advertising research as usually belonging to one of the three major literary genres used to classify fictional works, namely drama. Wells (1989) distinguishes between lectures and dramas and believes that advertisements are made up of both these components. The lectures usually centre on the characteristics of the product; there may be usually a speaker, which we call a testimonial, directly addressing the consumer, or a voice-over, presenting the product information and using persuasive language and exhortation. Dramas, on the contrary, can be classified as narratives, where actors speak not directly to the audience but to each other. Stern (1994) reinforces this view stating that there is indeed a dichotomy between "drama" and "lecture" or "argument" resting on narrative "telling" in lecture versus non-narrative "showing" in drama. This is reflected in lectures using narrators who describe events, whereas dramas allow characters to perform or show the events directly. Stern (op.cit) further distinguishes between classical dramas and vignettes. Whereas a classical narrative has a very strict causal structure, like a mini drama, the vignette (or montage)

narrative is more like a loose gathering of stories, the so-called slice-of-life type (Hoven, 2009: 11). Finally, we distinguish a further type of narrative structure: image sequence, defined as a sequence of images. This last category can be described negatively as not being either a drama nor a montage.

After an analysis of the ads conforming our database, we concluded that all of them contained a lecture, either as an introduction to the ad or as a colophon. Besides, the lecture type could be combined with one of the three main types of structure mentioned above. Indeed, according to Pennock-Speck (2005), "most TV commercials are made up of some kind of mini drama and most, but not all, include a voice-over". I now describe in detail these structures:

- Minidrama: a story with a plot. Plot is defined as all the events in a story. Gustav Freytag, a German dramatist and novelist, considered plot a narrative structure that divided a story into five parts, like the five acts of a play. These parts are: exposition (of the situation); rising action (through conflict); climax (or turning point); falling action; and resolution. Technically it can be composed by a collage of different images or a long take or different takes.

- Montage: different images shown as a collage. Defined at Merriam-Webster as "the production of a rapid succession of images in a motion picture to illustrate an association of ideas". The images can be unrelated and aim to achieve some particular artistic or emotional effect or general theme, or they can be related images of the same person aiming at generating an association of ideas. Montages have an artistic touch and leave room for the spectator to reflect on what is happening before telling them what the advertisement is about. It can be quite problematic to differentiate montage from both mini-dramas and image sequences.

- Image sequence is defined negatively as a sequence of images that do not tell a story with a plot or do not constitute a montage. These images are usually a long take or different takes of the same scene where some activity is unfolding. However, the cause-effect found in mini-dramas is not encountered here.

All these types can also include a demonstration and a testimonial. A demonstration is defined as a scene where we are shown how the product works or what the results of using the product are (for instance, by comparing before and after). When a testimonial is included, it is indicated in the field "Dialogue/Character". The mere application of a skin cream, for example, is not classified as a demonstration.

This gives us the following types of advertisements: Lecture + image sequence, Lecture + minidrama, Lecture + montage, Minidrama + lecture, Montage + lecture, Image sequence + lecture, Lecture + minidrama + demonstration, Lecture + image sequence + demonstration and Lecture + montage+ demonstration. The order of the elements is decided depending on when they appear. If the ad starts with a voice-over, or on-screen text, or this is found very near the start of the ad, then we classify it as a "lecture+". If the voice-ove or on-screen text appears only at the end, it is a mini-drama/montage/image sequence + lecture. We can see the distribution of the types of advertisements in Figure 1:

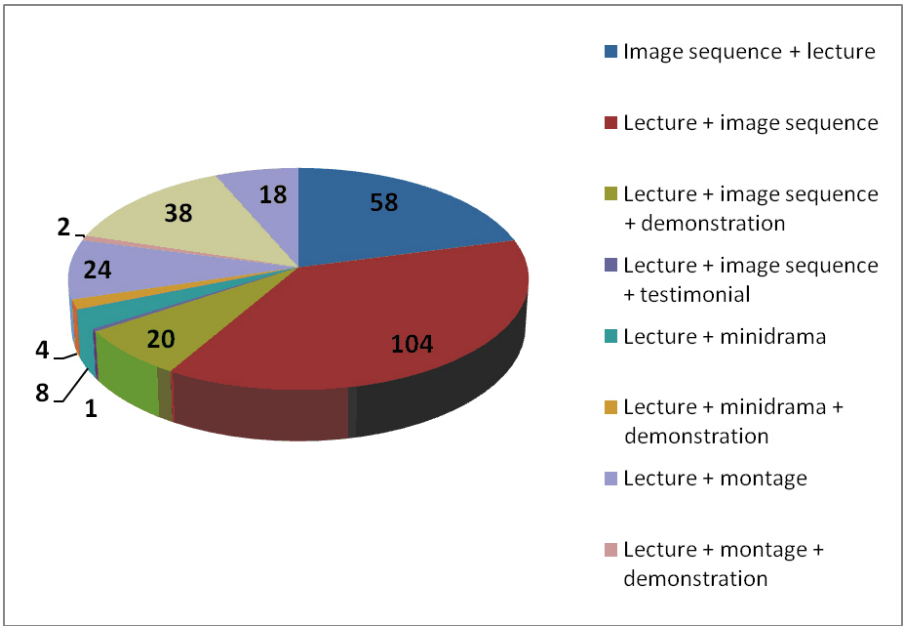


Figure 1: Distribution of the different types of commercials (distinct)

Product classification

In order to classify the advertised products in a structured way, we analysed different product classifications, such as the one used by Eurostat[22] and the International Standard Industrial Classification (ISIC)[23]. Finally, we decided that, for our purposes, the best classification was the one contained in the Karr Collection of Television Commercials, a guide to the television commercials in the Lawrence F. Karr Collection held in the collections of the Motion Picture, Broadcasting, and Recorded Sound Division of the Library of Congress[24]. Since our database is not yet as extensive the Karr Collection, we used a simplified version to reduce the number of product and sub-product types as the list attached in appendix 2 shows.

Organization of the MATVA corpus - Criteria, text selection and classification

One of the main elements of the database is the textual corpus, which is composed of four key elements: the voice-over texts, the transcriptions of the ads, the on-screen texts and song lyrics.

For the elaboration of this corpus we referred to the document "Preliminary recommendations of Corpus Typology" written within the EAGLES initiative by J. Sinclair (1996), where four characteristics that should define a corpus were considered:

22 Available at http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM&StrGroupCode=CLASSIFIC&StrLanguageCode=EN [Last accessed 4/19/2011]

23 Available at http://unstats.un.org/unsd/cr/registry/regct.asp [Last accessed 4/19/2011]

24 Available at http://www.loc.gov/rr/mopic/findaid/karr/karr13.html [Last accessed 03/23/2011]

- Quantity: the current MATVA database is formed by 277 different aids that belong to two full broadcasting days in June 2009. There are four types of elements that conform the corpus: the voice over text, the transcription of the dialogs, on-screen texts and song lyrics. Without taking into account the latter, the voice over text contains 18116 tokens and 1833 types; the transcribed dialogs 12138 tokens and 1157 types, whereas the on-screen texts have 13700 tokens and 1726 types. This makes up a total of 43954 tokens and 3062 types. It is planned to extend the corpus on the short term by including the two days in March, which is in progress.

- Quality: all the material is gathered from real TV ads and can be considered as authentic.

- Simplicity: though the original material is audiovisual, transcriptions are stored in plain text format and annotation has been conceived to be clearly identifiable and separable from text.

- Documented: as we have mentioned before, the corpus is formed by the transcriptions and is well documented with full details about the ads: times, dates, types of voices, accents etc.

With respect to the criteria that need to be considered when designing a corpus, Tourruela & LListerri (1999) define three features: its goal, limits, and the type of corpus. The goal of the MATVA corpus is to serve as a valuable resource for research of linguistic, extralinguistic and paralinguistic aspects of advertisements in the UK. The corpus might also serve as a tool to study the cultural aspects of marketing and advertising. Regarding the limits, it was established that the ads that constitute the corpus would be made up of British televisions gathered randomly during 2009 and would include ads in the English language. That is, the days in march and June have no particular significance but coincided with research visits to the UK. As to the type, our corpus is homogenous considering the texts, since texts are well structured in four blocks (dialogs, voice-over texts, on-screen texts and song lyrics) and all of them belong to a single unit which is the commercial.

Regarding the notion of representativity we consider that it is still too early to analyse this aspect, since as we have mentioned before the corpus is still under construction. Nevertheless, it will be an issue that we will tackle in further stages of this project.

Transcriptions and corpus annotation

The corpus component of the database contains the transcriptions of the oral elements of the advertisements (dialogs, voice-overs, and song lyrics) as well as the on-screen texts. These were made partly by members of the research group and partly by native British informants.

Llisterri (1999) deals with the different proposals for the orthographic transliteration of oral corpora. We generally follow recommendations made by the EAGLES Expert Advisory Group on Language Engineering Standards:

- The words spoken are represented in accordance with standard orthographic conventions.
- The only contractions used are those accepted as standard in the Oxford English Dictionary.
- Sentence boundaries are marked by a full stop and capital letter.
- Direct quoted speech or quotations from written texts are placed in single quotation marks.
- Apostrophes are used in accordance with standard conventions in possessives and in contractions.

Regarding the annotation of the corpus we developed a system of tags to reflect the different communicative agents involved in the advertisement. Tag conventions can be seen in Appendix III.

Technical methodology

With regards to the implementation of spreadsheets and databases, the data were first recorded in a spreadsheet where the different columns represented the fields of the database and each row was a record. Now we are working on a database format and plan to migrate to a platform where the database can be accessed and edited online in order to optimize the workflows. A snippet of the database can be seen in Appendix IV.

Conclusions and future prospects

Television commercials are a complex product made up of different elements: textual, visual, oral and musical. In order to study the effects of all these elements on the audience, we devised the construction of a database with a corpus component in order to enable the analysis of the pragmatic-cognitive effects of these para- and extralinguistic elements on the audience of English TV ads.

We recorded four days of British television broadcasting and designed a database that enabled and facilitated the study and research of the commercials. The construction of the database is still in process and will be extended with further recordings of British and America TV ads. The annotation schema is also under development and will be extended if new needs arise.

Besides, we foresee the creation of a comparable database with Spanish TV ads following the same methodology in order to carry out comparative studies as well as cross-cultural research.

Appendix I: database fields

1. **Ad number**: a unique number that identifies each record of the database.
2. **Ad ID**: a unique ID for each advertisement, that can be repeated more than once.
3. **Date**: date when the advertisement was broadcasted
4. **Time**: time when the advertisement was broadcasted.
5. **Duration**: duration of the advertisement.
6. **Previous Programme**: programme shown before the broadcasting of the advertisement.
7. **Next Programme**: programme shown after the broadcasting of the advertisement.
8. **Product name**: name of the product being advertised.
9. **Product type**: type of the product being advertised.
10. **Sub-product type**: subtype of the product being advertised.
11. **Voice-over text**: boolean value that indicates if there is voice-over text or not.
12. **Gender**: gender of the voice-over text
13. **Age**: age of the voice-over text
14. **Accent**: accent of the voice-over text.
15. **Quality**: quality of the voice-over text.
16. **Type of ad**: type of advertisement according the classification clarified in 4.1
17. **Description**: description of the events happening in the advertisement
18. **Voice over text**: transcription of the voice-over text
19. **Dialogue/Character**: existence of dialogue and character of dialog.
20. **Dialogue:** transcription of the dialogue
21. **On-screen Text**: transcription of the on-screen text.
22. **Song lyrics:** transcription of the song lyrics if they have been written for the advertisement or name of the song.
23. **Observations**: notes

Appendix II: list of product types based on the Kerr classification

- Household appliances and tools
- Electronic goods
- Computer and office
- Automobiles and Automobile products
- Clothing and shoes
- Entertainment
- Music, DVD & board games
- Toys & games
- Baby products
- Food
- Alcoholic Beverages
- Non-Alcoholic Beverages
- Household and garden
- Housewares
- Medicine and Medical products
- Personal products: hygiene, beauty, perfumes, etc.
- Pet products
- Publications
- Insurance and financial services
- Telecommunications
- Public transport
- Retailers, wholesalers
- Restoration and accommodation
- Charities, NGOs, public service announcements, party-political broadcasts, tourism
- Jewellery, watches, bags and accessories

APPENDIX III: TAG CONVENTIONS

Table 1: MATVA Tag Conventions

1st Level: Species								
Human (hu)	Animal (aa)	Object (ob)	Puppet (pu)	Undefined (un)				
	2nd Level: Gender							
	Female (fe)	Male (ma)	Girl (gr)	Boy (bo)				
		3rd Level: age						
		Mature (mt)	Young (yo)	Teenage (tn)	baby (bb)			
			4th Level: Role					
			Mother (mo)	Father (fa)	Child (ch)	Daughter (da)	Son (sn)	
			Friend (fr)	Sister (si)	Brother (br)	Partner (pt)	Worker (wk)	
			Boss (bs)	Testimonial (ts)	celebrity (cl)	Product Character (pc)		
				5th Level: Origin (optional)				
				Name of the Region and Country				

One level can have more than one value and will be separated by "-"

The levels are separated by "/"

If one level stays emtpy, write "-"

If one level needs specification, put it in bracktets

If there are various characters in a conversation, number them at the end with brackets

Ex: young female worker --> hu/fe/yo/wk

Ex: puppet male mature --> pu-hu/ma/mt/-

Ex: puppet dog mature --> pu-aa(dog)/-/mt/-

Ex: mature male father --> hu/ma/mt/fa(1)

Ex: young woman testimonial celebrity --> hu/fe/yo/ts-cl(Koby Bryant)

Ex: teenage girl from USA --> hu/gr/tn/-/North America-USA

VO: before a voice-over

OST: before on screen text

SL: before song lyrics

OSP: On screen presenter

Tags will be bracketed with <> signs for a better localisation when searching text patterns.

<ts/hu/fm/mt/nc1:> Keeping an eye on your weight? Bitesize shredded wheat. <ts/hu/fm/yn/nc2:> 100% wholegrain wheat. That's all. That's all that's in it. <ts/hu/fm/yn/nc 3:> There's nothing else in there. <ts/hu/fm/yn/nc 4:> It's great for watching what I'm eating. <ts/hu/fm/mt/nc1:> Very good for keeping an eye on your weight.

<ch/hu/ml/tn/of:> Mum, look at the state of this! What am I going to do? <ch/hu/fm/mt/pt:> Wear the white one. <ch/hu/ml/tn/of:> The white one? I'll wash it myself. <ch/hu/fm/mt/pt:> Bye bye. <ch/hu/ml/tn/of:> Oh Mum!

Figure 2: Example of corpus annotation

APPENDIX IV: MATVA DATABASE

34	NHSQuitSmoking01	06/24/09	08.00-12.30	5,00	GMTV	GMTV	NHS Quit smok	Charities, NGC		Yes	M	25 to 40	RP	serious	Lecture + imag	We see a still p	For
35	Baker'sCompleteDogFooc	06/24/09	08.00-12.30	30,00	GMTV	GMTV	Baker's Compl	Pet products		Yes	M	25 to 40	North	authoritative/l	Lecture + minii	We see three c	You
36	OlayRegeneristEyeDerma	06/24/09	08.00-12.30	30,00	GMTV	GMTV	Olay Regeneris	Medicine and I	Personal prodi	Yes	F	25 to 40	RP+	husky/sensual	Lecture + imag	We can see a v	Nev
37	OptimaxLaserEyeTreatme	06/24/09	08.00-12.30	30,00	GMTV	GMTV	Optimax Laser	Medicine and I		Yes	M	25 to 40	RP+	assertive/happ	Lecture + mon	We hear two d	Hav
38	LorealExcellenceCremeFi	06/24/09	08.00-12.30	20,00	GMTV	GMTV	L'Oreal Excelle	Personal prodi		Yes	M	25 to 40	RP		Lecture + imag	We see Andie	Exc
39	Holland&BarrettGlucosan	06/24/09	08.00-12.30	20,00	GMTV	GMTV	Holland & Barr	Medicine and I		Yes	F	25 to 40	RP	flirty/happy	Lecture + imag	We see a bottl	At I
40	BazukaExtraStrengthWart	06/24/09	08.00-12.30	20,00	GMTV	GMTV	Bazuka Extra S	Medicine and I		No	novo	novo	novo	novo	Lecture + imag	A woman talks	
41	SomerfieldStrawberries0	06/24/09	08.00-12.30	10,00	GMTV	GMTV	Somerfield Str	Retailers, who	Food	Yes	F	Over 40	South-West	assertive/happ	Lecture + imag	We see a wom	Hav
42	PGTipsDancingInKitchen0	06/24/09	08.00-12.30	10,00	GMTV	GMTV	PG Tips (Danci	Non-Alcoholic		Yes	M	25 to 40	North	happy/husky	Image sequen	A man is danci	PG
43	NHSQuitSmoking02	06/24/09	08.00-12.30	5,00	GMTV	GMTV	NHS Quit smok	Charities, NGC		Yes	M	25 to 40	RP	serious	Lecture + imag	We see a still p	For
44	MatalanFamilyOnBeachN	06/24/09	08.00-12.30	5,00	GMTV	GMTV	Matalan (Fami	Retailers, who	Clothing and s	No	novo	novo	novo	novo	Image sequen	We see a famil	
45	PhiladelphiaFridgeLikeBc	06/24/09	08.00-12.30	10,00	GMTV	GMTV	Philadelphia (I	Food		Yes	F	25 to 40	RP	flirty/husky/se	Lecture + imag	We see the ch	Phi
46	LloydsPharmacyHayFever	06/24/09	08.00-12.30	30,00	GMTV	GMTV	Lloyds pharma	Medicine and I		Yes	F	25 to 40	North	assertive/happ	Image sequen	We can see a v	Hay
47	SMAManHoldsBaby01	06/24/09	08.00-12.30	20,00	GMTV	GMTV	SMA (Man hol	Baby products	Food	Yes	F	25 to 40	RP	breathy/seriou	Image sequen	A man holding	Unc
48	HillarysShuttersRedBlinds	06/24/09	08.00-12.30	30,00	GMTV	GMTV	Hillarys (Shutt	Housewares		Yes	F	Up to 25	RP	assertive/play	Lecture + mon	We see the na	Sav
49	DellInspironPinkVespa01	06/24/09	08.00-12.30	30,00	GMTV	GMTV	Dell Inspiron.	Computer and		Yes	M	25 to 40	RP	assertive/opti	Lecture + mon	Some moving c	For
50	AdiosDancingInStreet01	06/24/09	08.00-12.30	15,00	GMTV	GMTV	Adios (weight	Medicine and I		Yes	F	Up to 25	RP+	assertive/happ	Image sequen	A girl wearing	Wh
51	DFSSofasHalfPrice01	06/24/09	08.00-12.30	30,00	GMTV	GMTV	DFS-sofas	Retailers, who	Housewares (e	Yes	F	Up to 25	RP+	assertive/opti	Lecture + mon	A young man s	If y
52	RonsealPaint01	06/24/09	08.00-12.30	10,00	GMTV	GMTV	Ronseal (paint	Household and		No	novo	novo	novo	novo	Lecture + imag	A man opens a	
53	Arm&HammerToothpaste	06/24/09	08.00-12.30	20,00	GMTV	GMTV	Arm and Hamn	Medicine and I		No	novo	novo	novo	novo	Lecture + imag	A young woma	
54	PGTipsDancingInKitchen0	06/24/09	08.00-12.30	10,00	GMTV	GMTV	PG Tips (Danci	Non-Alcoholic		Yes	M	25 to 40	North	happy/husky	Image sequen	A man is danci	PG
55	MatalanFamilyOnBeachN	06/24/09	08.00-12.30	3,00	GMTV	GMTV	Matalan (Fami	Retailers, who	Clothing and s	No	novo	novo	novo	novo	Image sequen	We see a famil	
56	BuxtonBottledWaterRefl	06/24/09	08.00-12.30	20,00	GMTV	GMTV	Buxton (Bottle	Non-Alcoholic		Yes	M	Up to 25	RP+	cheeky/happy	Image sequen	Image of a dro	Bux
57	Sky+Box01	06/24/09	08.00-12.30	40,00	GMTV	GMTV	Sky + Box	Telecommunic	Electronic goo	Yes	F	25 to 40	RP	assertive/happ	Lecture + imag	We see the pla	To I
58	MacMillanCancerClock01	06/24/09	08.00-12.30	30,00	GMTV	GMTV	MacMillan Can	Charities, NGC		Yes	M	25 to 40	North	deep/serious	Lecture + mon	We see the ref	Wh
59	LurpakButterSaturdayMo	06/24/09	08.00-12.30	30,00	GMTV	GMTV	Lurpak butter (	Food		Yes	M	Over 40	USA	deep/husky/se	Lecture + mon	We can see so	Sat
60	ShreddedWheatBiteSizeC	06/24/09	08.00-12.30	30,00	GMTV	GMTV	Shredded whe	Food		Yes	F	25 to 40	RP+	assertive/opti	Lecture + mon	A woman asks	And
61	PersilYoungMan01	06/24/09	08.00-12.30	40,00	GMTV	GMTV	Persil (Young r	Household and		Yes	M	Over 40	RP	authoritative/	Minidrama + le	A young boy h	Per
62	SafestyleWindows&Door	06/24/09	08.00-12.30	30,00	GMTV	GMTV	Safestyle (win	Housewares		Yes	M	25 to 40	RP+	assertive/happ	Lecture + imag	We first see th	List
63	SkyHD01	06/24/09	08.00-12.30	10,00	GMTV	GMTV	Sky HD (short	Telecommunic	Electronic goo	Yes	F	25 to 40	RP	assertive/opti	Lecture + imag	We can see a c	Got
64	MatalanFamilyOnBeachV	06/24/09	08.00-12.30	10,00	GMTV	GMTV	Matalan (VO v	Retailers, who	Clothing and s	Yes	F	25 to 40	RP	assertive/opti	Image sequen	We see a famil	GM
65	PampersNappiesMatureV	06/24/09	08.00-12.30	30,00	GMTV	GMTV	Pampers (nap	Baby products	Personal prodi	Yes	F	25 to 40	RP+	breathy/husky	Lecture + mini	We see a wom	You
66	DettolSurfaceCleanserNe	06/24/09	08.00-12.30	30,00	GMTV	GMTV	Dettol Surface	Household and		Yes	F	25 to 40	RP+	husky/serious	Lecture + imag	A needle pierc	Fac
67	InjuryLawyers4UDogTrack	06/24/09	08.00-12.30	30,00	GMTV	GMTV	Injurylawyers4	Insurance and		No	novo	novo	novo	novo	Lecture + imag	We see a dog r	
68	ThinkBingoCucumber01	06/24/09	08.00-12.30	15,00	GMTV	Jeremy Kyle	Thinkbingo.co	Entertainment	Telecommunic	Yes	F	Up to 25	North	assertive/happ	Image sequen	We see a wom	For
69	ThinkBingoHeadRisingFrc	06/24/09	08.00-12.30	10,00	Jeremy Kyle	Jeremy Kyle	Thinkbingo.co	Entertainment	Telecommunic	Yes	F	Up to 25	North	assertive/happ	Image sequen	We see a lot o	Get
70	NationalAccidentHelpline	06/24/09	08.00-12.30	30,00	Jeremy Kyle	Jeremy Kyle	National Accid	Insurance and		No	novo	novo	novo	novo	Lecture + imag	We see helme	
71	ButlinsFreeChildPlaces01	06/24/09	08.00-12.30	30,00	Jeremy Kyle	Jeremy Kyle	Butlins (Free C	Entertainment		Yes	F	25 to 40	RP+	assertive/happ	Lecture + imag	We see a famil	Dor

Table 2: Snippet of the MATVA Database

References

Hoven, H. A. (2009). *Narrative television commercials as a route to persuasion*. Master thesis: VU University of Amsterdam.

Kassarjian, H. H. (1977). Content Analysis in Consumer Research. *Journal of Consumer Research*, *4*(1), 8-18. doi: 10.1086/208674.

Llisterri, J. (1996). Preliminary Recommendations on Spoken Texts. Retrieved from http://www.ilc.cnr.it/EAGLES96/spokentx/spokentx.html.

Llisterri, J. (1999). Transcripción, etiquetado y codificación de corpus orales. *Panorama de la investigación en lingüística informática. RESLA, Revista Española de Lingüística Aplicada, Volumen monográfico*, 53-82.

Luna, D., & Gupta, S. F. (2001). An integrative framework for cross-cultural consumer behavior. *International Marketing Review*, *18*(1), 45-69. doi: 10.1108/02651330110381998.

Pennock-Speck, B. (2005). Styling the voice, selling the product. *Proceedings of the 4th international contrastive linguistics conference*, 973-980.

Pennock-Speck, B., & Del Saz Rubio, M. M. (2009). Constructing female identities through feminine hygiene TV commercials. *Journal of Pragmatics*, *41*, 2535–2556.

Sinclair, J. (1996). *EAGLES Preliminary recommendations on Corpus Typology*. Retrieved from http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.htm.

Stern, B. B. (1994). Classical and Vignette Television Advertising Dramas: Structural Models, Formal Analysis, and Consumer Effects. *Journal of Consumer Research*, *20*(4), 601-615.

Torruella, J., & Llisterri, J. (1999). Diseño de corpus textuales y orales. In J. M. Blecua, G. Clavería, C. Sánchez, & J. Torruella (Eds.), *Filología e informática. Nuevas tecnologías en los estudios filológicos* (pp. 45-77). Editorial Milenio y Universidad Autónoma de Barcelona.

Wells, W. D. (1989). Lectures and Dramas. In P. Cafferata & A. M. Tybout (Eds.), *Cognitive and affective responses to advertising*. Lexington (MA): Lexington Books.

Design and development of the Bulgarian sense-annotated corpus

Koeva, Svetla

Leseva, Svetlozara

Rizov, Borislav

Tarpomanova, Ekaterina

Dimitrova, Tsvetana

Kukova, Hristina

Todorova, Maria

Institute for Bulgarian Language – Bulgarian Academy of Sciences

This paper describes the methodology of compilation and annotation of the Bulgarian Sense-Annotated Corpus - a manually annotated corpus of over 100,000 words in which each lexical unit (LU) is assigned a sense according to the Bulgarian wordnet. The paper gives a brief outline of the corpus representation, the functionalities of the annotation tool Chooser, and sketches the linguistic conventions and practical considerations adopted in the process of corpus annotation. Finally, the paper describes one of the major applications of the Bulgarian Sense-Annotated Corpus as a training corpus for a word-sense disambiguation system for Bulgarian.

Keywords: design and development of sense-annotated corpora, semantic annotation, word-sense disambiguation

El artículo describe la metodología de la compilación y la anotación del Corpus Semánticamente Anotado Búlgaro - un corpus anotado de modo manual y consta de más de 100 mil palabras donde en cada unidad lingüística se le ha atribuido un significado conforme al Wordnet Búlgaro. El artículo presenta, asimismo, el programa de anotación Chooser. Han sido descritas las convenciones lingüísticas y las soluciones prácticas adoptadas en el proceso de la anotación. Al final, el artículo describe una de las aplicaciones esenciales de BulSemCor como un corpus de entrenamiento orientado a desarrollar el sistema de desambiguación de la lengua búlgara.

Palabras clave: diseño y desarrollo de corpus anotados semánticamente, anotación semántica, desambiguación semántica

1. Introduction

The Bulgarian Sense-Annotated Corpus *BulSemCor* (Koeva, Leseva, Tarpomanova, Rizov, Dimitrova, & Kukova, 2010) is compiled according to the general methodology established by the *SemCor project* (Landes, Leacock & Tengi, 1998). It is a subset of the Brown Corpus of Bulgarian (BCB) semantically annotated with a corresponding synonym set (synset) in the Bulgarian wordnet *BulNet* (Koeva, 2010a). Unlike the bulk of sense-annotated corpora where only (sets of) content words are annotated (Ng & Lee, 1996; Pianta, & Bentivogli, 2003; Wu, Jin, Zhang & Yu, 2006, to mention but a few), in BulSemCor each LU has been assigned a sense. Section 2 gives an overview of the input corpus, its structure and the format required by the annotation tool (presented in Section 3). Section 4 discusses the principles of annotation, along with the linguistic assumptions and practicalities behind the annotation process. Section 5 sketches the results, while Section 6 focuses on the application of the corpus for word sense disambiguation (WSD).

2. Input corpus – Structure and representation

The annotation corpus consists of two BCB subsets with an overall 811 text units of 100+ words each, adding up to 101,062 tokens. The first subset was selected according to the density of highest frequency open-class lemmas with heuristics applied to provide a balance between different parts of speech and a better coverage of lemmas. The second one was sampled according to the density of open-class lemmas not included in the BulNet at that time with priority given to polysemous lemmas (Koeva, 2010b:16).

The corpus is represented in a flat xml format. The text is encoded as a list of xml tags labeled *word.* The relevant information is stored in separate attributes: word form (“w”), lemma (“l”), sense (“s”), annotator (“u”), time stamp (“t”), sentence end (“e”) (Rizov, 2010: 43). A special attribute is reserved for a parent ID that links the individual tokens of a compound (“p”). An annotated unit contains the following basic information: *<word l=”замова» p=»-1529023764» s=»1100001720» t=»1298483182» w=»замова»/>*.

Minimum restrictions are imposed on the extension of the specified file format, so that it permits addition of flat and/or hierarchical annotation schemata without affecting the current one, thus enabling other levels of annotation.

3. Annotation tool

Chooser is an OS independent multi-functional system for linguistic annotation (Figure 1), adaptable to annotation schemata for different language levels. The basic *annotation functionalities* are: (i) fast and easy-to-perform selection; (ii) run-time access to information for the candidate senses such as definition, frequency, the associated wordnet synsets with all the pertaining info – synonyms, gloss, semantic relations, notes on usage, form, etc.; (iii) identification of MWEs with contiguous and non-contiguous constituents and supplying information for them at run-time. The tool provides a number of *input data editing* functionalities including (i) editing of word forms; (ii) editing of lemmas; (iii) insertion and deletion of tokens. The basic functions are enhanced with *flexible text navigation strategies* - forward and backward navigation over: (i) all words; (ii) non-

annotated words; (iii) all instances of a word; (iv) all instances of a sense. Finally, a *flexible search strategy* allowing both exact match search according to word form or lemma, and regular expression search is integrated.

The tool interface features a fully-fledged visualization of the BulNet synsets for the candidate senses available for a selected LU through coupling with the system for wordnet development and exploration Hydra. A unified BulNet representation in Chooser and Hydra is implemented with the purpose of facilitating both the annotation process and the parallel enlargement of the wordnet.

Chooser provides multiple-user concurrent access and dynamic real-time update in the knowledge base, so that all changes, such as newly-encoded synsets, literals, relations, are updated in both systems and made available to all the users immediately (Rizov, 2010).

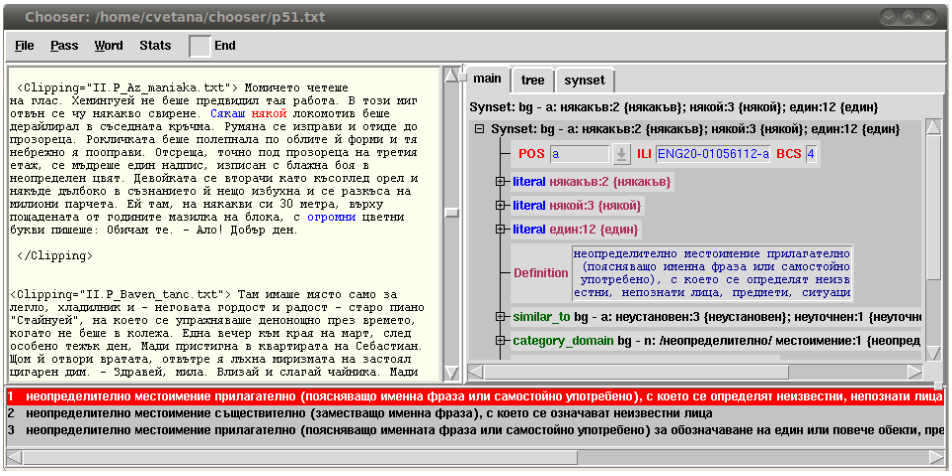


Figure 1. Chooser – *left-pane* – the annotated corpus unit with the current LU in red; *bottom pane* – list of the candidate senses for the current LU with the selected one in red; *right-hand pane* – the synset view for the selected sense.

4. Annotation

4.1. *Levels of annotation*

The first stage in the annotation process is automatic lemmatization followed by manual post-editing. Through lemmatization a two-fold purpose is accomplished: (i) association of word forms with a canonical form is ensured (i.e. *morphological annotation*); (ii) each LU in BulSemCor is mapped to all the synsets in BulNet that feature a literal (synset member) with an identical lemma (Koeva, 2010b: 9). Thus, the relevant synset senses and all the pertaining information are associated with the words in the corpus.

BulSemCor also comprises *morpho-syntactic annotation* – each LU is assigned the POS tag of the synset it is annotated with.

Syntactic annotation has not been consistently applied. Partial information for the dependencies between compounds' elements is available through syntactic head marking of compounds in BulNet (Koeva, 2010b: 10).

Semantic annotation proper consists in the association of a LU in BulSemCor with the most appropriate synset in BulNet. The annotated item is associated with all the linguistic information in the synset, including synonyms, gloss, POS value, the semantic, morpho-semantic and extralinguistic relations pertaining to the synset, the semantic and derivational relations pertaining to the literal, etc. The BulNet synsets are associated with their equivalents in Princeton WordNet (PWN) through unique identifiers (IDs). In such a way, the annotated LUs are mapped to their translation equivalents in English and, with PWN serving as a hub, to all the other wordnets.

4.2. *Methodology*

The annotation process involves two major tasks: (i) defining the boundaries of the LUs in the corpus; (ii) choosing the most appropriate sense for a LU from a list of candidates.

Annotation is applied to: (i) *single words* – formed by one or by two or more stems; (ii) *MWEs* –(dis)continuous sequences of two or more single words, separated by space(s).

While single word identification is on the whole unproblematic, automatic lemmatization is not always straightforward, as different senses of a word may have different canonical forms, consider *mladi,* a plural form of the adjective *young*, which functions as a collective noun: *{mladezh:3, mladi:1, mladi hora:1, mladost:3}* ({young:7; youth:2} - "young people collectively"). MWEs pose a number of challenges with respect to: (i) *delimitation* - the boundaries of a MWE are not necessarily straightforward, consider *contextual ellipsis, reduced and expanded variants*, etc.; (ii) *lemma definition* – the form of the individual elements in the MWEs may differ from their canonical forms as single words; (iii) *morpho-syntactic value* – the POS of a MWE may not be the same as the POS of the head; (iv) *semantic value* – the concept expressed by a MWE may not be the sum total of the concepts expressed by its elements.

The appropriateness of candidate senses is assessed according to: (i) interchangeability of the LU in the corpus with the rest of the synonyms in the synset; (ii) appropriateness of the definition; (iii) the position of the synset in the wordnet structure.

In cases where no appropriate candidate is found among the list of BulNet synsets one checks whether a synset denoting a relevant sense exists in PWN, and, if so, it is encoded in BulNet; otherwise, a new BulNet-unique synset is created, as with most closed-class words (see Section 4.3) and language and culture specific words.

Certain numerical and time expressions that do not constitute lexemes are annotated with a corresponding ontological category, such as *date, number*, etc.

4.3. Linguistic conventions

PWN synsets are distributed into four classes – nouns, verbs, adjectives and adverbs – which roughly correspond to the relevant parts of speech. While the morpho-syntactic categorization is on the whole unproblematic, certain asymmetries between English (as represented in PWN) and Bulgarian (BulNet) needed to be accommodated in the process of encoding and annotation.

Numerals: According to the classification principles adopted, numerals are treated as either nouns or adjectives, as follows: *cardinal numerals* are encoded as synsets with a POS value "adjective" (POS: a), when denoting quantity of some entity, and with a POS "noun" (POS: n) when denoting numbers. This distinction is also made in BulSemCor, e.g. in *dve godini* "two years", *dve* "two" is annotated with the adjective synset, while in *chisloto 7* "number 7" *7* is annotated as a noun. *Ordinal numerals* are mostly classified as adjectives.

Adverbials: Certain words and expressions categorized in Bulgarian as adverbials, receive treatment as adjectives or as adverbs in English depending on the type of phrase they modify – NPs or VPs, and are encoded in PWN accordingly. The PWN distinction is preserved in BulNet, with a synset note (Snote) added to the adjective synset marking its POS as "adverb" (POS: b). Respectively, the expression *na zhivo* "live" is annotated with an adjective synset in *predavane na zhivo "*live performance", and with an adverb synset in *predavat na zhivo* "(they) perform live".

The following conventions regarding POS categorization were adopted in both PWN and BulNet.

Adjective-like participles: Participles used as adjectives are encoded as synsets with a POS tag "adjective". The BulNet entry encodes a Snote marking them as participles.

Substantives: *Substantivized adjectives* and *participles* are encoded as synsets with a POS tag "noun". The BulNet entry encodes a Snote that marks them as substantives.

BulSemCor has necessitated the expansion of BulNet with closed-class words resulting in an enlarged lexical-semantic net that covers all the 10 traditional parts of speech. To the best of our knowledge, this is the only wordnet to cover the function word classes. The following classes have been included – *pronouns* (POS: pron), *prepositions* (POS: p), *coordinating* and *subordinating conjunctions* (POS: conj), *particles* (POS: particle), *interjections* (POS: ij). Most of the newly-encoded synsets are provided with English translation equivalents in the Snote node. Function words are integrated into the wordnet structure through the *[category_domain]* relation pointing to the synset denoting the relevant category: *{preposition}*, *{coordinating conjunction}*, *{subordinating conjunction}, {particle}*. For pronouns this is the pronoun type *{personal pronoun}, {possessive pronoun}*, etc.

Some function words have nevertheless been encoded in PWN as members of synsets pertaining to one of the content word classes. For instance, certain pronouns are included as adjectives, nouns, or adverbs depending on the part of speech they substitute for, some adverbial conjunctions are encoded as adverbs, etc. In these cases the corresponding

BulNet synsets explicitly indicate the POS value according to the traditional grammatical conventions by means of a Snote.

5. Results

The main results achieved in the work on BulSemCor include: definition of an annotation schema, implementation of an annotation tool, elaboration of a methodology and annotation conventions, compilation of an input corpus, development of a sense-annotated corpus, BulNet enlargement (for an overview cf. Koeva, 2010b: 7-42).

Some of the quantitative parameters of BulSemCor are presented below - the overall number of tokens and annotated LUs, along with their distribution into single words and MWEs (Table 1), as well as the distribution of annotated lexical units according to POS (Table 2):

Table 1. Overall numbers of tokens and annotated units

Total number of tokens	Annotated words	Single unique tokens	Annotated single words	Annotated MWEs	Unique tokens in MWEs
101062	99480	88196	86842	5797	12866

Table 2. POS distribution of annotated LUs

POS	Nouns	Verbs	Adj	Adv	Preps	Conj	Pron	Part	Interj
Number	31058	17041	12012	7935	14772	7265	6810	2570	17

6. Applications

The principal application of BulSemCor is in the training and evaluation of a multi-component WSD system. At present five independent "weak" classifiers and an ensemble one (combining all of them) are used in the disambiguation. Each classifier provides a confidence distribution over the senses for a particular single word or MWE (lists of pairs: <sense, confidence>, where the sum of the confidences is 1, are generated). Two knowledge-based classifiers based on the *Lesk* (Lesk, 1986) and the *Degree* (Navigli, & Lapata, 2010) algorithms, respectively - are employed to disambiguate LUs using information encoded in BulNet and the context of the LU in the corpus. Lesk determines the best choice using the overlap between the context of a LU and the bag of words for every possible sense available for it. The Degree algorithm creates a subgraph of the knowledge-base relational structure (in our case wordnet) according to the context of a LU. The senses receive confidence values according to their vertex 'degree', e.g. the arcs connected with it.

Two Hidden Markov Model disambiguators - one for forward and one for backward processing of the sequences in the text – are also used (Rabiner, 1989). The training data

is insufficient to provide every possible sequence of words and senses. Even though smoothing with information from BulNet has been applied, the searching algorithm of the optimal senses (Viterbi) needs to be restarted. Thus, a second (backward) HMM is provided (for a well-trained model the forward and backward HMMs will be identical). The fifth weak classifier assesses the confidence for a particular sense according to its frequency in BulSemCor. The ensemble classifier uses a weighted sum of the five weak ones.

BulSemCor was split into three parts in proportion 2:1:1. The largest one is reserved for training – generation of the HMM, tuning of the Lesk and Degree classifiers, etc. The second segment serves in the estimation of the weights of the ensemble classifier for each unique LU in the corpus – every weak disambiguator receives confidence value for a given LU according to its performance in the training. On account of data sparseness words that are not attested must be processed according to the overall confidence distribution (computed on the basis of all the words). The last portion of the corpus is used for evaluation of the system's performance. The current version outperforms the calculated random sense baseline (~40%) by 24 points with an overall precision of ~65%. The ensemble disambiguator shows a good overall improvement in terms of precision outperforming the best of the weak classifiers by approximately 5 points (~65% vs. ~60%). Although some of the algorithms process only part of the words in a given text, the coverage of the system is near 100%.

7. Conclusion

The WSD system is going to be employed in the annotation of the Bulgarian National Corpus (Koeva, Blagoeva, & Kolkovska, 2010) – a large corpus of approximately 320 million words. The annotation is to be performed automatically, with possible manual post-editing. One of the major applications will be in the further improvement of the system's performance.

8. Acknowledgements

This paper has been supported by the European Social Fund 2007-2013, Human Resources Development Operational Programme, within the Operation *Support to the development of PhD students, post-doctoral students, post-graduate students and young scientists* (BG051PO001-3.3.04), grant No. BG051PO001-3.3.04/27 of 28 August 2009 – *Mathematical Logic and Computational Linguistics: Development and Permeation*.

References

Koeva, S. (2010a). Bulgarian Wordnet - Current State, Applications and Prospects. In *Bulgarian-American Dialogues* (pp. 120-132). Sofia: Prof. Marin Drinov Academic Publishing House.

Koeva, S. (2010b): Balgarskiyat semantichno anotiran korpus – teoretichni postanovki. In S. Koeva (Ed.), *Balgarskiyat semantichno anotiran korpus* (pp. 7-42). Sofia: Institute for Bulgarian Language.

Koeva, S., Leseva, S., Tarpomanova, E., Rizov, B., Dimitrova, T., & Kukova, H. (2010). The Bulgarian Sense-Annotated Corpus – Results and Achievements. In M. Tadic, M. Dimitrova-Vulchanova & S. Koeva (Eds.), *Proceedings of the FASSBL-7 Conference* (pp. 41-48). Zagreb.

Koeva, S., Blagoeva, D., & Kolkovska, S. (2010). Bulgarian National Corpus Project. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner & D. Tapias (Eds.), *Proceedings of the Seventh Conference on International Language Resources and Evaluation (*pp. 3678-3684). La Valletta, Malta.

Landes, S., Leacock, C., & Tengi, R. (1998). Building Semantic Concordances. In C. Fellbaum (Ed.) *Word-Net: An Electronic Lexical Database* (pp. 199-216). Cambridge, Mass.: MIT Press.

Lesk, M. (1986). Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In V. DeBuys (Ed.), *Proceedings of the 5th Annual Conference on Systems Documentation* (pp.24-26), Toronto, Ontario, Canada.

Navigli, R., & Lapata, M. (2010). An Experimental Study on Graph Connectivity for Unsupervised Word Sense Disambiguation. In *IEEE Transactions on Pattern Anaylsis and Machine Intelligence*, 32(4) (pp. 678–692).

Ng, H.T., & Lee, H.B. (1996). Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 40-47).

Pianta, E., & Bentivogli, L. (2003). Translation as Annotation. In *Proceedings of the AIIA 2003 Workshop "Topics and Perspectives of Natural Language Proceesing in Italy"* (pp. 40-48), Pisa, Italy.

Rabiner, L. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Proceedings of the IEEE 77* (2) (pp. 257–286).

Rizov, B. (2010). Sistema za anotirane Chooser. In S.Koeva (Ed.), *Balgarskiyat semantichno anotiran korpus* (pp. 51-64). Sofia: Institute for Bulgarian Language.

Wu, Y., Jin, P., Zhang, Y., & Yu, S. (2006). A Chinese Corpus with Word Sense Annotation. In *ICCPOL* (pp.414-421).

Designing a dependency representation and grammar definition corpus for Finnish

Atro Voutilainen, Krister Lindén, Tanja Purtonen

Department of Modern Languages, University of Helsinki

We outline the design and creation of a syntactically and morphologically annotated corpora of Finnish for use by the research community. We motivate a definitional, systematic "grammar definition corpus" as a first step in a three-year annotation effort to help create higher-quality, better-documented extensive parsebanks at a later stage. The syntactic representation, consisting of a dependency structure and a basic set of dependency functions, is outlined with examples. Reference is made to double-blind annotation experiments to measure the applicability of the new grammar definition corpus methodology.

Parsebank, grammar definition corpus, dependency grammar

Presentamos el primer diseño y creación de un corpus del finlandés anotado sintáctica y morfológicamente para su uso por la comunidad científica. En este trabajo se motiva un "corpus de definición gramatical" sistemático y que servirá como base para un proyecto de anotación de tres años, como ayuda para la creación de corpus anotados sintácticamente (treebanks o parsebanks) amplios, de mejor calidad y mejor documentados en una fase subsiguiente. La representación sintáctica, consistente en una estructura de dependencias y un conjunto básico de funciones de dependencia, es presentada con ejemplos. En este trabajo se hace referencia a los experimentos de anotación doblemente ciegos (double-blind) para medir la aplicabilidad de la nueva metodología para el corpus de definición gramatical.

1. Background

This paper outlines the first main step - motivation and design of a grammar definition corpus - in a multiyear project at University of Helsinki (as part of the pan-European CLARIN research infrastructure effort) to provide (i) open-source morphological and dependency syntactic language models and analysers for the Finnish language and (ii) publicly available morphologically and dependency syntactically annotated large text corpora of Finnish (e.g. Finnish Wikipedia and EuroParl corpora) for R&D uses in Finland and other countries.

More specifically, we outline an effort to create a **grammar definition corpus** and related documentation of linguistic descriptors ("stylesheet") of Finnish. This corpus consists of 19,000 example sentences extracted from a comprehensive descriptive Finnish grammar (Hakulinen, Vilkuna, Korhonen, Koivisto, Heinonen & Alho, 2004), and annotated according to a linguistic representation (a morphological and dependency syntactic grammar with a basic dependency function palette). To our knowledge, this effort if the first one based on a comprehensive, systematic set of sentences illustrating the syntactic structures of a natural language in considerable depth. This grammar definition corpus will be used as a basis for creating and documenting (i) formal language models and parsers for use in automatic corpus annotation and (ii) large syntactically annotated text corpora for R&D related to the Finnish language.

The structure of this paper is as follows. Section 2 discusses the terms "treebank", "parsebank" and "grammar definition corpus". Section 3 outlines descriptive solutions related to Finnish language analysis. Section 4 focuses on the dependency syntactic representation used in the grammar definition corpus. Section 5 tells about the work process and deliverables.

2. Treebank, Parsebank, Grammar Definition Corpus

A Treebank can be described as a set of sentences syntactically annotated by trained linguists. A hand-annotated Treebank is restricted in size, of high annotation quality and consistency, and represents running text sentences and/or selected sentences illustrating various syntactic structures of the language. The PARC 700 Dependency Bank is a good example of a manually annotated Treebank, with a set of 700 text sentences annotated manually according to a form of Lexical Functional Grammar (King, Crouch, Rietzler, Dalrymple & Kaplan, 2003). Far larger annotated resources of English are documented in (Cinková, Toman, Hajič, Čermáková, Klimeš, Mladová, Šindlerová, Tomšů & Žabokrtský, 2009; Marcus, Santorini & Marcinkiewicz, 2004). Additionally, Wikipedia ("Treebank") lists a large number of treebank projects for many languages.

A Parsebank can be characterized by a large amount of sentences that have been mechanically annotated (with a parser), and the annotating parser has repeatedly been modified by sampling the output to correct mistakes and gradually create a better Parsebank.

In order to create a high-quality Parsebank, we need documentation and examples on the linguistic representation and its use in text analysis. A hand-annotated set of sentences is

useful, but in order to approximate the structures that are used in a large corpus of text in a more comprehensive and systematic way, we need a more exhaustive and systematic set of sentences to be analysed and documented e.g. as a guideline for creating a Parsebank. We use a large descriptive grammar as a source of example sentences to reach a high and systematic coverage of the syntactic structures in the language. A hand-annotated, cross-checked and documented collection of such a systematic set of sentences – in short, *a Grammar definition corpus* – serves as an inventory of high and low frequency syntactic constructions in the language.

However, sample sentences in a descriptive grammar usually are kept as simple and short as is convenient for illustrating the grammatical construction in point. To start approximating the variation possibilities within each grammatical construction, additional running-text corpora from different genres are needed for annotation – but following the guidelines set at the definitional phase.

3. Finnish in outline

Morphology. Finnish has a rich inflectional system with thousands of forms for each verb, adjective and noun. Some combinations clearly have a special function and the need for reducing these to a single base form is more a question of how useful the connection with the valency or frame information of the base form is.

One of the tasks of morphology is to provide the inflected words with base forms and a set of morphological tags. If the word in non-inflecting or has a deficient paradigm, we have opted for the form given by the descriptive grammar (Hakulinen *et al*., 2004) .

Participles can in general be formed from all verbs, so one natural form for participles is the base form of the corresponding verb. However, some participles have clearly taken on an adjectival or nominal meaning of their own and may therefore also have the participle form as their base form. This will introduce systematic ambiguities in some cases. In Finnish there is the present participle (*-va*) , the past participle (*-nut*) , the agent participle (*-ma*) and the negation participle (*-maton*) that may introduce such ambiguities. Ambiguities between lexicalised and systematic analyses can be resolved in lexicalised parsing grammars as documented in Voutilainen (2003), so emergence of such ambiguities is not considered problematic.

Derivational endings more often than not introduce a new meaning to a stem so there will be fewer mistakes by not stripping away a derivational ending. For identified derivational endings, it is still useful to indicate the derivation, e.g. *ärsyttävästi* DRV=STI (irritatingly), even if the word is not reduced to a potential base form such as *ärsyttävä* (irritating) or *ärsyttää* (irritate).

The same reasoning with regard to valency and frames also applies to newly coined derivations and it is a task for further investigations how transparent productive derivations are. From a technical point of view, a base form is simply an index to a separate semantic unit with its own syntactic behaviour. If two forms of a word have similar syntactic preferences, they may as well be reduced to the same base form.

Syntax. Finnish syntax is characterised by (relatively) free constituent order. The rich Finnish morphology provides for means to express constraints on how syntactic units can be combined with each other. A parsing grammar for Finnish syntax requires extensive lexical information of valency/frame type. Such information needs to be identified from existing resources or extracted from large morphologically analysed corpora.

There are also some other features in Finnish grammar that need a principled (or at least operational) classification (similar challenges occur in other languages too): (i) analysis of so-called special clause types (where the potential subject has an untypical case); (ii) continuum from auxiliaries to semiauxiliaries to main verbs (a similar continuum exists in other languages too, e.g. English (Quirk & al 1985: 136-147); (iii) nominalisation (continuum from verbs to nouns). The grammar definition corpus drawn from Hakulinen *et al*. illustrates continua such as these with numerous well-ordered example sentences, which helps make a systematic categorisation.

4. Dependency representation in outline

In this section, we outline the dependency grammar representation used in the grammar definition corpus mostly by examples and short notes. A larger documentation of the linguistic representation ("style sheet") will be published separately.

Our dependency syntactic representation follows common practice in many ways. For instance, the regent of the sentence is the main predicate verb of the main clause, and the main predicate has a number of dependents (clauses or more basic elements such as noun phrases) with a nominal or an adverbial function. More simple elements, such as nominal or adverbial phrases, have their internal dependency structure, where a (usually semantic) head has a number of attributes or other modifiers. In our representation, grammatical markers (such as determiners, conjunctions, auxiliaries and adpositions) are described as dependents (with an attributive or phrase marker or auxialiary function); as a result, semantically "heavier" words get a head status in dependency analyses. In this respect, our representation follows that used in the Prague Dependency Treebank (while e.g. the Danish Dependency Treebank follows almost the opposite policy of granting grammatical categories a head status).

The dependency function palette is fairly ascetic at this stage. The dependency functions for nominals include Subject, Object, Predicative and Vocative; adverbials get the Adverbial function; modifiers get one of two functions, depending on their position relative to the head: premodifying constructions are given an Attributive function tag; postmodifying constructions are given a Modifier function tag. In addition, the function palette includes Auxiliary for auxiliary verbs, Phrasal to cover phrasal verbs, Conjunct for coordination analysis, and Idiom for multiword idioms.

The present surface-syntactic function palette can be extended into a more fine-grained description at a later stage; for instance, the Adverbial function can be divided into functions such as Location, Time, Manner, Recipient and Cause. Such a semantic classification is best done in tandem with a more fine-grained lexical description (entity classification, etc).

Here are some sample analyses in tabular format. The leftmost column gives a numerical address the each token (word or punctuation mark); note that position ”0” is given as regent of the main predicate verb of the main clause. The second column from the left shows the dependency relation by indicating the position of the regent of the current word. The third column from the left shows the dependency function of the dependent. The fourth column shows the word-form itself. The fifth column shows the base form of the word (including compound boundary marker ”#”). The sixth column shows the morphological tags, e.g. word-class and inflection tags.

The quantifier *kaikki* (all) is analysed as Attribute (attr) of the Subject (subj) noun *peruslagerit* (basic lagers); the main predicate verb of the sentence *ovat* (are) is linked (axiomatically) to ”0”, and has also another dependent, the Predicative (pred) *samanlaisia* (similar), which has a modifying adverb *hyvin* (very) labelled as Attribute.

Table 1. “All basic lagers are very similar.”

1	2	attr	Kaikki	kaikki	all	PRON NOM PL
2	3	subj	peruslagerit	peruslager	basic-lager	N NOM PL
3	0	main	ovat	olla	be	V ACT IND PRES PL3
4	5	attr	hyvin	hyvin	very	ADV
5	3	pred	samanlaisia	samanlainen	similar	A PTV PL

Sometimes, the question arises whether to relate elements to each other on syntactic or on semantic criteria. As an example from English, consider the sentence ”I bought three litres of milk”. On syntactic criteria, the head of the object for the verb ”bought” is ”litres”, but semantically one would prefer ”milk”. Our dependency representation relates elements to each other based on semantic rather than inflectional criteria, and this has resulted in some analyses that we look at next. Note that in the following examples, base forms and morphological tags are omitted for simplicity.

Titles, roles, given names and other non-final parts of names generally are given an Attribute function rather than a nominal head function when they are followed by a suitable semantic head, e.g. surname. Also quantifiers are analysed as Attribute of the quantified expression. For example, *joukon* (group of) is analysed as Attribute of *ihmisiä* (people).

Table 2. “The resing place employs a group of people.”

1	2	subj	Taukopaikka	tauko#paikka	rest-place	N NOM SG
2	0	main	työllistää	työllistää	employ	V ACT IND PRES SG3
3	4	attr	joukon	joukko	group-of	N GEN SG
4	2	obj	ihmisiä	ihminen	people	N PTV PL

Adpositions (prepositions and postpositions) are analysed as Phrase mark (rather than regent) of the adjacent nominal phrase. For instance, the preposition *ennen* (before) is

analysed as Phrase mark of the noun *paluutaan* (his return). As an additional advantage, adpositional phrases receive a more similar dependency analysis with e.g. locative nominal phrases where the locative case is given morphologically (locative suffix) rather than syntactically (with an adposition). In both cases, the nominal phrase is regarded as the head category that can serve a nominal or adverbial function in the sentence.

Table 3. "Koivisto had not received all of his receivables before his return."

1	2	subj	Koivisto	Koivisto	Koivisto	N NOM SG
2	3	aux	ei	ei	not	NEG
3	4	aux	ollut	olla	have	V ACT SG3
4	0	main	saanut	saada	receive	V ACT PCP PAST SG
5	6	attr	kaikkia	kaikki	all	PRON PTV PL
6	4	obj	saataviaan	saatava	receivable	N PTV PL POSS
7	8	pmark	ennen	ennen	before	PREP
8	4	advl	paluutaan	paluu	return	N PTV SG POSS

Also conjunctions (coordinating and subordinating) are analysed as Phrase mark for the unit that they introduce. In the case of the coordinating conjunction, e.g. *mutta* (but), the regent of the Phrase mark function is the (head of) the following conjunct. The conjunct itself is linked to the other (preceding) conjuct head.

5. Annotation and deliverables

The manual tagging of the syntactic dependencies and functions was done by three linguists with background in Finnish linguistics working on separate sections of the grammar definition corpus, after a week's training period. The data for annotation was given in a spreadsheet format, with the columns for dependency relation and dependency function to be populated by the annotators.

During the annotation period, 1-2 weekly meetings were arranged to discuss and resolve e.g. borderline cases between different analyses. In addition, the annotators cross-checked each other's output to detect possible interannotator inconsistencies. The highest consistency would probably have been reached using double/triple-blind method combined with negotiations (Voutilainen, 1999), but this method was not used due to resource and time limitations.

As a result of the discussions, the documentation of the dependency syntactic representation was extended and made more specific. Problematic cases and outright misanalyses were often detected by the annotators when checking their own annotations; additional cases and inconsistencies were found as a result of daily cross-checks between the annotators. In case of genuinely problematic cases, the annotators were instructed not to force an arbitrary analysis, but to leave the problematic part of the sentence unanalysed, and to bring it to the weekly meetings. The work on syntactically

annotating the grammar definition corpus of the 19,000 grammar sentences by hand took approximately 5 person months.

The 19,000-sentence grammar definition corpus and documentation has been published (contct details to be provided); additional corrected versions will follow through 2011-2012.

A limited amount of running text representing different genres and taken from various public sources has also been annotated manually according to the dependency syntax specification resulting from the grammar definition phase. This step provides additional high-quality annotated corpus for researchers (e.g. to serve as additional learning and testing material for building language models for rule-based and statistical parsers). In addition, this step will help experiment with the usability of the developed grammar scheme in the analysis of real-world text; in terms of coverage and consistency, for instance. The manually annotated corpus will be published during 2011.

Initial experiments on interannotator agreement using the double-blind method and negotiations with limited data (three texts from different genres amounting to over 200 sentences) have been carried out to assess the pros and cons of using a systematic set of example sentences from a descriptive grammar as the initial data in a treebank (anonymous citation, to be provided). The main observations were that after negotiations, the interjudge agreement at word level (labelled dependency relations) was close to 99%. During the negotiations it was found that also complex syntactic phenomena, including various mid or low frequency special sentence types, were generally annotated quite consistently among the annotators, even before the negotiation phase took place. This supported the hypothesis that a grammar definition corpus would cover a high number of syntactic constructions in the language, and the resulting treebank and documentation should guide annotation of sentences containing these syntactic phenomena.

During the experiments it was also found that annotations were unsystematic mostly in expressions including numerals and referring to temporal or areal phenomena, which are typically poorly covered (maybe as linguistically "uninteresting phenomena") in traditional descriptive grammars. In the case of such semi-structured phenomena, the need to negotiate a consistent analysis to be documented in the annotator's manual and exemplified in the grammar definition corpus, became evident.

6. Work to do

The ongoing project will deliver also large corpora from public sources (such as the Finnish EuroParl corpus) analysed automatically following the dependency syntax specification described above. The automatic analysis (or alternative analyses) will result from language models and parsers made according to the grammar definition corpus and its documentation. The accuracy of the automatic analysis will be lower than is the case with the manually analysed corpora, but the much higher volume of text will enable e.g. quantitative linguistic studies.

References

Cinková, S., Toman J., Hajič J., Čermáková K., Klimeš V., Mladová L., Šindlerová J., Tomšů K. & Žabokrtský, Z. (2009). Tectogrammatical Annotation of the Wall Street Journal. *Prague Bulletin of Mathematical Linguistics*, 85-104.

Hakulinen, A., Vilkuna, M., Korhonen, R., Koivisto, V., Heinonen, T. & Alho, I. (2004). *Iso suomen kielioppi*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Haverinen, K., Ginter, F., Laippala, V., Viljanen, T. & Salakoski, T. (2009). Dependency Annotation of Wikipedia: First Steps towards a Finnish Treebank. In Marco Passarotti, Adam Przepiórkowski, Savina Raynaud and Frank Van Eynde (Eds), *Proceedings of The Eighth International Workshop on Treebanks and Linguistic Theories (TLT8*) (pp. 95-107). Milano: EDUCatt.

Jäppinen, H., Lehtola A. & Valkonen K. (1986). Functional structures for parsing dependency constraints. In *Proceedings of the 11th conference on Computational linguistics*. Association for Computational Linguistics (pp. 461-463). Bonn: Institut for angewandte Kommunikations- und Sprachforschung e.V.

Karlsson. F., Voutilainen, A., Heikkilä J. & Anttila A. (1995). *Constraint Grammar: A Language-Independent Framework for Parsing Unrestricted Text*. Berlin / New York: Mouton de Gruyter.

King, T., Crouch, R., Rietzler, S., Dalrymple, M. & Kaplan, R. M. (2003). The PARC 700 Dependency Bank. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora, held at the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*. Budapest: ACL.

Marcus, M., Santorini B. & Marcinkiewicz M. (2004). Building a large annotated corpus of English: the Penn Treebank. In G. Sampson & D. McCarthy (Eds.), *Corpus Linguistics: Readings in a Widening Discipline.* New York: Continuum.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Tapanainen, P. & Järvinen T. (1997). A non-projective dependency parser. In *Proceedings of the fifth conference on Applied natural language processing.* Washington, DC: ACL.

Voutilainen, A. (2003) Part-of-Speech Tagging. In Ruslan Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (pp 219-232). Oxford and New York: Oxford University Press.

Discurso, análisis literario y corpus

An approach to native and non-native writers' use of interactional metadiscoursal features in scientific abstracts in English within the field of Agricultural Sciences

Mª Milagros del Saz Rubio

Universitat Politècnica de València

Abstract

This paper approaches the use of interactional metadiscoursal features in a corpus of 48 scientific abstracts written in English by English and Spanish native speakers with a view to unveiling differences in the rhetorical choices, use and frequency of distribution of such features. Drawing on Hyland's (2005) pragmatic-oriented framework of metadiscourse, the categories of hedges, boosters, attitude and engagement markers, and self-mentions were electronically searched with the help of the software WordSmith 5.0 and their rhetorical functions identified and mapped onto the different sections that integrate the research abstract (cf. Santos 1996; Swales 1990). Results reveal that interactional metadiscourse is most frequently enacted through the use of hedges, boosters and attitude markers, although differences can be pointed out regarding the native and non-native use of such features in the different sections that make up the research abstract, thus pointing to rhetorical variation in the functions of these items.

Key words: interactional metadiscourse, research abstracts.

Resumen

Este artículo aborda el uso de elementos metadiscursivos de tipo interaccional en un corpus de 48 resúmenes científicos escritos en inglés por un grupo de autores nativos de la lengua inglesa y española con la finalidad de delimitar si existen diferencias en las elecciones retóricas de los mismos, así como en el uso y distribución de tales elementos metadiscursivos en las diferentes partes que integran el resumen. Tomando el modelo de metadiscurso de orientación pragmática de Hyland (2005) como punto de partida, se han contabilizado de forma electrónica las categorías metadiscurvisas de los matizadores, enfatizadores, marcadores de actitud y de compromiso y las auto-menciones gracias al empleo del software WordSmith 5.0. Además se han identificado sus funciones retóricas dentro de cada uno de los movimientos que integran el resumen científico (véase Santos 1996; Swales 1990). Los resultados indican que las categorías metadiscursivas más empleadas son los mitigadores, enfatizadores y los marcadores de actitud, aunque se han constatado diferencias en cuanto al uso que de éstas hacen los dos grupos de hablantes, lo cual apunta a la existencia de variedad retórica en cuanto a las funciones que tales unidades desempeñan en las diferentes secciones que constituyen el resumen científico.

Palabras clave: metadiscurso interaccional, resúmenes científicos.

1. INTRODUCTION

The relevance of academic writing is nowadays more than justified as demonstrated by the large body of research in this area. Authors such as Berkenkotter, Huckin & Ackerman (1991) have brought to attention the importance of mastering a *specialized literacy*, especially for students or researchers entering the academic disciplines. This literacy can be defined as the ability to make use of the discipline-specific rhetorical and linguistic conventions in order to fulfill the purpose as writers. Mastering academic writing thus involves an awareness of the existence and structure of specific genres, as a key element for acculturation and success. Therefore, to engage in the writing of a genre such as the research article or abstract, inevitably calls for awareness of its specific conventions, as well as of the role of the writer and the purpose of the writing task. This situation can be certainly more complex for researchers who need to write and publish their research in an L2, since mastering the grammar, lexicon or syntax is not enough to guarantee them communicative competence. What is more the discourse or rhetorical structures of scientific texts in different languages could vary to a great extent due to the existence of cultural influences. Taking all this into consideration, the main aim of this paper is to assess whether there is variation in the rhetorical preferences of native English and Spanish-speaking researchers when producing research article abstracts for publication in the same English journals within the field of Agricultural sciences. To do so, a corpus of 48 abstracts, 24 written by native English speakers (NESs) and 24 by native Spanish speakers (NSSs), has been analyzed and a quantitative and qualitative analysis of the interactional metadiscoursal features they employ, drawing on Hyland's (2005) framework, is carried out using WordSmith Tools 5.0.

It is expected that results will help unveil different or similar rhetorical preferences in the writing conventions of both sets of writers regarding the use of metadiscoursals units. Finally, the results obtained here can have implications for the teaching of academic writing to non-native speakers of English. As such, they will be taken as a starting point for the design and elaboration of meaningful writing activities aimed at raising awareness of the conventions and expectations which operate in the genre of the research article in English in the field of Agricultural Sciences.

2. HYLAND'S INTERPERSONAL MODEL OF METADISCOURSE

The view that academic writing is purely informative and objective is no longer sustained, as it is generally agreed that writing a research paper is not only a matter of transmitting facts, but also, and perhaps more importantly, of interpreting these facts, while aiming to persuade readers of the truth of certain claims. Therefore, the fact that academic writing participates of a persuasive component has been long propounded (Stotesbury, 2003; Swales, 1990; Hyland, 2000). In this vein, the abstracts that accompany research papers are not only examples of expository prose, but rather, texts which are profoundly persuasive in themselves. Recently, there has also been an increasing interest in the study of interpersonal or evaluative features which contribute to creating a relationship with the reader. These features are key mechanisms that especially non-native writers need to

achieve and master if they want to publish in international scientific publications while persuading readers, peer reviewers and editors of the acceptance of their claims.

Thus, for the analysis of the interactional metadiscoursal features employed in this corpus I have taken Hyland's interpersonal model of metadiscourse (2005) as a point of departure. Hyland differentiates between two main types of metadiscourse which are, drawing on the Hallidayan distinction, the *interactional* and *interactive* type (Halliday, 1985). The former is concerned with how writers convey their opinions, interact with their readers and intrude on the text (cf. Vande Kopple, 1985). On the contrary, interactive metadiscourse is concerned with guiding the reader through the text, while managing the information flow. In spite of these differences, Hyland in his 2005 framework states that "all metadiscourse is interpersonal in that it takes account of the reader's knowledge, textual experiences and processing needs and that it provides writers with an armory of rhetorical appeals to achieve this" (Hyland, 2005: p.41). Due to time and space limitations, only the interactional categories will be analyzed here in an attempt to unveil how writers construct a relationship with their readers, reach the audience while intruding and commenting on their own message by making their views implicit through the use of interactional metadiscourse items. Likewise, attention will be paid to the rhetorical choices of both NESs and NSSs in the use, frequency and distribution of such elements throughout the abstract.

3. MATERIALS AND METHODOLOGY

For the current study, I have gathered a corpus of 48 multi-authored abstracts published during 2006-2009 in peer-reviewed journals devoted to a wide range of scientific topics within the field of Agricultural Sciences. Research article abstracts were selected and downloaded from journals on the basis of their impact factor. 24 of these abstracts were written by authors who are native speakers of English and thus affiliated within English-speaking institutions, whereas the other remaining 24 abstracts were written by Spanish authors affiliated within Spanish-speaking research institutes or universities. I first proceeded to identify the metadiscoursal items under analysis carrying out an electronic computer search with WordSmith 5.0. The metadiscoursal items searched were those listed in Hyland (2005). However, on a second stage, I complemented the electronic search with a manual codification of these elements as some items are used in a non-metadiscoursal role, or it can be the case that new categories are not included in the already existing frameworks or taxonomies of analysis and thus may escape the researcher's view.

4. RESULTS AND DISCUSSION

In this section I will first comment on some general findings concerning both the use and frequency of interactional metadiscoursal items in the research abstracts surveyed to later focus on the different metadiscoursal categories most commonly employed by NESs and NSSs. The rhetorical functions of these features will also be mapped onto the different

moves that integrate the research abstracts, drawing on Santos (1996) and Swales' (1990) classification of such moves:

- Move 1-Situating the research or STR;
- Move 2-Signaling the gap (STG);
- Move 3-Presenting the research or PTR;
- Move 4- Describing the methodology or DTM;
- Move 5-Summarizing the findings (reporting the main findings of the study) or STF;
- Move 6- Discussing the research (interpreting the results or findings, giving recommendations, or implications/applications of the study), or DTR).

A total of 330 metadiscoursal items have been identified in the corpus analysed, which represent 3% of the language employed (12,754 words). The distribution of metadiscourse in both sets of the abstracts is fairly similar: 177 metadiscoursal items in the abstracts written by NESs, (3% of the language employed or 6,255 words). In the case of those abstracts written by NSSs, metadiscoursal items (153 items) represent 2% of the language employed (6,499 words), so there is a slightly higher percentage in the use of metadiscourse in the case of native speakers of English. Among the categories surveyed, neither engagement markers nor self-mentions have been found. Thus, interactional meanings are best accounted for through the use of hedges (45%), attitude markers (41%) and boosters (14%).

The next step was to see how these metadiscoursal items were patterned onto the communicative moves that make up the abstract in order to see if writers from both nationalities make a similar or different use of such devices in the different abstract moves.The distribution of interactional metadiscourse throughout the different sections of the abstract shows a higher concentration of metadiscoursal features in the STF and DTR sections (34% and 40%, respectively), as they are the most interpretive sections, followed by STR (14%), GS (7%), PTR (5%) and DTM (1%).

4.1. Interactional Metadiscourse in abstracts by NESs and NNESs

A comparison of the use that NESs and NSSs make of interactional metadiscourse items indicates that there is no striking significant difference in the percentages of use of metadiscoursal elements, except for the fact that Spanish writers show a preference for attitude markers to convey interactional meanings (44%) rather than for hedges (40%), and make a greater use boosters (16%) if compared to English writers. In contrast, English writers employ hedges (49%) as the most frequent category, followed by attitude markers (39%) and boosters (12%). The following tables depict the percentages of use of the different interactional categories within the different sections of the abstract. For time and space reasons I will just highlight the most relevant findings concerning the use of interactional metadiscoursal items in the two sets of abstracts:

Table 1 Interactional metadiscourse in NESs

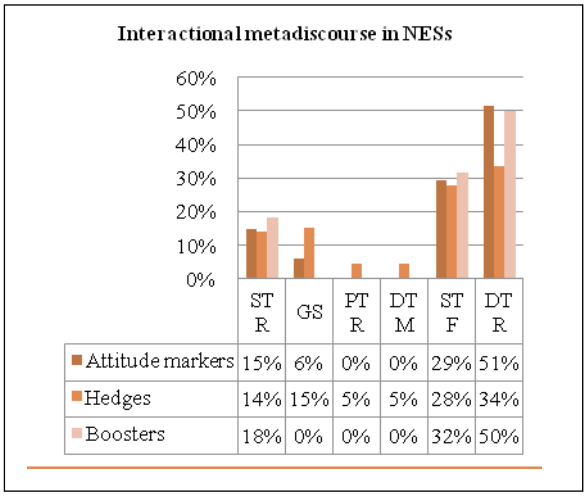
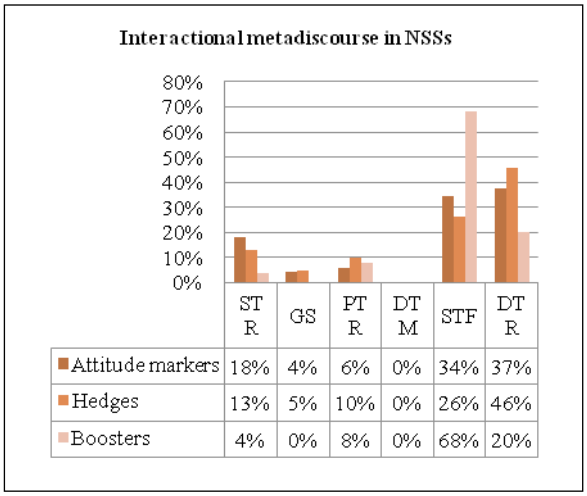


Table 2: Interactional metadiscourse in NSSs



Attitude markers are used by NESs in all the abstract moves except when they present their piece of research (PTR) and in the DTM section, two of the most objective moves. In contrast, Spanish writers tend to employ attitude markers to present their piece of research, especially to highlight its importance, relevance or to claim that their objectives are essential, necessary or interesting:

Example 1: Abstract 16 (PTR) (NSSs)

Furthermore, as a preliminary study, it was interesting *to know the effects of biodiesel fuel on a common-rail high pressure injection system, those more useful in modern light duty diesel engines, as a consequence of its different physicochemical properties compared with conventional diesel fuel. As the real goal of the study is to compare fairly performance and emissions from the engine, it was* essential *to know any injection effects owed to fuel's own characteristics that finally would affect those parameters that will be evaluated […].*

Regarding the use of hedging devices, both native English and Spanish writers employ them quite frequently to present findings, and to evaluate or discuss the research (PTF, DTR), as it is in these moves where writers can assess the relevance, impact or application of their findings and where they portray their attitude towards the results:

Example 2: Abstract 18 (DTM) (NSSs)

A priori, it could be thought *that viscosity and density values will be the most significant parameters capable of altering the injection rate. (abstract 16, Spanish)*

Example 3: Abstract 16 (DTM) (NESs)

These results imply that lysergic acid may be involved in the fescue toxicosis syndrome.

One significant difference concerns the use of hedges by English speakers to signal the gap (15% compared to just 5% in the Spanish set), that is, they use mitigating devices in the 'delicate' communicative category of signaling that there is a problem, gap, deficiency or inconsistency in previous studies. This is a move which initially calls for a great degree of hedging devices to conform to politeness norms:

Example 4: Abstract 8 (STG, NESs)

The potential ecological benefits of an alternative agronomic practice such as alley cropping are numerous, but the practice is unlikely to be adopted unless *it is economically viable.*

Another use of hedges concerns the presentation of hypotheses in the PTR section. However, Spanish writers have been found to employ hedging devices twice as much as English writers (10% and 5% respectively), when hypotheses are presented, especially in the form of modal verbs such as *should, could, would* and *may/might*:

Example 5: Abstract 22 (PTR, NSSs)

We hypothesize that (1) burned forest stands should *exhibit lower net mineralization rates than unburned ones; (2) these differences would be greatest during the growing season; (3) differences between soil variables* might *also be observed among plots from different years since the last fire.*

Example 6: Abstract 23 (PTR, NESs)

We tested the hypothesis that moderate amounts of dietary HTS could *reduce markers of oxidation on muscle of rats without having detrimental effects in growth.*

Regarding the use of boosters, the main difference between the two sets of abstracts concerns their appearance in the STR move, that is, the move where writers contextualize their research and present claims of increasing specificity regarding the topic of study. English writers tend to commit themselves to the truth of the claims reported through the use of boosters, whereas their use by Spanish writers is far less prolific (only 4% compared to 18%). This somehow portrays the idea that Spanish writers maintain a more objective tone in their argumentation when reporting claims or claiming topic centrality and thus less involvement is conveyed on their part if compared to their English counterparts:

Example 7: Abstract 17 (STR, NSSs)

Conjugated linoleic acid (CLA) exerts a strong *positive influence on human health but intake of these fatty acids is typically too low, and increased consumption of CLA is recommended.*

Example 8: Abstract 18 (STR, NESs)

Turfgrass irrigation strategies must *be* clearly *defined in response to increasing concerns over quality water availability.*

On their part, English writers make use of boosters in the DTR section whereas Spanish writers use them in the PTF move. A possible explanation is that English writers make a more clear distinction between the two sections whereas in the case of the Spanish abstracts it was far more difficult to separate two of them, as these two moves seem to coalesce.

5. CONCLUSIONS

The findings portrayed above─albeit tentative─ have shown that both sets of writers make a fairly similar use of metadiscoursal devices in an attempt to build and foster a relationship with the reader. However, the main differences found concern the distribution and specific use of these features in each of the communicative moves that make up the research abstract. Although writers from both nationalities seem to make a similar use of these metadiscoursal items, they differ in the amount and distribution of hedging devices, which are more frequently employed by native speakers of English. The underuse of such devices has been already pointed out and could be the result of a lack of explicit instruction on the part of Spanish academics. Likewise, the lack of mitigating devices in the signaling the gap move, where a certain amount of hedging devices is employed by English writers, points to the need of explicit instruction for Spanish academics as they seem to lack the resources to convey problems, deficiencies or inconsistencies while attending to politeness conventions (Myers 1989). However, further research is necessary before the findings obtained in the preliminary study carried out here can be considered conclusive.

6. REFERENCES

BERKENKOTTER, C. HUCKIN, T.N. & ACKERMAN J. (1991). Social context and socially constructed texts. In C. Bazerman & J. Paradis (Eds.), *Textual Dynamics of the Professions* (pp. 191-215). Madison: University of Wisconsin Press.

BUNTON, D. (1999). The Use of Higher Level Metatext in PhD Theses. *Journal of English For Specific Purposes, 18*, S41–S56.

HALLIDAY, M.A.K. (1985). *An Introduction to Functional Grammar*. London: Edward Arnold.

HYLAND, K. (2000). *Disciplinary discourses: Social interactions in academic writing*. London: Longman.

HYLAND, K. (2005). *Metadiscourse*. London: Continuum.

MYERS, G. (1989). The pragmatics of politeness in scientific articles. *Applied Linguistics, 10*, 1-35.

MORENO, A. (1997). Genre constraints across languages: causal metatext in Spanish and English RAs. *English for Specific Purposes, 16 (3),* 161–179.

SANTOS, M.B. (1996). The textual organization of research paper abstracts in applied linguistics. *Text, 16(4)*, 481-499.

STOTESBURY, H. (2003). Evaluation in research article abstracts in the narrative and hard sciences. *Journal of English for Academic Purposes*, *2(4)*, 327–342.

SWALES, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

VALERO-GARCÉS, C. (1996). "Contrastive ESP rhetoric: metatext in Spanish-English economic texts", *English for Specific Purposes*, 15 (4), 279–294.

VANDE KOPPLE, W. (1985). Some Exploratory Discourse on Metadiscourse. *College Composition and Communication, 36*, 82–93.

Evaluative adjectives in a corpus of Greek opinion articles

Georgia Fragaki
University of Athens

The aim of this paper is to contribute to the description of the category of evaluative adjectives, drawing on a corpus of opinion articles (450,576 words) from the Corpus of Greek Texts (CGT), a reference corpus of Greek. Existing attempts to describe evaluation in text regard adjectives as an important device of evaluation, without, however, systematically studying the categories of adjectives involved in evaluation. The most common assumption is that these come from the category of descriptive adjectives or have the typical features of descriptive adjectives (e.g. positive or negative meaning, gradability). In this paper it is claimed that several adjective categories can assume an evaluative function in Greek, among which a special category of evaluative adjectives, whose exclusive function is evaluation. This category consists of four subgroups, namely modal adjectives, comment adjectives, intensifying adjectives and adjectives of importance.

Keywords: adjective, evaluation, categorization, opinion articles, Greek

El objetivo de este trabajo es contribuir a la descripción de la categoría de los adjetivos de evaluación, a partir de un corpus de artículos de opinión (450.576 palabras) de Corpus de Textos Griegos. Los intentos existentes para describir la evaluación en el texto consideran los adjetivos como un dispositivo importante de la evaluación, sin estudiar de manera sistemática las categorías de los adjetivos que participan en la evaluación. El supuesto más común es que estos proceden de la categoría de los adjetivos descriptivos o tienen las características típicas de los adjetivos descriptivos. En este trabajo se afirma que varias categorías de los adjetivos pueden asumir una función de evaluación en griego, entre los cuales una categoría especial de los adjetivos de evaluación, cuya única función es la evaluación. Esta categoría consiste en cuatro subgrupos: adjetivos de modalidad, adjetivos de comentario, adjetivos de intensificación y adjetivos de importancia.

Palabras clave: adjetivo, evaluación, clasificación, artículos de opinión, griego

1. Adjectives and evaluation in the literature

Evaluation in text is manifested through a variety of devices, including words, phrases, lexical or grammatical categories and patterns. Adjectives are typically considered as "a very important and frequent means [of evaluation]" (Hewings, 2004: 253) or as "one of the most prototypical and canonical exponents of evaluation" (Swales & Burke 2003: 3). In addition, the presence of adjectives in a text has been correlated with the degree of subjectivity in it, as shown in computational linguistic studies (see e.g. Hatzivassiloglou & Wiebe, 2000). However, the features of the adjectives involved in evaluation have not been systematically studied so far in the literature.

One of the recurring remarks in the literature is that adjectives are usually found in evaluative patterns with link verbs (Hunston & Francis, 1999: 188-189; Hunston & Sinclair, 2000). However, this is not a safe way to find which adjectives are typically used as devices of evaluation, since it is the pattern that has the evaluative role and not the adjective; thus, any adjective used in an evaluative pattern would acquire an evaluative meaning (cf. Hunston, 2011: 120).

Studies dealing with evaluative adjectives, like Hewings (2004), suggest a semantic classification of these into categories such as *interest*, *suitability*, *comprehensibility*, *accuracy*, *importance*, *sufficiency*, *praiseworthiness* and *perceptiveness*. This categorization is probably driven by the specific genre studied, namely peer reviews of journal article submissions, as well as by the focus on the evaluated entities. It is also significant that the adjectives belonging to these categories are distinguished into positive and negative, reflecting thus the view that evaluation is characterized by polarity. Similarly, Swales and Burke (2003) classify evaluative adjectives in academic texts into semantic categories such as *acuity*, *aesthetic appeal*, *assessment*, *deviance*, *relevance size* and *strength*. These are distinguished into polarized (strongly positive or negative) and centralized adjectives (not extreme positive or negative), something which shows that gradability and polarity are again regarded as main characteristics of evaluative adjectives. However, in both approaches evaluative adjectives are not defined as a set of items that have common features or differ from other types of adjectives. Furthermore, the proposed sub-classifications are closely related to the specific text types studied and exclusively based on meaning. For this reason, the classifications proposed are not easily generalizable.

It can be argued that in terms of the traditional distinction between descriptive and classifying adjectives, evaluative adjectives in the literature are usually considered a subclass of descriptive adjectives (e.g. Hewings, 2004: 253). Even when there is no claim about the membership of evaluative adjectives, they are regarded as having the typical features of descriptive adjectives, i.e. positive or negative meaning and the resulting relations of antonymy, the ability to form relative or superlative degree, gradability etc. (e.g. Hunston & Francis, 1999: 188-189; Hunston & Sinclair, 2000: 91).

This paper presents an attempt to define the category of evaluative adjectives as a part of a wider categorization of adjectives in a corpus of Greek opinion articles. This categorization is made on the basis of specific formal, functional and semantic criteria with a view to examining whether there is a particular category of adjectives with an evaluative role *par*

excellence. The application of corpus-based criteria helps avoid the circular argument that evaluative adjectives are those that are employed as devices of evaluation in text. More specifically, the following two questions are the focus of this paper:

- Which are the features and functions of the adjectives that can be used for evaluation in text?

- Which parameters of evaluation are expressed through adjectives?

2. Data and methodology

The data used in this study include opinion articles from three Greek newspapers, *To Vima, I Kathimerini* and *Rizospastis* (450,576 words in total). The source of data is the Corpus of Greek Texts (CGT), a reference corpus of Modern Greek, which includes about 28,000,000 words from a variety of text types (Goutsos, 2010). The opinion articles selected were published between 1996 and 2003 and concern social, political, financial and leisure topics.

CGT is not tagged for grammatical categories and thus adjectives were manually extracted from the frequency list of all running words in the corpus on the basis of specific criteria. The features and functions of all adjectives occurring at least 15 times were analysed, using information from concordances. The span used for these concordances was 5 words to the left and right of the node, but wider context was also consulted as appropriate.

Features and functions of the adjectives in the corpus were used as criteria for the categorization of adjectives. These include the syntactic role of the adjective (attributive, predicative or used in phrases with a link verb and a clause complement), pre- or post-nominal position of the adjective in the case of attributive use, adjective modification by an adverb as an indication of gradability, ability to form comparatives and superlatives attested in the corpus, coordination with other adjectives and position in the noun phrase, adjective complements, collocations (frequent, restricted or fixed) and semantic criteria.

3. The category of evaluative adjectives in the corpus

The employment of the criteria described above has led to the categorization of the adjectives in the corpus into ten categories (for more details see Fragaki, 2010):

a. classifying adjectives, which constitute the largest category in the corpus, used to classify a modified noun (e.g. *κοινωνικός*, social),

b. descriptive adjectives, used for the attribution of a property to a modified noun (e.g. *κακός*, bad),

c. evaluative adjectives, which is the third largest category, used to express local or textual evaluation (e.g. *λεγόμενος*, alleged),

d. deictic adjectives, used for placing an event in relation to the deictic centre (e.g. *επόμενος*, next),

e. relational adjectives, used for relating two entities with respect to a property which they do not name (e.g. *κοινός*, common with),

f. specializing adjectives, used for restricting or generalizing the reference of the modified noun (e.g. *ειδικός*, particular),

g. indefinite adjectives, which express indefiniteness (e.g. *διάφοροι*, several)

h. colour adjectives (e.g. *μαύρος*, black)

i. verbal adjectives, which are equivalent to verbs in Greek, have a verbal meaning or characteristics (e.g. *γεμάτος*, full)

j. quantitative adjectives, used to define the amount of an entity (e.g. *πολύς*, much, many).

In the proposed categorization the categories of descriptive and evaluative adjectives are clearly distinguished, contrary to the assumption in the literature that evaluative adjectives are a subclass of descriptive adjectives. The members of the two categories have different features and functions in the corpus. In particular, descriptive adjectives are used to denote properties such as size (e.g. *μεγάλος* vs. *μικρός*, big vs. small), age (e.g. *παλ(α)ιός* vs. *καινούρ(γ)ιος*, old vs. new), value (e.g. *καλός* vs. *κακός*, good vs. bad) etc. They are also organized in relations of antonymy, as seen in the pairs above. Other features include positive or negative semantic orientation, gradability and the ability to form inflected or periphrastic comparatives and superlatives. Moreover, keeping in mind that the attributive is the most frequent use for Greek adjectives, descriptive adjectives are more frequently predicative (9.1%) than other adjective categories in the corpus. For instance, in example 1, the descriptive adjective *difficult* is used predicatively and is modified by the intensifier *extremely*:

> 1. *Οι πολίτες αυτής της χώρας* […] *έχουν συνειδητοποιήσει ότι τα προσεχή δύο χρόνια θα είναι ιδιαίτερα* δύσκολα ...
>
> The citizens of this country […] have realized that the next two years will be extremely *difficult* …

The adjective here attributes the property of difficulty to the noun phrase *the next two years*.

On the other hand, evaluative adjectives denote the author's stance towards an evaluated entity rather than ascribe properties to it. They are placed away from the modified noun in the case of multiple modification, in which classifying or descriptive adjectives are usually found in between. They also show high frequency in patterns with link verbs and thus can assume an organizing role in discourse. In the following example, the evaluative adjective *alleged* does not attribute any property to the noun *upgrade*:

> 2. *Πέρυσι ψηφίσθηκε ο νόμος για τη* λεγόμενη *«ανωτατοποίηση» των ΤΕΙ.*
>
> Last year the law for the *alleged* "upgrade" of Polytechnics passed.

Here the author doubts that Polytechnics have really improved to be at the same level with other universities, something which is also indicated by the use of inverted commas for *"upgrade"*.

On the basis of the criteria mentioned above, four sub-categories of evaluative adjectives can be distinguished: modal adjectives, comment adjectives, intensifying adjectives and adjectives of importance. It is notable that two of these categories (modal and adjectives of importance) concur with Hunston's (1994) and Thompson & Hunston's (2000) parameters of evaluation.

Modal adjectives indicate deontic or epistemic modality, i.e. the degree to which something is necessary, possible or sure, according to the author, e.g. *αναγκαίος* (necessary), *πιθανός* (possible). These are recurrently used in phrases with link verbs such as *είναι δυνατό(ν) να* (it is possible to) and *είναι/θεωρείται βέβαιο(ν) ότι/πως* (it is/it is considered certain that).

Comment adjectives, including adjectives like *περιβόητος* (famous, notorious) and *αποκαλυπτικός* (revealing), are used to make a (usually negative in the corpus) comment on an evaluated entity. For example, by *αποκαλυπτικά*, translated as *revealing* in 3, the author makes a forward evaluation regarding the extract from the *Guardian*.

> 3. *Όσο για τις φορολογικές περικοπές* [...] *που εξήγγειλε ο Πρόεδρος Μπους είναι* αποκαλυπτικά *τα όσα γράφει «Ο Γκάρντιαν»: «Ο Πρόεδρος Μπους κατάφερε κάτι μοναδικό: την απόλυτη ικανοποίηση των ολίγων* […]»
>
> As for the tax cuts […] announced by President Bush what "The Guardian" writes is *revealing*: "President Bush achieved something unique: the total satisfaction of the few […]"

The predicative use of the adjective here facilitates its textual role. Although the adjective *αποκαλυπτικός* in Greek does not have an inherent negative or positive meaning, in this context it is used to show the author's agreement with the *Guardian* and, consequently, to make a negative comment on President Bush. This adjective is commonly used in our data in order to predict or summarize negative comments, something which is indicative of its negative semantic prosody.

Intensifying adjectives like *πλήρης* (total) and *ολόκληρος* (entire) convey the extreme degree to which something happens, according to the author, whereas adjectives of importance such as *σημαντικός* (important) and *κρίσιμος* (crucial) are used to express the degree of importance of what is referred to. Thus, in example 4 the author uses the adjective of importance *βασικός* (basic) in order to attract the reader's attention to the goal of the government, which is presented as important:

> 4. *Ο* βασικός *και κύριος στόχος της κυβέρνησης είναι το μεγαλύτερο δυνατό πετσόκομμα του σημερινού ασφαλιστικού συστήματος.*
>
> The *basic* and main goal of the government is the biggest possible slash of the current social security system.

The high importance is stressed by the coordinated adjective of importance, *κύριος* (chief).

4. The role of adjectives in evaluation

Our corpus study has shown that there is a special category of adjectives that have evaluation as an exclusive function. These adjectives can be clearly distinguished from descriptive adjectives on the basis of specific criteria. This does not mean that only evaluative adjectives can be used as devices of evaluation in text. Descriptive adjectives are also commonly used for the attribution of a (good or bad) property to an evaluated entity. However, their use does not always imply the presence of evaluation, since, among else, some descriptive adjectives (usually, the most frequent) are neutralised because of their recurrent use in formulas or collocations. For example, the adjective *καλός* (good) is frequently used in social formulas like *καλή σας μέρα* (good day to you) etc. More generally, all adjectives have the potential to function as devices of evaluation by being used in patterns or by acquiring features that are characteristic of evaluative adjectives. For example, the modification of a classifying adjective such as *Balkan* by an adverb of degree and its predicative use trigger its evaluative role in: *αυτό* […] *είναι ένας* πολύ βαλκανικός *τρόπος θεώρησης των προβλημάτων*, (this […] is a *very Balkan* way of viewing the problems).

This view can be summarized in Figure 1 below, showing that evaluative adjectives are used for evaluation relating to the parameters of comment, modality, intensification and importance, while descriptive adjectives are mainly restricted to serve the parameter of value.

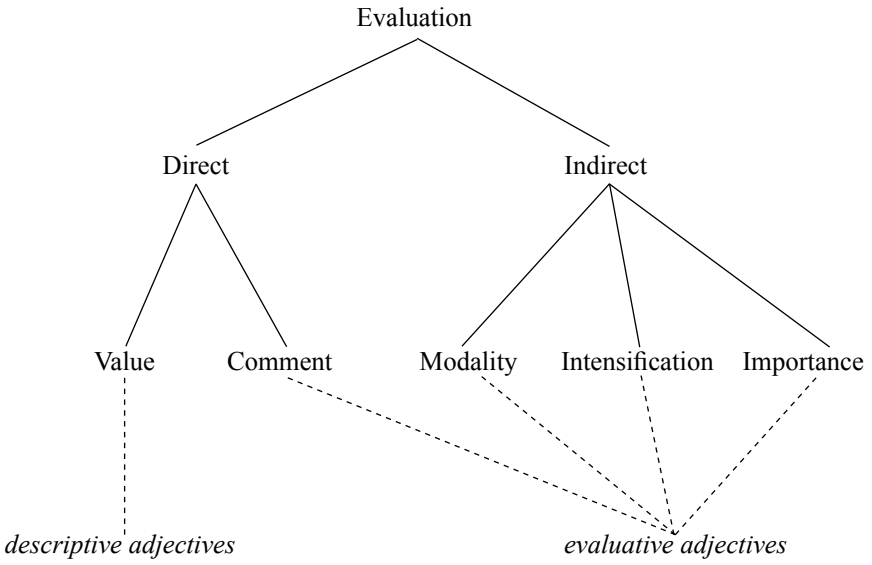


Figure 1. Adjectives and parameters of evaluation in the corpus

The parameters of value and comment contribute to direct evaluation, in the sense that they are bipolar, ascribing a good or bad property or making a positive or negative comment on the evaluated entity, respectively. The parameters of modality, intensification and importance can be thought of as providing indirect evaluation, since the adjectives serving these parameters do not contribute directly to the positive or negative evaluative frame of the text (cf. *attitudinal frame* in Bublitz, 2002). For example, the use of the pattern with the modal adjective *αναγκαίος* (necessary) in 5 below expresses the author's stance towards the necessity of keeping in Athens the former airport (*Elliniko*) as an alternative airport to the current one.

> 5. *Ωσάν να μην υπάρχει στο Ελληνικό ένα πλήρες, οργανωμένο και εύρυθμο αεροδρόμιο* […] *Ενώ* είναι *πολλαπλώς* αναγκαίο να *διατηρηθεί ως εναλλακτικός αερολιμένας! Ούτε το εθνικό συμφέρον ούτε η σύγχρονη διεθνής πρακτική ούτε καν η κοινή λογική δείχνουν να προβληματίζουν τη σημερινή κυβέρνηση* [...]
>
> As if there isn't in Elliniko a complete, organized and fully working airport […] Whereas *it is* for many reasons *necessary to* be kept as an alternative airport! Neither the national interest nor current international practice or even common sense seems to make the present government think [...]

The choice of this strong deontic modality indirectly serves the main goal of the author, which is to make a negative evaluation of the *present government* (see e.g. the last two lines). The strong support to keeping the former airport, which contrasts with the government's view, reinforces the text's negative evaluative frame.

5. Conclusions

The study of evaluative adjectives with the use of corpora can offer a systematic view of evaluation based on the categorization of adjectives according to their different features and functions. In particular, it has been proposed that several categories of adjectives can have an evaluative role, although a separate category of evaluative adjectives has been identified with this specialized role. Thus, relating adjective categories to functions can help us predict their potential and their specific role as devices of evaluation in text.

References

Bublitz, W. (2002). Emotive prosody: How attitudinal frames help construct context. In E. Mengel, H.-J. Schmid & M. Steppat (Eds), *Anglistentag Bayreuth. Proceedings* (pp. 381-391). Trier: Wissenschaftlicher Verlag Trier.

Fragaki, G. (2010). A corpus-based categorization of Greek adjectives. *Online proceedings of the 5th Corpus Linguistics Conference. 21-23 July 2009. University of Liverpool*. Retrieved from http://ucrel.lancs.ac.uk/publications/CL2009/.

Goutsos, D. (2010). The Corpus of Greek Texts: a reference corpus for Modern Greek. *Corpora* 5(1), 29-44.

HATZIVASSILOGLOU, V. & WIEBE, J. (2000). Effects of adjective orientation and gradability on sentence subjectivity. *Proceedings of 18th International Conference on Computational Linguistics (Coling 2000), Saarbrücken, Germany.* Retrieved from www.aclweb.org/anthology-new/C/C00/C00-1044.pdf.

HEWINGS, M. (2004). An "important contribution" or "tiresome reading"? A study of evaluation in peer reviews of journal article submissions. *Journal of Applied Linguistics* 1(3), 247-274.

HUNSTON, S. (1994). Evaluation and organization in a sample of written academic discourse. In M. Coulthard (Ed.), *Advances in Written Text Analysis* (pp. 191-218). London: Routledge.

HUNSTON, S. (2011). *Corpus Approaches to Evaluation. Phraseology and Evaluative Language.* London: Routledge.

HUNSTON, S. & FRANCIS, G. (1999). *Pattern Grammar. A Corpus-driven Approach to the Lexical Grammar of English.* Amsterdam/Philadelphia: Benjamins.

HUNSTON, S. & SINCLAIR, J. (2000). A local grammar of evaluation. In S. Hunston & G. Thompson (Eds), *Evaluation in Text. Authorial Stance and the Construction of Discourse* (pp. 74-101). Oxford: Oxford University Press.

SWALES, J. M. & BURKE, A. (2003). "It's really fascinating work": Differences in evaluative adjectives across academic registers. In P. Leistyna & C. F. Meyer (Eds), *Corpus Analysis: Language Structure and Language Use* (pp. 1-18). Amsterdam: Rodopi.

THOMPSON, G. & HUNSTON, S. (2000). Evaluation: An introduction. In S. Hunston & G. Thompson (Eds), *Evaluation in Text. Authorial Stance and the Construction of Discourse* (pp. 1-27). Oxford: Oxford University Press.

Political language in 140 symbols: Twitter use by Barack Obama and Dmitry Medvedev

Anna Ivanova
University of Seville

Abstract

This paper presents a cross-cultural analysis of Twitter use by Barack Obama and Dmitry Medvedev during the period June-January 2010-2011. Nearly maximum use of available Twitter symbols, low lexical density of the corpus and its restriction to political topic reveal an extensive use of Twitter by both presidents for professional purposes only. High results for Gunning-Fog index classify both corpora as technical texts which possible audience is expected to have a University degree. Thus, we conclude that Twitter platform is used by both presidents as an advertisement tool to give an additional promotion to their and their cabinets' actions, and new technologies are used to tell basically the same "old" story but in a more modern way.

Keywords: Twitter, Barack Obama, Dmitry Medvedev, wordcloud, collocations, Gunning-Fog index

Este artículo presenta un análisis intercultural del uso de Twitter por Barack Obama y Dmitry Medvedev durante el periodo junio-enero del 2010-2011. El uso casi total de los símbolos disponibles de Twitter, la baja densidad léxica del corpus y su restricción al tema político, revelan un uso extenso de Twitter por ambos presidentes sólo en el ámbito profesional. Los altos valores del índice de Gunning-Fog clasifican ambos corpora como textos técnicos cuya posible audiencia se espera que tenga estudios universitarios. De este modo, concluimos que ambos presidentes usan la plataforma Twitter como una herramienta publicitaria que les proporcione a ellos y a sus gabinetes una promoción adicional. Por lo tanto podemos decir que las nuevas tecnologías se usan para contar básicamente la misma "vieja" historia pero siguiendo un nuevo camino.

Palabras claves: Twitter, Barack Obama, Dmitry Medvedev, nube de palabras, colocaciones, índice de Gunning-Fog

1. Introduction

Twitter is a social microblogging network that permits posting messages with the length restriction to 140 symbols. It was launched in 2006, and by January 2011 there were 175,000,000 registered users all over the world. According to Dom Sagolla (2009: n. pag.), "this constraint has created a marketplace of ideas that may only be expressed in a short format of words, symbols, and hypertext links". The dramatic popularity of Twitter has led to a nearly continuous growth of its users all over the world. It has become popular not only among ordinary people: many actors, public persons, including politicians, also use Twitter as a way to have their say in the Internet. Thus, Twitter serves as an effective platform for communication and, mainly: "Twitter supporters see it as a potential solution for many information sharing problems" (Golbeck, 2010: 1612).

As a new type of communication, Twitter has drawn the attention both of press (Fildes, 2010; Cain Miller, 2010) and scholars. The most recent research deals with the collaborative nature of Twitter (Honeycutt & Herring, 2009), its use for educational purposes (Grosseck & Hollotescu, 2008), its role for informal work communication (Zhao & Rosson, 2009), etc. In the study of Twitter use by the U.S. Congress, Golbeck et al. (2010: 1612) find that "[…] Congress people are primarily using Twitter to disperse information, particularly links to news articles about themselves and to their blog posts, and to report on their daily activities". In other words, they use Twitter as a tool of self-promotion and/or additional advertisement in the Internet. However, in spite of the textual character of Twitter service, very little has been said about its linguistic side. The process of information sharing on Twitter goes through visual channel using text as its chief component. The main characteristic feature that singles out Twitter text from all other online texts lies in its restriction to 140 characters. That is why, I find it challenging to reveal how politicians are using this social platform as a means of online discourse. Mainly, I will concentrate on its linguistic side employing lexical analysis of Barack Obama's and Dmitry Medvedev's Twitter accounts. As the first world leader to use this microblogging service, Obama now is on the 4th place of the most popular Twitter users in the world[25]. At the same time, his Russian colleague adopted this experience opening Twitter account during his official visit to the USA in June 2010 and became the first Russian leader who used new technologies during his government.

Next section explains my steps in more detail.

2. Lexical analysis of Twitter use by Barack Obama and Dmitry Medvedev

2.1. General description of @BarackObama and @MedvedevRussiaE

The corpora under the study was collected from the official Twitter accounts of Barack Obama (@BarackObama) and Dmitry Medvedev (@MedvedevRussiaE) during June-January 2010-2011. The tweets appear in a reverse chronological order starting with the

25 According to www.twitaholic.com due to 30/01/2011.

most recent and finishing with the first published ones. The total number of messages posted during the above period by both presidents is 831: @BarackObama – 510 (mean=64), @MedvedevRussiaE – 321 (mean=40). Diagram 1 below demonstrates a decrease in Twitter use by the Russian president after July 2010, so, by January 2011, the difference in the number of messages equals 73. On the contrary, his American colleague sticks to a steady rhythm by posting 1,6 times more messages.

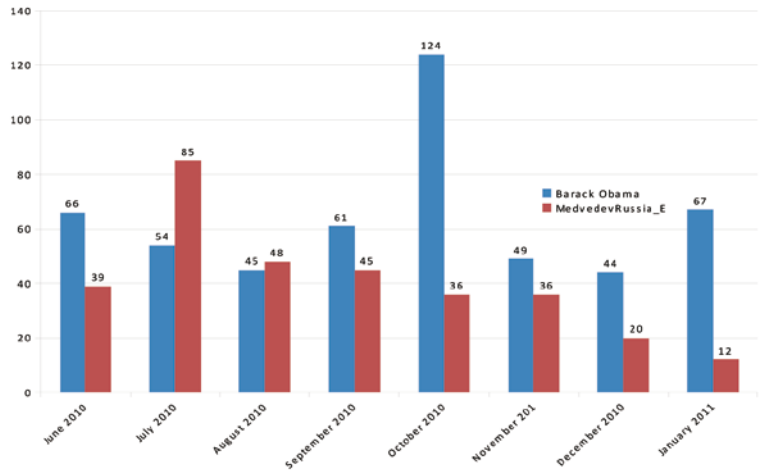


Diagram 1. Quantitative representation of Twitter use by Barack Obama and Dmitry Medvedev during the period June-January 2010-2011/Messages per month

Also, Diagram 2 shows no coincidence between Twitter use and presidents' work weeks, i.e., there is a tendency to post messages both during the week and at week-ends.

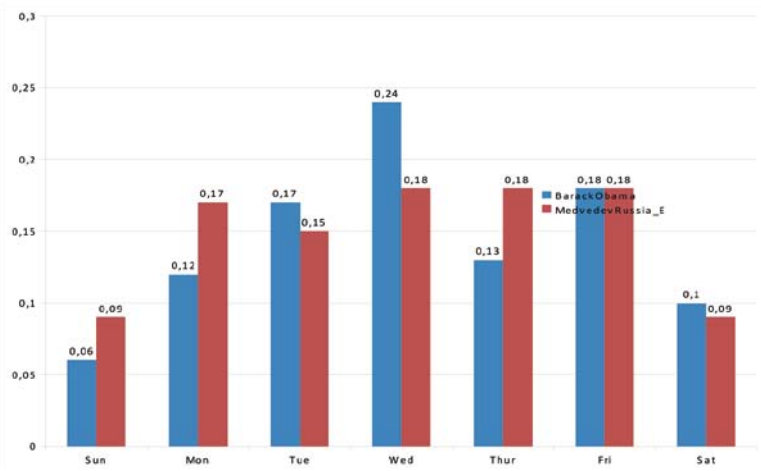


Diagram 2. Twitter use and presidents' work weeks

Golbeck et al. (2010) find that 0.72 of all tweets posted by the members of the U.S. Congress include links to the external web sites. It is not surprising if we take into account information sharing as the main Twitter objective. The results show similar situation in Obama's case where 0,68 of all tweets contained links; while in Medvedev's case this number was much low - 0.27 (0.61 of them are photos). These links are included into the bulk of a tweet (max. 140 characters), which, therefore, lessens its textual component. The analysis demonstrates that the mean for characters per message is 120 (range: 41-140; mode=139; StDev=21,63) for @BarackObama, and 116 (range: 16-140; mode=140; StDev=24.86) for @MedvedevRussiaE, i.e. nearly maximum use of available space. If we put together the results for the number of tweets with links and the mean for characters per each tweet, we see a clear prevalence of language component over the links (which have no contextual meaning) on Medvedev's account. Thus, he is using Twitter more for "talking" than for redirecting a user to a presidential personal web site or any other online platform. The last option refers to Obama's Twitter. However, taking into account linguistic focus of this study, the special emphasis was put on the language component of both Twitters as examples of political discourse online.

In the following section I conduct descriptive lexical analysis of both corpora.

2.2. Lexical characteristics of @BarackObama and @MedvedevRussiaE

The first step in the lexical analysis of @BarackObama and @MedvedevRussiaE was a compilation of corpus word list using a free software Simple Concordance Program 4.09. Table 1 presents basic results for the corpus word lists where a type would be a lexical unit as it would appear in a dictionary, and a token - the number of occurrences of that lexical unit in the entire corpus. "Types" and "tokens" are statistical units of SCP 4.09 to measure the length of the corpus. The relation types/tokens gives us the lexical variability and complexity of a corpus. It goes between 1 and 0. The richer the corpus is in lexical terms, the closer the value types/tokens is to 1, and the opposite.

Table 1. Basic lexical analysis of @BarackObama and @MedvedevRussiaE

	@BarackObama	@MedvedevRussiaE
Total vocabulary	1804 types	1794 types
Project wordcount	9327 tokens	5772 tokens
Types/tokens	0,19	0,31

Table 1 demonstrates that the correlation types/tokens is relatively low for both corpora which means that presidents' online discourses are concentrated around one topic with its zero further development across the corpora. Further on, high text readability index results (Gunning Fog=14,8 for @BarackObama, and 16.8 for @MedvedevRussiaE) imply that possible readers of presidents' Twitters are expected to have a University degree. This finding means it cannot be a simple/day-to-day online conversation with relatively easy and casual language component; rather, the theme should be more "technical". That is why, the next step of the analysis will be to explain this "technical" topic behind Obama's

and Medvedev's Twitters through word clouds as, in my opinion, they give a better visual representation of the corpus than traditional word lists.

Figure 1. Word clouds for Barack Obama's (left) and Dmitry Medvedev's (right) Twitter accounts

Taking into account that the bigger the word is in a word cloud, the more frequent it is in the corpus, one can clearly see that "watch" (N=97) and "live" (N=95) are the most popular words in Obama's corpus, while in Medvedev's one they are "Russia" (N=30) and "today" (N=29). It explains previous results about the prevalence in the percentage of links in Obama's messages, as both "watch" and "live" will probably be followed by an external video link. Another interesting feature of these two word clouds lies in the size (=frequency) of the most popular words, i.e. "watch" and "live" are much bigger (=more frequent) than other words in the cloud, as, e.g., "vote", "economy", etc.; while "Russia" and "today" do not differ in size so much with e.g. "law", "people" or "new". This observation is directly connected with the results on lexical density of the corpora where @MedvedevRussiaE had a bigger number than @BarackObama. The clouds also demonstrate that the lexical level of both corpora is restricted to political topic, which fully explains the above high readability index result.

However, if we also consider functional words, we will see that the most used word in presidents' Twitters is a personal pronoun "we" (N=128 for @BarackObama; N=63 for @MedvedevRussiaE). This fact indicates inclusive character of presidents' Twitter language which covers political side of their actions. Having discovered this, I am then interested in exact words used with the most popular one "we". For this purpose I built the concordance lists to show the most popular "we" collocates within the span 4:4. From Table 2 below it is seen that in Obama's corpus, <we 128> most frequently goes together with <can 22>, which as we remember was part of his campaign slogan "Yes We Can", where <we 128> includes both the president and the nation, as well as with collocates <win 13> and <need 12>. Consequently, the collocate <move 10> was mainly used with <forward 10> and <America 9>, while <fight 7> - with <your 8>.

In Medvedev's case the first four collocates <need 9, must 8, discussed 6, agreed 4> are explicit indications to the president's professional activities where <we 63> with <need 9> would probably be inclusive both for him, his cabinet and Russia (i.e. Russian people), while WE with <must 8>, <discussed 6> and <agreed 4> would include only President

and his cabinet. The rest of the collocates, i.e. <issues 4, summit 4, working 4, energy 3, global 3, russia 3> also deal with professional side of president's life.

@BarackObama	@MedvedevRussiaE
can 22	need 9
our 13	must 8
win 13	discussed 6
need 12	agreed 4
move 10	issues 4
forward 10	summit 4
you 10	working 4
America 9	energy 3
your 8	global 3
fight 7	Russia 3

Table 2. WE collocate list for @BarackObama and @MedvedevRussiaE

Thus, the corpora visualization through word clouds together with "we" collocates indicate strong political orientation of Obama's and Medvedev's Twitter accounts. One might suggest it was quite obvious from the very beginning; however, judging by other Twitter accounts of public people (e.g., Ashton Kutcher, an American actor), I claim that the lexical component is not restricted to their professional life only, on the contrary, it includes more private information as a way of self-promotion and fans attraction. Apparently, this is not the case of this study where both the American and the Russian presidents keep their Twitters exclusively for professional purposes to inform people about their political actions only.

3. Conclusions

This present paper was a contribution to Presidential Studies with the main focus on the language component of political discourse online using as the examples Barack Obama's and Dmitry Medvedev's Twitter platforms.

The central point of this study was to infer the language of state leaders while communicating with their audience in popular microblogging network using 140 symbols. The careful analysis of the corpora showed nearly maximum use of available characters per message which in its turn indicates presidents' extensive use of Twitter service. However, general statistics for the period of June-January 2010-2011 reveals considerable decrease in the number of posted messages on Medvedev's account. The reason for that is not clear; though one might suggest that this kind of online behavior is not typical for a public person who normally tries to stick to the same rhythm in Twitter use.

The most frequent words and "we" collocates of the corpus show a clear political orientation of both Twitters with no introduction of a new topic. Low lexical density revealed its

static and monothematic character concentrated mainly on the professional side of Barack Obama and Dmitry Medvedev. Gunning Fog index classified corpora as texts which expect readers with a University degree. Thus, Obama's and Medvedev's Twitters have restrictions at two levels: internal (lexical) and external (perception by the audience). In other words, lexical component determines a priori the future public of a text, or in this case, the future readers of presidents' Twitters. Having said that, I conclude that Twitter accounts of the U.S. and the Russian Presidents do not serve for the wide audience of this social platform and the Internet in general; rather, they are aimed at a restricted number of population due to its topic and composite readability factor. In political communication, Twitter is then defined as a powerful advertisement tool (for the wide spread of the global network) to give additional information about the presidents and their cabinets's actions.

To test the above conclusions, I propose to continue linguistic research of Twitter use as a means of online political communication. An extention of the present corpora by including other political leaders' Twitters will be a good way to go towards a more profound cross-cultural research in this area. Additionally, it would be challenging to reveal Twitter characteristics on the grammatical and syntactic levels to elicit, e.g., any cases of their simplifications caused by space restriction of each tweet. These results will help draw a more complete picture of political language on Twitter platform. Moreover, being part of the linguistic studies on Twitter, they will also contribute as an attempt to single out and describe Twitter text as an independent type of online texts.

References

Cain Miller, C. (2010). Twitter Unveils Plans to Draw Money from Ads. *The New York Times*. Retrieved from http://www.nytimes.com/2010/04/13/technology/internet/13twitter.html?_r=1&partner=rss&emc=rss

Fildes, J. (2010). Twitter scrambles to block worms. Retrieved from http://www.bbc.co.uk/news/technology-11382469

Golbeck, J., Grimes, J. M., & Rogers, A. (2010). Twitter Use by the U.S. Congress. *Journal of the American Society for Information Science and Technology 61*(8), 1612-1621.

Grosseck, G., & Holotesku, C. (2008). Can we use Twitter for educational activities? Retrieved from http://www.morsmal.org/documents/members/admin/Can-we-use-Twitter-for-educational-activities.pdf

Honeycutt, C., & Herring., S. (2009). Beyond Microblogging: Conversation and Collaboration via Twitter. Paper presented at the *42nd Hawaii International Conference on System Sciences.* Los Alamitos, CA, January, 1-10.

Sagolla, D. (2010). 140 characters. Retrieved from http://www.140characters.com

Zhao, D., & Rosson, M. B. (2009). How and Why People Twitter: The Role that Micro-blogging Plays in Informal Communication at Work. Paper presented at the *ACM 2009 international conference on Supporting group work*, New York, May, 243-252.

Electronic deconstruction of an argument's rhetorical structure using its discussion forum supplement

Kieran O'Halloran

Open University, UK

Abstract:

A recent technological innovation is the appending of electronic discussion forums to online arguments, such as carried by online newspapers. This facility allows readers to post responses to an argument and to debate issues raised in it. Such discussion forums can be regarded as supplements to these arguments.

I highlight the utility value of this electronic supplementarity for critical reading of arguments. I show how, using corpus linguistic software, keyword analysis of a discussion forum appended to an argument can illuminate whether or not the rhetorical structure of the argument is unstable, is in a state of deconstruction. Since I use corpus linguistic method to conduct this content analysis, I refer to this approach to critical reading as *Electronic Deconstruction*. The theoretical stimulus for this approach comes from the work of the French philosopher, Jacques Derrida.

Keywords:

Argumentation, cohesion, critical discourse analysis, critical reading, electronic supplementarity, Jacques Derrida, keywords, rhetorical structure.

1. Introduction

In the last few years, one technological innovation has been the appending of electronic discussion forum facilities to online arguments, such as in online versions of newspapers. The facility allows readers to post responses to an argument and to debate issues raised in it. Such discussion forums can be regarded as supplements to these arguments.

The aim of this article is to highlight the utility value of this electronic supplementarity for critical reading of arguments. I show how a content analysis of a discussion forum appended to an argument can illuminate whether or not the rhetorical structure of the argument is unstable, is in a state of deconstruction. Since I use corpus linguistic method to conduct such content analysis - specifically 'keyword analysis' - I refer to this approach to critical reading as *Electronic Deconstruction*. The theoretical stimulus for this approach comes from the work of the French philosopher, Jacques Derrida. To be clear, this is an appropriation of Derrida's work. Though Derrida is synonymous with an approach to critical reading called 'Deconstruction', I am not doing Derridean Deconstruction.

In the next section, I outline keyword analysis. In Section 3, I highlight how a discussion forum can be treated as a 'supplement' in a sense related to that used in Derrida (1976[1967]). Section 4 includes an extract from the argument I analyse and a keyword analysis of the discussion forum which supplements this argument. I highlight keywords which are absent or marginal from the argument. In Section 5, by replacing absence / marginalisation in the argument with keywords from the forum, I highlight how this leads to instability in the cohesive structure of the argument and, in turn, its rhetorical structure.

2. Keyword analysis

2.1 Exclusions and marginalisations

All arguments place certain concepts at the centre of attention, with others excluded or marginalised. This may be because it is in the author's interests to exclude and marginalise the following: certain concepts which are habitually discussed in relation to the argument's topic as their inclusion may weaken the rhetorical power of the argument. A reader may intuitively detect such absences / marginalisations in relation to how a topic is habitually discussed. However, it could be difficult for a reader to avoid charges that they have detected absences and marginalisations arbitrarily; there are, after all, a potentially large number of absences from any text and not every absence is a candidate for repression. Since any argument will have a centre of attention, logically it must have a margin; but not everything which is marginal in an argument can be considered to be marginalised. Furthermore, what if readers are not so familiar with the topic of the argument? They would not then be in a position to assess, or perhaps even notice, possible exclusions and marginalisations.

This is where an electronic discussion forum attached to an argument is valuable. While the quality of comment can be variable, and the interpersonal dimension can vary from polite to abusive, nevertheless the anonymity of online response offers participants, in this 'community of interest', freedom and room to manoeuvre conceptually. Given this, a discussion forum - should it be sufficiently large - can provide illumination of what normal usage is likely to be in relation to the topic of the originating argument. By normal

usage I am referring to *what* concepts are normally used in discussion of a particular topic whether these concepts are assented to or not. Should certain concepts be salient in the forum as a whole but absent from, or at best marginal in, the original argument, this can offer insights into what the argument might be said to repress or marginalise from normal conceptual discussion of the topic of the argument. Since the discussion forum is directly related to the original argument, then arbitrariness will have been considerably reduced in producing these insights.

2.2 Keywords

Salient concepts in an online discussion forum, or any collection of electronic texts, can be revealed through keyword analysis using appropriate corpus linguistic software. A keyword is 'a word which occurs with unusual frequency in a given text…by comparison with a reference corpus' (Scott 1997: 236). Keywords are established through statistical measures such as log likelihood (see Dunning, 1993). A log likelihood value of ≥ 7 ($p < 0.01$)[26] confers keyness on a word. The larger the log likelihood value, the greater the salience of the keyword. Importantly, the log likelihood value, as a statistical measure, reduces arbitrariness in what is selected as salient.

Comparison of keywords in a discussion forum with words in the original argument can be illuminating. For this article, the following types of keyword are relevant:

• keywords in the forum which are absent from the original argument. These are candidates for the status of repressed concepts from the argument.

• keywords in the forum which are used infrequently in the original argument. These are candidates for the status of marginalised concepts from the argument.

Keyword analysis should not only be quantitative. They also need to be qualitatively explored tounderstand how they are being used. Qualitative exploration of keywords in the discussion forum can, in turn, strengthen judgements as to marginalised / repressed candidacy in the original argument.

In order to mobilise how I use keywords in a discussion forum supplement for purposes of Electronic Deconstruction (see Section 5), I take as stimulus how Derrida conceives of the supplement.

3. Discussion forum as supplements

3.1 Derrida's supplement

We normally think of the word 'supplement' as meaning something extra, an add-on. And when things are added to, they usually get bigger. For Derrida, the supplement is more subtle. This is because he illuminates how a supplement adds only to replace a lack of something. So, while the supplement may seem like an add-on and thus *outside* that which is supplemented, in fact it becomes simultaneously *inside* that which it is added to.

26 In reporting statistical significance, $p < 0.01$ indicates a 1 in 100 likelihood that the result could occur purely by chance.

Derrida writes that every supplement:

> …harbors within itself two significations whose cohabitation is as strange as it is necessary […] [The supplement] adds only to replace. It intervenes or insinuates itself *in-the-place-of.* (Derrida, 1976[1967]: 144-5)

For Derrida, then, the 'logic of supplementarity' is an undecideable inside-outside relation (Derrida, 1976[1967]: 215). Take, for example, a shop sign emblazoned outside a bicycle shop. It is outside the shop not part of the inside. It is an add-on, an extra, signalling the nature of the shop and in so doing increases the scope of the shop to include material and information outside of it. However, in being outside the shop it adds to replace a 'lack' inside the shop - the shop cannot function unless it can attract custom. The shop sign is thus simultaneously an add-on and an essential part of the shop - it is both outside and oddly inside the shop as well.

3.2 Discussion forums as Derridean supplements

If we apply the logic of the supplement to online discussion forums appended to arguments, then a discussion forum is not just outside the original argument. It is not just an add-on, an extra. On the logic of the supplement, keywords in the forum absent from the argument can be perspectivised as 'lacking' in it. Furthermore, since keywords are generated non-arbitrarily, we have in turn a non-arbitrary basis for intervening in the argument; we can use these keywords to 'add to replace' what can be perspectivised as deficiency in normal discussion of the argument's topic, intervening to replace an absence *inside* the argument with keywords *outside* the argument. Via the logic of the supplement, the border between an argument and its discussion forum supplement is porous.

3.3 Using the supplement to investigate deconstruction in an argument

The next stage is to trace the extent to which this intervention in the argument 'to add to replace' leads to instability in its cohesion. An argument's rhetorical structure is dependent on effective cohesion. If cohesion is disturbed by this intervention, then the rhetorical structure of the argument is unstable. If the rhetorical structure deconstructs in this way, this can offer insights into repression or marginalisation in the argument *relative* to the particular supplement.

4. The data

4.1 The online argument: "The New Atheism"

On December 30th 2007, an argument appeared in the British newspaper *The Guardian* entitled 'The New Atheism'. Its author, Brendan O'Neill, used this expression to capture

a number of books published in 2006 and 2007 which set out atheistic arguments (e.g. Richard Dawkins' 'The God Delusion' (2006) and Christopher Hitchens' 'God is not Great' (2007)). The argument totals 926 words and consists of 42 sentences. The whole argument and the discussion forum appended to it can be found at:

http://www.guardian.co.uk/commentisfree/2007/dec/30/thenewatheism. The first two paragraphs below contain the gist of the argument; in Section 5, I perform an Electronic Deconstruction of paragraph 2:

1. 'New atheism' was the surprise political hit of 2007.

2. *God*-bashing books by Hitchens, Dawkins and other thinkers who come out in a rash

when they hear the word 'religion' flew out of the bookshops.

3. Philip Pullman's anti-divine Golden Compass hit the big screen.

4. Everywhere, *God* was exposed as a fraud and *God* botherers were given an intellectual lashing.

5. I am as atheistic as it gets.

6. But I will not be signing up to this shrill hectoring of the religious.

7. The new atheists have given atheism a bad name.

8. History's greatest atheists, or the 'old atheists' as we are now forced to call them, were humanistic and progressive, critical of religion because it expressed man's sense of higher moral purpose in a deeply flawed fashion.

9. The new atheists are screechy and intolerant; they see religion merely as an expression of mass ignorance and delusion.

10. Their aim seems to be, not only to bring *God* crashing back down to earth, but also to downgrade mankind itself. [emphasis added]

In the complete argument, there are 21 instances of 'religion' but only 2 instances of '(religious) belief'. 'Religion' is clearly a centre in the argument and 'religious belief' a marginal. As such, I might be tempted to treat 'religious belief' as a marginalised category, but this would be an arbitrary selection since, naturally, there are other infrequent words in the argument. I turn to the discussion forum to reduce arbitrariness of judgement here.

4.2 Keywords in the discussion forum

In the discussion forum appended to the argument, there are 365 individual posts, which in total come to 69, 252 words.[27] The software I use to find keywords in the forum is WMatrix (Rayson, 2008). Using this online tool, I compare a corpus of the discussion forum posts with a corpus of around 1 million words of written English, which WMatrix

27 At the time of publication of O'Neill's argument, *The Guardian* had a policy of closing a forum after 3 days.

accesses online.[28] In order to make my examination manageable, I use the keyword cloud function which shows the 100 highest keywords (see Figure 1; see also the Appendix for the log likelihood values for these keywords as well as their frequencies).[29]

Figure 1: Keyword cloud showing the 100 highest keywords in the discussion forum; keywords with higher log likelihood values are in larger font size.

4.3 Repression of 'faith' and marginalisation of 'belief'

As Figure 1 shows, 'faith' is a significant keyword in the forum. However, it is absent from the argument. The semantically close, 'belief', is also a significant keyword in the forum as are its cognates, 'beliefs', 'believe' and 'believers'. When I inspected the forum qualitatively, I found that 'faith' and 'belief' are used mostly in a way equivalent to religious belief. Generally speaking, these terms are used in the forum in relation to questioning faith / belief in a supernatural power or arguing against the new atheist position, especially Dawkins' perspective that belief in God is a delusion ('Dawkins', 'delusion' and 'God' are all keywords - see Figure 1). Posts are not just from atheists and agnostics but theists too.

'God' occurs four times in the argument, but only in the opening two paragraphs (see italicised in Section 4.1 extract). All instances of 'God' are metaphorical. Since the argument is in a newspaper, the use of metaphor here would seem to have an interpersonal function to help attract the reader into the argument by use of colourful imagery. However, one of the most common expressions in the forum containing either of the keywords 'belief' or 'believe' is 'belief/ve+in+supernatural being' such as 'belief in God'. Out of 439 instances of 'belief/ve' in the forum, a quarter (113 instances) are realised in this expression.

Given all this, we have non-arbitrary grounds for supposing that 'faith' is not only absent

28 WMatrix has online access to one reference corpus - the British National Corpus (BNC) Sampler. This consists of around 1 million words each of the BNC Sampler spoken corpus and the BNC Sampler written corpus (the whole of the BNC consists of 10 million words of spoken and 90 million words of written English). On the rule of thumb that 'we should at least try to obtain reference corpora which reflect some aspect of the smaller corpus or text sample we are studying' (Baker, 2006: 43), I chose to compare the argument with the BNC Sampler written corpus. Information can be obtained on the composition of the BNC Sampler written corpus at http://ucrel.lancs.ac.uk/bnc2sampler/sampler.htm

29 I have not used WMatrix to analyse the argument itself since, given the size of the text, judgements of relative frequency and salience of lexis are straightforward.

from the argument but is, in fact, repressed and, furthermore, that 'religious belief' is not only marginal in O'Neill's argument but is, in fact, marginalised. On the logic of the supplement, then, we can perspectivise the argument's God metaphors as lacking mention of 'belief' / 'faith'. That is, though on first sight the metaphors seem to have a journalistic interpersonal function, in fact they can be construed as having an excluding ideational function in proscribing the concepts of 'belief' / 'faith'. All my judgements here are, of course, made relative to this particular supplement since I have no way of accessing the author's mind.

In Section 5, I perform an electronic deconstruction of paragraph 2 of the argument - where the rhetorical opposition of new atheism versus old atheism is first introduced - by using the logic of the supplement to do the following: add the 'belief' keyword in the discussion forum ('belief' being equivalent to 'faith' in the forum) in order to replace its relative absence in the argument. Since 'belief' is keyword, this intervention is non-arbitrary. I trace the effects of this intervention on the cohesive structure of the argument and show how its rhetorical structure deconstructs.

5. Electronic deconstruction

Firstly, I 'add to replace' the lack of 'belief [in]' before 'God' in sentence 10. Again, on the logic of the supplement, I 'add to replace' the absence of 'religious belief' in sentence 9 (I could have used 'faith' also). I thus cross out 'religion' in sentence 9:

9. The new atheists are screechy and intolerant; they see ~~religion~~ *religious belief* merely as an expression of mass ignorance and delusion. [emphasis added]

10. Their aim seems to be, not only to bring *belief [in]* God crashing back down to earth, but also to *downgrade* mankind itself. [emphasis added]

Sentences 9 and 10 are now linked by the specific category of '(religious) belief'. Cohesion in sentence 9 is troubled through this intervention since a tension is now created between 'religious belief' and 'intolerant'. Intuitively, it is difficult to see how it is possible to be intolerant of a mental state - that is, the mental states of religious believers. The evidence in Table 1 supports this intuition. While there is plenty of evidence of forms of the lemma, TOLERANCE, collocating with 'religion(s)' and 'religious' and with some very significant t-scores, there is little evidence of forms of the lemma TOLERANCE collocating with the category, 'religious belief(s)'. The argument's instability, thus, starts to become apparent through intervening on the basis of the discussion forum supplement.

While both sentences 9 and 10 refer to new atheism, old atheism is first mentioned in sentence 8:

8. History's greatest atheists, or the 'old atheists' as we are now forced to call them, were humanistic and progressive, critical of religion because it expressed man's sense of *higher* moral purpose in a deeply flawed fashion. [emphasis added]

Notice the 'old atheism / new atheism' binary opposition is associated with another binary opposition, high /low, i.e., 'higher moral purpose' (sentence 8 – see italics) and 'downgrade mankind' (sentence 10 – see italics).

The original sentences 8 and 9 link lexically through the general category, 'religion'. But, following my intervention on the basis of the supplement, like-for-like cohesion no longer exists. The intervened cohesive structure of sentences 8-10, thus, vibrates with instability which, in turn, reveals deconstruction in the rhetorical structure of old atheism as high / new atheism as low.[30]

Collocation values for frequency and t-score						
	religion(s)		religious		religious belief(s)	
	freq.	t-score	freq.	t-score	freq.	t-score
intolerance	40	6.3	267	16.3	5	2.2
intolerant	35	5.9	18	4.2		
tolerance	106	10.2	386	19.6	8	2.8
tolerant	76	8.7	46	6.7	4	1.9
toleration	21	4.6	234	15.3		
Intolerance	4	2.0	17	4.1		
Tolerance	27	5.2	69	8.3		
tolerated	31	5.5	14	3.6		
tolerate	20	4.4	14	3.6		
tolerating	6	2.4	3	1.7		
Toleration	4	2.0	18	4.2		

Table 1 Frequency and t-score values for collocation in the 1.5 billion word corpus, UKWaC, of 'religion(s)', 'religious', 'religious belief(s)' with the lemma TOLERANCE for -5+5 word span; values are for both lower-case and initial capital letter instances of TOLERANCE. T-scores over 2 are 'normally taken to be significant' (Hunston, 2002: 72); t-scores over 10 are very significant (Hunston, 2001: 16).

6. Conclusion

I have exploited the utility value of an electronic supplement to the argument which originated it in order to offer insights into what concepts the argument can be reasonably said to repress and marginalise. On the logic of the supplement, I replaced deficiency in the argument. Drawing from a discussion forum, I showed how the argument's lexical cohesive structure became undone by this 'adding to replace' and thus that the rhetorical structure deconstructed. On this basis, I am able to illuminate how the category of

30 One can only speculate why O'Neill seems to repress the specific category, 'faith', or marginalise the specific category, 'religious belief', via use of the general category, 'religion'. One possibility is that since he professes to be 'as atheistic as it gets' (sentence 5), it would not serve his argument to mention too often the more specific categories 'religious belief' or 'faith'. Since these elements of religion are the most vulnerable to attack from atheists - i.e., including O'Neill - it would be better for him to (attempt to) repress or marginalise these concepts in the argument in order to help avoid contradicting himself.

'religion', at 21 instances, works in organising a centre in O'Neill's argument. It does this through being strategically vague, allowing the marginalisation / exclusion of the specific terms 'religious belief' / 'faith', which deconstruct part of the rhetorical structure of the argument when included. It should be stressed, though, that how an argument is shown to undo itself is relative to the supplement used and, in turn, to the keywords generated.

Because the procedure for locating salient concepts in the forum is statistically informed, it reduces arbitrariness in making judgements of repressions and marginalisations as well as in selecting interventive points into the argument. However, notice I say 'reduces'. Inevitably, I have to make some arbitrary choices, e.g. of the number of keywords to examine so as to make my use of the supplement manageable. Exploring more keywords could lead to further revelation of instabilities in the argument relative to the supplement used.[31] Finally, this article has argued that in the new era of electronic supplementarity, argument borders are porous. As such, it has shown a new way of viewing the stability or otherwise of rhetorical structures in arguments.

Appendix

Table A1 The 100 highest keywords, including frequency ('Freq') and log likelihood values ('LL'), in the discussion forum

Keyword	Freq	LL
religion	336	1665.17
atheists	227	1249.02
religious	264	1127.97
god	200	1100.46
atheism	141	764.05
atheist	117	632.37
belief	141	604.86
i	1093	545.37
that	1232	525.72
Dawkins	90	495.21
n't	445	463.60
you	707	384.81
science	84	363.00
faith	86	339.49
beliefs	77	337.06
think	188	316.27

31 To augment rigour, the above analysis could be compared with other relevant supplements such as a corpus of discussion forum posts from atheist websites (e.g. http://richarddawkins.net), and ideally comments posted close to when O'Neill's article was published, to see if similar keywords are generated or not.

do	365	310.95
is	1195	304.30
what	327	249.59
Brendan	47	249.02
article	77	246.86
believe	114	232.50
'm	142	222.99
religions	44	220.18
Hitchens	40	220.09
Marx	44	210.46
people	239	207.66
Darwin	37	203.58
not	569	197.82
intolerant	36	189.02
cif	34	187.08
folks	34	187.08
does	146	185.47
delusion	32	167.24
believers	34	166.65
christians	44	165.24
humanity	31	161.80
hectoring	29	159.57
universe	34	150.52
it	884	146.14
read	71	146.01
just	187	145.03
catholic	45	142.57
scientific	41	136.10
why	117	130.57
shrill	25	129.21
human	57	125.86
actually	61	122.95
moral	40	122.48
about	230	120.18

Keyword	Freq	LL
argument	42	117.73
O'Neill	21	115.55
supernatural	27	114.91
christian	45	113.10
are	561	110.79
@	20	110.05
reason	63	107.19
intellectual	29	105.08
theists	19	104.54
Brendan_O'Neill	18	99.04
spirituality	18	99.04
their	323	98.59
they	425	96.08
as	562	94.48
or	387	94.46
Grayling	17	93.54
those	154	90.70
agree	38	88.74
metaphysical	16	88.04
your	200	85.10
christianity	26	82.71
spiritual	26	82.71
arguments	32	82.43
point	61	80.24
silly	26	78.35
've	75	76.02
have	432	75.98
islam	19	75.18
rational	19	75.18
secular	19	75.18
thinking	44	72.89
morality	18	72.84
--	13	71.53

Richard_Dawkins	13	71.53
hardtimethinking	13	71.53
enlightenment	17	70.62
ignorance	19	70.57
irrational	15	70.48
ideas	42	70.15
exist	30	68.31
so	201	67.76
new	180	66.73
intolerance	15	66.71
attacking	19	66.57
faustroll	12	66.03
trying	47	65.78
evolutionary	14	65.24
but	453	64.65
say	91	64.01
catholics	16	63.05

References

Baker, P. (2006). *Using Corpora in Discourse Analysis*, London: Continuum.

Dawkins, R. (2006). *The God Delusion*. London: Black Swann.

Derrida, J. (1976[1967]). *Of Grammatology* [trans G.C. Spivak]. Baltimore: Johns Hopkins. University Press.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61–74.

Hitchens, C. (2007). *God is not Great*. New York: Atlantic Books.

Hunston, S. (2001). Colligation, Lexis, Pattern, and Text. In M. Scott and G. Thompson (Eds.), *Patterns of Text: In Honour of Michael Hoey* (pp. 13–33). Amsterdam: John Benjamins.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Rayson, P. (2008). Wmatrix: A web-based corpus processing environment. Computing Department, Lancaster University. http://ucrel.lancs.ac.uk/wmatrix.

Scott, M. (1997). PC analysis of key words – and key key words. *System* 25(2), 233–45.

Building a comparable corpus (English-Spanish) of newspaper articles on gender and sexual (in)equality (GENTEXT-N): Present and future applications in the analysis of socio-ideological discourses

José Santaemilia
Sergio Maruenda
Universitat de València

Abstract:

As part of the work of the research group GENTEXT[32], we are building a 35 million-word, comparable (Spanish-English), highly-specialised corpus (GENTEXT-N) from Spanish and British newspapers, which serves to analyse, document and offer insights into the complex socio-ideological debates generated by a number of sexual-equality legal measures that have been adopted in Spain and the United Kingdom over the last few years.

To this end, a combination of qualitative and quantitative analyses (as advocated, among others, by Baker & McEnery 2005, Baker et al 2008, Caldas-Coulthard & Mund 2010) is essential if we wish to grasp both the linguistic and the ideological underpinnings of the heterogeneous texts we are investigating. In this paper we will explore two of the potentialities of corpus analysis –the study of frequent words/keywords and the study of collocations and semantic/discourse prosodies, as applied to our corpus.

Keywords: gender – sexual (in)equality – critical discourse analysis (CDA) – corpus analysis – keywords – semantic prosodies – discourse prosodies

Resumen:

Como parte de las investigaciones realizadas por parte del grupo GENTEXT, estamos compilando un corpus comparable (español-inglés), muy especializado (GENTEXT-N), y procedente de periódicos españoles y británicos, que cuenta con unos 35 millones de palabras, y que pretende analizar, documentar y explicar los complejos debates socio-ideológicos generados por una serie de medidas legislativas aprobadas en España y en el Reino Unido durante los últimos años, con el objetivo de alcanzar la igualdad sexual. Para ello, es esencial utilizar análisis tanto cuantitativos como cualitativos (como proponen, entre otros, Baker & McEnery 2005, Baker et al 2008, Caldas-Coulthard & Mund 2010), si queremos llegar a los fundamentos lingüísticas e ideológicos de los heterogéneos textos que estamos investigando. En este artículo nos centramos en dos de las potencialidades del análisis de corpus: el estudio de las palabras clave o más frecuentes, así como el estudio de los colocados y la prosodia semántica/discursiva, aplicadas a nuestro corpus.

Palabras clave: género – (des)igualdad sexual – análisis crítico del discurso – análisis de corpus – palabras clave – prosodia semántica – prosodia discursiva

32 "Género y (des)igualdad sexual en las sociedades española y británica contemporáneas: Documentación y análisis discursivo de textos socio-ideológicos", under a Research Project from the *Ministerio de Ciencia e Innovación* (FFI2008-04534/FILO).

0. Context: Legal measures towards sexual equality

Over the last few years a number of legal measures have been adopted recently both in Spain and in the UK on key gender-related issues –abortion, gender-based violence or homosexual marriages. These measures, along with the growing recognition of social and sexual rights in Western Europe, have sparked a heated debate within both Spanish and British societies. These debates are reproduced, generated, amplified, diminished, perverted or exploited by mass media, political parties or religious institutions.

1. GENTEXT: Gender, Language & Sexual (In)Equality Research Group

Our research group (GENTEXT – Gender, Language and Sexual (In)Equality), based at the *Universitat de València*, has set out to study this growing and complex discursive reality. The aim of our research is to document and analyze the concepts, the discursive processes, the ideological tensions and the semantic negotiation behind all these sexual equality measures and subsequent social debate.

At present, our corpus (GENTEXT-N) is made up of newspaper articles from Spanish and British dailies (embodying liberal vs conservative positions), which include Spanish dailies *El País* and *El Mundo*, and British dailies *The Guardian* and *The Times*. The corpus contains around 35 million words, and our research is centred around the analysis of key terms and concepts such as *gender-based* or *domestic violence*, *homosexual(ity)*, *gay*, and *abortion*.

2. The study: Basic assumptions

Some of the main assumptions under which our study is based are that:

– we believe that large, ad-hoc corpora can offer invaluable insights into discursive practices of a social nature.

– we analyse a large amount of texts, but we always need to pay attention to context, which is inextricably linked to social actors, to historical circumstances, to ideological factors, to power asymmetries, etc.

– the driving force of this new language/discourse which is appearing around gender-related legislation and its ensuing social debate is a combination of socio-ideological tensions and semantic/pragmatic negotiation.

All the information provided by the different newspapers is the result of a large-scale process of negotiation at the centre of which is *discourse*, which is defined by Burr (1995: 48) *–à la Foucault–* as

> a set of meanings, metaphors, representations, images, stories, statements, and so on that in some way together produce a particular version of events (…) Surrounding any one object, event, person, etc., there may be a variety of different discourses, each within a different story to tell about the world, a different way of representing it to the world.

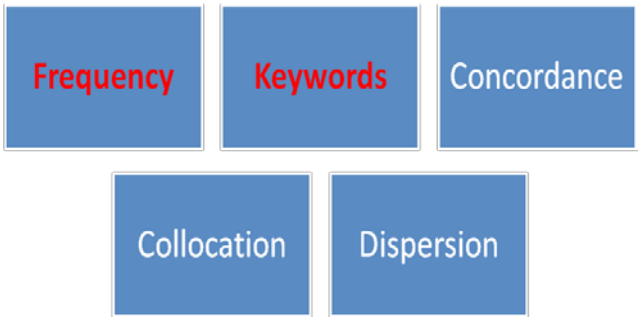
Our basic methodological assumption is that, in order to analyse this complex reality, we need a combination of methods (which are plural themselves): *corpus linguistics* (CL), *critical discourse analysis* (CDA) and *lexical pragmatics* (LP). Due to the space limitations of this publication, in this paper we are going to concentrate on the term *homosexual(idad)*, as found in all the issues of the Spanish dailies *El País* and *El Mundo* for the year 2005, the year in which Bill no. 121/000018 (*Proyecto de Ley por la que se modifica el Código Civil en materia de derecho a contraer matrimonio*) was passed (on 30 June 2005), which ammended the Civil Code to grant full marriage rights to gay couples.

Our main research question is: how far can CL + CDA + LP help to uncover (ideological) discourse(s)? And how? To address this complex question, in this paper we will concentrate on the study of frequent words/keywords; and the study of collocations and semantic/discourse prosodies, as applied to a segment of our corpus.

3. Main corpus processes. (Most) frequent vs key words

CL is paramount in today's linguistic and discursive research as it can pinpoint areas of interest for more local and refined analysis. But CL is not a single or unified methodology of analysis; rather, as stated by Baker (2010: 19) it is "a *collection* of methods". In Table 1 we can see the main processes or techniques of CL:

Table 1. Main processes of techniques of corpus linguistics.



Depending on our aims, some or all of these processes are likely to be combined in order to produce more reliable results. In this section, we will focus on *frequency* and *keywords* from our *ad hoc* corpus. Table 2 shows the total number of words in the two sub-corpora:

Table 2.- Total number of words (El País and El Mundo)

El País (2005)	***El Mundo*** (2005)	**TOTAL**
327,243	353,051	680,294

3.1. Frequency

Though valuable, mere frequency means very little. Let us draw two frequency lists, one from *El País* and the other from *El Mundo*, which can be compared in search of initial similarities or differences:

Table 3.- The ten most frequent words in El País and El Mundo sub-corpora.

Matrimonio	940	Matrimonio	917
Civil	818	Homosexuales	713
Homosexual	721	Homosexual	684
Ley	598	Ley	587
Homosexuales	590	Años	661
Años	484	Dos	517
Dos	478	Todo	496
Personas	406	Gobierno	490
Iglesia	389	Mismo	472
Gobierno	340	España	458

To generate Table 3, we can use *WordSmith* or *AntConc* or any other similar software. Grammatical words (*de, la, que, el, en, y, a* and so on) are usually left out, as they thought to add little semantic value; however, they are revealing sometimes (e.g. *entre, contra, para, como*).

Focusing on a specific word which appears frequently in our corpus is a good analytical procedure. Let us concentrate on the lexeme *homosexual*. Table 4 gives us a few statistics of its morphological variants:

Table 4.- Number of occurrences of the lexeme homosexual

	EP	EM
Homos	---	2
Homosexual	731	738
Homosexuales	598	781
Homosexualidad	188	155
Homosexualidades	1	---
Homosexualismo	1	1

The figures for *homosexual* are illustrative, but they are even more so if we compare them with those for its opposite, *heterosexual*, which we find in Table 5:

Table 5.- Number of occurrences of the lexeme heterosexual

	EP	EM
Heterogays	---	2
Heteros	3	7
Heterosexismo	---	4
Heterosexual	92	68
Heterosexuales	99	74
Heterosexualidad	20	13

When we compare, then frequencies become more important. But higher frequency doesn't mean necessarily that the term is preferable or even more important. The term *homosexual* is much more frequent than *heterosexual* –i.e. it is 'marked'– but the interpretation is far from straightforward. For Baker (2010: 126), "homosexuality is marked *because* society views it as unusual". Heterosexuality, by contrast, is unmarked, in need of no explanation.

In Tables 4 and 5, the presence or absence of certain terms in specific dailies is also significant: terms such as *homos*/*heteros* or surprising coinings such as *heterogays* might be indicative of ideological stance or resistance, which could only be made evident through analysis of context. Other terms or pairs of terms could also be explored (*hombre* v *mujer*, *matrimonio* v *pareja*, *hijo* v *hija*, etc.), perhaps yielding revealing results.

To summarise, frequency counts are potentially important analytical weapons which can help the analyst to uncover evidence for bias, to reveal the main of a text or a corpus, or more generally to suggest areas worth examining in our material.

3.2. Keywords

Another type of analysis we may carry out is a statistical keyword analysis, using *WordSmith Tools 5.0*. As compared with frequency counts, it constitutes a more sophisticated type of analysis. Keywords are the result of comparing two corpora against each other, one of them being a small corpus and the other a larger reference corpus. A *keyword* is, in short, "a word which occurs statistically more frequently in a single text or corpus than in another text or corpus" (Baker, 2010: 133). Lists of *positive* or *negative* keywords can be compiled –*positive* being those words with unusually high frequency, and *negative* those with unusually low frequency.

The resulting keywords indicate saliency or aboutness, and point to the favoured lexical areas or concepts. And they rely on *expectations* about language or ideology, thus yielding extremely helpful information for language learning, genre and register analysis. Its potentialities, which need to be explored further, lie mainly the analysis of lexical keywords as indicators of discursive or rhetorical strategies, from naming strategies to argumentation, from predication to discursive construction of key gendered or sexualised terms –e.g. *homosexualidad, familia* or *matrimonio*. They are also a privileged tool for the critical analysis of ideology, as keywords are not neutral or statistical lists of words, but rather privileged rhetorical devices used to implant common sense in our ways of thinking. Its applications can be found in genre or register analysis, as already mentioned, but also in the growing field of (teaching and learning) specialized language and translation.

4. From collocation to discourse/semantic constellations: A study on naming practices

Apart from the initial focus on keyword in these gender-sensitive texts, we also examine the potentialities of collocations and semantic/discourse prosodies for our research (Louw 1993, Stubbs 2001). As for the former, it will be revealing to document the collocations certain keywords (e.g. *homosexual, abortion* or *violence*) give rise to. As for the latter, semantic/discourse prosodies help transcend the collocational or even sentential scope to reveal discursive patterns and, consequently, to trace evaluative relations (with participants in discourse) in terms of ideological standpoint (see Martin & White 2005). These constitute a network or constellation of semantic concepts which contribute to shaping and (de)legitimising citizens' discourses and rhetorical frameworks within communities of practice.

Here we would like to focus on the study of 'naming practices' for 'homosexual marriage' as a result of the same-sex marriage law (2005) in Spain. Our main aim is to see whether CL can help to unveil ideological discourses, and more specifically, to explore evaluation and stance by means of semantic/discourse prosodies. The phenomenon of *discourse prosody* was initially discussed by Sinclair (1991) in terms of positive/negative connotation of a word on the basis of its recurrent co-texts: if a word is typically associated with negatively evaluated entities, then it is likely to have a negative 'prosody'. For Sinclair (2000), the idea behind the concept of *prosody* is that an awareness of both the referential and the evaluative/attitudinal aspects of meaning is necessary for accurate deployment of the lexical item. So prosody can shed light on point of view or ideology in discourse.

Hunston (2002) uses the example of *illegal immigrant*. The recurrent combination of these two items may help spread the assumption that going from one country to another is wrong, that all immigration is wrong or even that all immigration is illegal. This assumption may be retrieved even when the term *immigrant* is used without *illegal*. It is evident that this may have a great impact on our attitudes, as can be seen in the opposition between e.g. *matrimonio homosexual* vs *unión homosexual*.

Our work offers an initial analysis of the key semantic sets regarding naming practices for:

- ✓ 'people' (e.g. *gay couples, homosexual couples, partner*, etc) and
- ✓ 'relationships' (e.g. *matrimonio homosexual, matrimonio gay, pareja de hecho, homosexual couples, civil partnership, same-sex partnerships,* etc).

A simple glance at our two sub-corpora evinced that there were disparate reactions to the new legislation, conceptualised in terms of at least two discourses:

(1) the provision of marriage rights to homosexual couples; and

(2) the depiction of the new law as a fierce attack on the institution of marriage and an infringement of traditional family rights.

Thus, whilst *El País* abounds in op-ed or think pieces (where their writers take up a stand against right-wing and Church leaders opposing the granting of civil rights to homosexual couples), *El Mundo* contributes to the debate generated among social actors and communities of practice by giving authoritative voice through attributed discourse.

Although we acknowledge the relevance and significance of discourse prosodies in our search for *traces* of ideological discourses, we believe that the notion needs to be redefined for the following reasons. On the one hand, it presupposes a semantically-based positive or negative orientation in *all* lexical items, thus minimising the role of pragmatics in deriving intended meaning. Qualitative analysis suggests, nevertheless, that the meanings of individual forms surrounding the term *matrimonio homosexual* are dynamic and constantly renegotiated in discourse through the exploitation of further evaluative slants. As Baker (2010: 128) points out

> We should not assume that everyone experiences and processes language in the same way. many people approach their encounters with certain types of language in a critical way and this may 'immunise' them to the ideologies inherent with certain collocational patterns.

For our purposes, and without further precision here, we prefer the term *discourse constellations* to refer to a form of organising the multiplicity of conceptual representations subject to ideological negotiation and social and political pressure in/between communities of practice. These are nebulous realizations of conflicting ideological concepts/discourses in today's societies and as such they are imprecise and constantly changing, in continuous struggle to become legitimised or *core*. If we consider the case of *homosexual marriage*, we can observe that official characterisations (i.e. laws) are peripheral and perhaps sociologically minoritarian vs other 'unofficial' ones (i.e. Popular Party and the Church) that hold enormous power and influence. Associated with the semantic sub-constellation of the Popular Party we find elements such as 'aberration', 'attack against the family', 'the true family', 'undermines (family and God)', 'confuses moral order and education'; etc. On the opposite shore, the sub-constellation of the left-wing party features elements such as 'coexistence', 'diversity', 'free', 'egalitarian', 'support to the homes', 'extending rights', 'legalisation', etc.

5. Conclusions

It is worth taking advantage of both quantitative (corpus linguistics) and qualitative (CDA, lexical pragmatics) approches in order to further investigate the legal, sexual or linguistic implications of our GENTEXT corpus. In particular, frequency/keyword lists and semantic/discourse constellations have proved useful in uncovering not only lexical or grammatical preference but also ideological bias or stance. In future analyses, we must ensure a more thorough integration of quantitative and qualitative methodologies, addressing the strengths and weaknesses of both.

Bibliography

Baker, Paul (2010). *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Baker, Paul *et al* (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse and Society*, (19)3, 273-306.

Baker, Paul & Tony McEnery (2005). A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts. *Journal of Language and Politics*, (4)2, 197-226.

Burr, V. (1995). *An Introduction to Social Constructionism*. London: Sage.

Caldas-Coulthard, Carmen R. & Rosamund Mund (2010). 'Curvy, hunky, kinky': Using corpora as tools for critical analysis. *Discourse & Society*, (21)2, 99-133.

Hunston, S. (2002), *Corpora in Applied Linguistics*. Cambridge: CUP.

Louw, William (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In Mona Baker *et al* (Eds.), *Text & Technology: In honour of John Sinclair* (pp. 157-176). Philadelphia/Amsterdam: John Benjamins.

Martin, J.R. & P.R.R. White (2005). *The Language of Evaluation: Appraisal in English*. London: Palgrave Macmillan.

Sinclair, John (1991). *Corpus, Concordance, Collocation*. Oxford: OUP.

Sinclair, John (2000). Lexical Grammar. *Naujoji Metodologija*, 24, 191-203.

Stubbs, Michael (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

METAPHOR identification in corpora: the case of 'as' in a business periodical corpus

Hanna Skorczynska

Universitat Politècnica de València

This study analyzes the phraseological patterns of 'as' used as a metaphor signal in a corpus of business periodical articles. The proposal made here is that the use of phrases instead of words in concordancing techniques may facilitate the identification of signaled metaphors in a corpus. The analysis of the co-text of 'as', when it signaled a metaphor in the corpus, revealed certain phraseological variation in the anticipation of metaphors. The main pattern detected was the combination of a verb with 'as'. 11 of these verbs were identified as recurrently used with 'as' in signaling a metaphor: 'act', 'call', 'describe', 'know', 'look', 'perceive', 'refer to', 'see', 'think of', 'use', and 'view'. 'View as' registered the highest probability of signaling a metaphor, while 'call as' was the least probable metaphor-signaling phrase.

Key words: metaphor, metaphor signal, metaphor identification, corpus

Este estudio analiza los patrones fraseológicos de 'as' usado como una señal metafórica en un corpus de artículos de revistas de negocios de divulgación. La propuesta que se formula es la utilización de combinaciones de palabras en lugar de palabras individuales en las técnicas de concordancia, lo cual podría facilitar la identificación de metáforas señalizadas en un corpus. El análisis del cotexto de 'as' cuando señalizó una metáfora reveló ciertas variaciones fraseológicas en la anticipación de la metáfora. El patrón principal detectado consistió en la combinación de un verbo con 'as'. Once de estos verbos fueron identificados como recurrentes al usare 'as' para señalizar una metáfora: 'act', 'call', 'describe', 'know', 'look', 'perceive', 'refer to', 'see', 'think of', 'use', and 'view'. 'View as' fue la unidad fraseológica que con más probabilidad señaliza una metáfora , mientras que 'call as' fue la menos probable.

Palabras clave: metáfora, señal metafórica, identificación de la metáfora, corpus

1. Metaphor identification in corpora

The identification of metaphors in corpora is neither a simple nor a straightforward task for a metaphor scholar. One of the reasons is that the context of an expression has to be analyzed in order to decide whether it bears a metaphorical or a literal meaning. This requires corpora to be searched manually, and a manual search is typically laborious and time-consuming. An additional problem with metaphor identification is that the corpus query software available is designed for word searches, while in the case of metaphors, what is searched for are not words, but meanings (Philip, 2010).

To overcome this problem, metaphor scholars working with corpora use a combination of the manual search and automatic techniques. In a two-stage procedure (Charteris-Black, 2004), first a small corpus or a sub-corpus of a larger collection of texts is manually analyzed in order to identify the so-callled 'key metaphors'. In the second stage, the key metaphors are searched for in the larger corpus using concordancing techniques. Deignan (2005) further proposes evaluating the metaphorical meaning of the concordances obtained by going back to the context of a particular concordance or to the text containing that concordance. The main drawback of this method is that the results obtained are limited to the key metaphors identified in the sub-corpus or in the small corpus.

Philip (2010) proposes a semi-automatic method to identify metaphor themes (Black, 1993) or conceptual metaphors in a specialized corpus. First, the key word programme (Scott, 2004) is used to identify the corpus key words. After that, the corpus frequency word list is manually analyzed to find the semantically incongruous items as compared to the semantic sets of the key words. The metaphor candidates determined in this way are then used to run a concordancer. The concordances obtained are manually analyzed to select metaphors used. Finally, the concordances are examined to find metaphor collocational patterns representing source domains, with the key words, representing target domains. In this way, the corpus metaphor themes can be identified.

In other metaphor identification procedures, words that would signal the presence of a metaphor in discourse were used as a short-cut to the signaled metaphorical material in a corpus (Skorczynska & Piqué, 2005). In this case, metaphor signals identified elsewhere (Goatly, 1997) were used as search words to run a concordancer. The concordances obtained were analyzed with regard to the metaphorical co-text of the signals searched for. The results obtained in this way are limited to the cases of signaled metaphors, leaving unidentified all those metaphors that were not deliberately pointed to in the corpus.

Despite the shortcomings of the method described, metaphor signals clearly provide a direct access to the signaled metaphorical material in large corpora if concordancing techniques are chosen to be used. This method can prove to be useful as a complementary search in other manual and partially manual metaphor identification procedures. In order to further turn it more effective for specialized corpora, the present study suggests using metaphor signaling phrases instead of the single words, which were proposed by Goatly (1997) and Wallington, Barnden, Barnden, Ferguson and Glasbey (2003). To this end, the co-text of a metaphor signal 'as' was examined in a corpus of business periodical articles and certain phraseological patterns were identified. Using phraseological chunks that

typically signal metaphors in a particular discourse could render metaphor identification less time-consuming and laborious.

2. Metaphor signals

Metaphor signals, also called metaphorical markers (Goatly, 1997), tuning devices (Cameron & Deignan, 2003) and flagging expressions (Steen, 2007), are words and phrases that anticipate metaphors in discourse and are used to cue the reader/listener into the metaphorical rather than the literal interpretation of an expression.

Early studies of metaphor co-text, and so of metaphor signaling focused on the differences between similes and metaphors (Miller, 1993; Ortony, 1993), and analyzed selected 'hedges' from the point of view of their effects on the perceived metaphoricity of an expression (Glucksberg & Keysar, 1993). More elaborated accounts of possible metaphor signals were published by Goatly (1997) and Wallington et al. (2003). Goatly (1997) proposed 20 categories of metaphor signals. For instance, words such as 'metaphor' and 'metaphorically' were classified as explicit metaphor signals; 'just', 'really' and 'literally' were labelled intensifiers, as they intensify the metaphor's effect; and 'sort of', as well as 'kind of' were described as superordinate terms because they refer to a metaphor as a superordinate semantic category. Wallington et al. (2003) published a less extensive, but a corpus-based categorization consisting of 12 types of metaphor signals, distinguished according to semantic criteria. The following citations show how metaphor signals can be used in written discourse:

> Laser pulses can form **a kind of** ruler of light, which scientists use to measure the frequencies of other lasers with great precision. (Cundiff, Ye & Hall, 2008: 74)
>
> Pygmy chimps or bonobos are both **literally** and **metaphorically** our kissing cousins. (Kaplan, 2006: 40)
>
> But it is in quantum theory that 1D physics **really** comes to life. (Brooks, 2009: http://www.newscientist.com/article/mg20327231.400-beyond-space-and-time-1d--walk-the-line.html)

Corpus-based studies of metaphor signaling provided evidence of its use in different types of discourse. Metaphor signaling was analyzed in spoken corpora (Cameron & Deignan, 2003; Low, Littlemore & Koester, 2008), in written corpora (Partington, 2006; Skorczynska & Piqué, 2005), as well as in comparative studies of both spoken and written language data (Wallington et al., 2003). On the whole, the findings obtained showed that metaphors are not very frequently signaled in discourse. However, the choice of signaling words as well as their frequency was reported to vary in the different corpora analyzed. For instance, Wallington et al. (2003) found that metaphors were more frequently anticipated in written than in spoken discourse. Skorczynska and Piqué (2005) observed differences in metaphor signaling in business periodical articles and business research papers. Partington (2006) showed that 'sort of' and 'kind of' were more often used as metaphor signals in newspapers than in briefings. Finally, Skorczynska and Ahrens (2010) analyzed business periodical articles, political speeches and popular science articles and

also reported clear differences in the use of 16 metaphor signals. The study mentioned also found out that some of the signals were used in phraseological patterns when the text that followed included a metaphor.

The present analysis continues this line of research by looking into the co-text of 'as' signaling metaphors in a sample of business periodical discourse. The rationale for this study is that in different genres, the phraseology of metaphor signals might vary, as this type of variation could be motivated by the communicative genre-related language functions.

3. Corpus and method

The corpus used in this study consisted of articles from three business periodicals (*Business Week, The Economist, Fortune*) published between 1997 and 2007. The articles in these periodicals are intended for both business practitioners and the general public. They are normally written by journalists specializing in business management and deal with current business events, as well as with general news. Table 1 summarizes the corpus statistics.

Table 1. The statistics of the business periodical article corpus

number of texts	322
mean text length	2,029
tokens in text	653,276
types	27,888

The concordancer of WordSmith Tools 4 (Scott, 2004) was run to obtain all of the concordances of 'as'. The concordances were manually analyzed in order to single out those containing the metaphors signaled by 'as'. The metaphor identification procedure by Pragglejaz Group (2007) was used at this stage of the analysis. Finally, all of the concordances including 'as' and a signaled metaphor were examined to detect patterns of phraseological sets with 'as'.

4. Results

The total of 4,772 occurrences of 'as' was analyzed for its possible metaphor-signaling uses (see Table 2). Of those, 260 were actually used to signal a metaphor. Its frequency per 100,000 words in the corpus of business periodical articles reached the figure of 39.8. The frequency of all of the occurrences of 'as' was 730.47, which is significantly higher in comparison to its metaphor-signaling uses. However, the average frequency of 'as' used as a metaphor signal is notably higher in this copurs than the average frequency of 16 different metaphor signals analyzed in Skorczynska and Ahrens (2010) in a smaller corpus of business periodical articles, where it registered the value of 5.3

per 100,000. From this perspective, the metaphor signal 'as' is more recurrent than other metaphor signals.

Table 2. Frequencies of 'as' in the corpus

'As'	Number of tokens	Average frequency per 100,000
Total uses	4,772	730.47
Metaphor-signaling uses	260	39.8

The analysis of the co-text of 'as' used as a metaphor signal showed that nearly on all of the occasions 'as' combined with a verb. Most of these combinations consisted in one-off uses of a wide range of verbs, such as deduct, speak, treat, cast, cite, invest, emerge, etc. The following citation illustrates such uses: "George Robertson, the new secretary-general of NATO, has **spoken of** America **as** suffering from 'a sort of schizophrenia' ". (The ageing alliance, 1999: 7)

Despite the clear verb variability in the pattern identified, in 47% of the occurrences, 11 verbs were used more recurrently. They are listed in Figure 1.

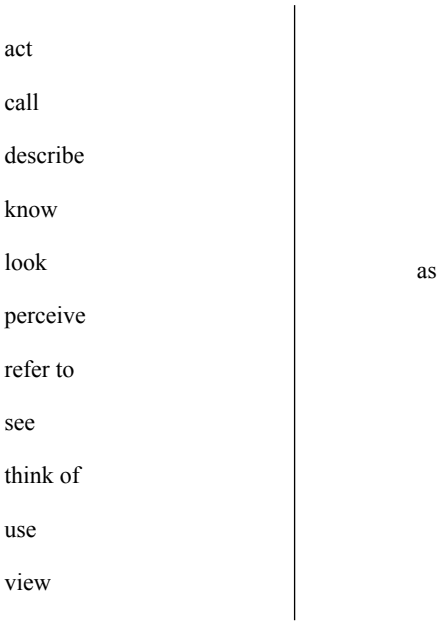


Figure 1. Verbs combining with 'as' in phraseological patterns signaling metaphors

Among the verbs included in Figure 1, four (look, perceive, see, view) are verbs of perception; three (call, describe, refer to) express verbal processes; two (act, use) are action verbs and the other two (know, think of) describe cognitive processes.

Table 3. The occurrences of the combination of a verb and 'as'

Verb + as	Total occurrences	Metaphor-signaling occurrences	Percentage of metaphor-signaling occurrences
view as	8	6	75%
refer to as	9	5	55%
describe as	29	14	48%
look as (if/ though)	29	13	45%
act as	25	11	44%
perceive as	8	3	38%
think of as	17	5	29%
see as	115	32	28%
know as	108	22	20%
use as	38	7	18%
call as	17	3	18%

The total occurrences of each combination of a verb and 'as' in the corpus together with the number of times when each combination signaled a metaphor are shown in Table 3. The last column includes the percentage of the metaphor-signaling occurrences with regard to the total occurrences. The percentages given actually express the probability with which the combinations identified signal a metaphor in this corpus. In this way, 'view as' appears to be the most probable metaphor-signaling phrase (75%), even though it is not at all frequent in the corpus. The second most probable word combination is 'refer to as' with 55% of metaphor-signaling uses, and a clearly low frequency in the corpus. The phrases with the highest frequencies, both in terms of the total and the metaphor-signaling uses, such as 'see as' and 'know as', are not highly probable metaphor-signaling combinations with 28% and 20% respectively. 'Call as', the last item in the table, is not very frequent neither with reference to the total number of its occurrences, nor as a metaphor-signaling phrase.

The quantitative data in Table 3 provide valuable information on the variations in the use of metaphor-signaling phrases including 'as' in the business periodical article and can be readily applied in the metaphor identification in other corpora representing this particular genre. The phraseology of metaphor-signaling expressions might be a characteristic of a genre, even though this observation should be supported by more representative corpus

data. However, as the business periodical article explains and comments on business issues, the type of metaphor-signaling phrases identified in this study seem to participate in the fulfillment of its communicative functions.

> As a veteran software architect, the director of IBM's Institute for Advanced Commerce **views** programming **as** all about suffering—from ever-increasing complexity. (The beast of complexity, 2001: 3)
>
> Over the years, central bankers have popularly been **referred to as** captains, admirals, pilots and lifeboatmen. (Woodall, 1999: 4)
>
> Payne, who owns a benefits-consulting company in New Jersey, **describes** herself **as** a "poster child for the Strategic Coach." (Jervey, 2003: 94)
>
> The real culprit, declared one critic, was "the Wall Street-Treasury Complex," which **used** the IMF **as** a battering ram to open new markets for U.S. financial firms. (Useem, 2001: 80)
>
> So **think of** the European Union not **as** a would-be superstate, but **as** a sort of chassis on to which countries can bolt themselves. (My continent, right or wrong, 1999: 5)

The examples above show how the metaphor-signaling phrases with 'as' render the metaphor's connotative and referential functions more noticeable and explicit in the genre in question.

5. Conclusions

This study offers useful data for the complementary methods in metaphor identification in corpora. Using phraseological sets, such as a verb combined with 'as', as search words in concordancing techniques may simplify the access to the metaphorical material in a particular corpus. Thousands of lines of concordances that normally need to be manually analyzed can be significantly reduced, and the procedure in this way becomes less time-consuming and laborious.

The probability data reported are especially worth considering for corpus searches, and they should be used in combination with the frequency data. It is possible that different genres register varying patterns of metaphor signaling. If such patterns were detected, metaphor identification in corpora could be improved.

References

(1999). My continent, right or wrong. *The Economist, 353*(8142), 3-5.

(1999). The ageing alliance. *The Economist*, *353*(8142), 6-10.

(2001). The beast of complexity. *The Economist, 359(*8217), 3-4.

Black, M. (1993). More about metaphor. In A. Ortony (Ed.), *Metaphor and Thought* (2nd ed.) (19-41). Cambridge: Cambridge University Press.

Brooks, M. (2009). Beyond space and time: 1D – Walk the line. *New Scientist*, *2723* Available at http://www.newscientist.com/article/mg20327231.400-beyond-space-and-time-1d--walk-the-line.html.

Cameron, L. & Deignan, A. (2003) Combining large and small corpora to investigate tuning devices around metaphor in spoken discourse. *Metaphor & Symbol*, *18*(3), 149–160.

Charteris-Black, J. (2004). *Corpus approaches to critical metaphor analysis*. Basingstoke, England: Palgrave-Macmillan.

Cundiff, S., Ye, J. & Hall, J. (2008) Rulers of light. *Scientific American*, *298*(4), 74-81.

Glucksberg, S. & Keysar, B. (1993). How metaphors work. In A. Ortony (Ed.), *Metaphor and Thought* (2nd ed.) (401-424). Cambridge: Cambridge University Press.

Goatly, A. (1997). *The Language of Metaphors*. London and New York: Routledge.

Jervey, G. (2003). Workaholics anonymous. *Fortune, 25*(3), 94.

Kaplan, M. (2006). Make love, not war. *New Scientist*, *192*(2580), 40-44.

Low, G., Littlemore, J. & Koester, A. (2008). Metaphor use in three UK University Lectures. *Applied Linguistics*, *29*(3), 428-455.

Miller, G. A. (1993). Images and models, similes and metaphors. In A. Ortony (Ed.), *Metaphor and Thought* (2nd ed.) (357-400). Cambridge: Cambridge University Press.

Ortony, A. (1993). Metaphor, language and thought. In A. Ortony (Ed.), *Metaphor and Thought* (2nd ed.) (1-16). Cambridge: Cambridge University Press.

Partington, A. (2006). Metaphors, motifs and similes across discourse types: Corpus-Assisted Discourse Studies (CADS) at work. In A. Stefanowitsch & S. Th. Gries (Eds.), *Corpus-Based Approaches to Metaphor and Metonymy* (267-304). Berlin, New York: Mouton de Gruyter.

Philip, G. (2010). Metaphorical keyness in specialised corpora. In M. Bondi & M. Scott (Eds.), *Keyness in Texts* (185-203). Amsterdam: John Benjamins.

Pragglejaz Group (2007). MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, *22*, 1-39.

Scott, M. (2004). *WordSmith Tools 4*. Oxford: Oxford University Press.

Skorczynska, H. & Piqué, J. (2005). A corpus-based description of metaphorical marking patterns in scientific and popular business discourse. *Metaphorik.de*, *9*, Available at http://www.metaphorik.de/09/skorczynskapique.pdf.

Skorczynska, H. & Ahrens, K. (2010). A corpus-based study of metaphor signaling variations in three genres. Paper presented at the *8th RaAM Confere*nce, Amsterdam, 30 June-3 July.

Steen, G. (2007). *Finding metaphor in grammar and usage: A methodological analysis of theory and research.* Amsterdam: John Benjamins.

USEEM, J. (2001). Globalization. *Fortune, 144*(11), 76-84.

WALLINGTON, A.M., BARNDEN, J.A., BARNDEN, M.A., FERGUSON, F.J. & GLASBEY, S.R. (2003). Metaphoricity Signals: A Corpus-Based Investigation. *Cognitive Science Research Papers*. Available at http://www.cs.bham.ac.uk/~jab/ATT-Meta/Papers/CSRP-03-05.pdf

WOODALL, P. (1999). Navigators in troubled waters. *The Economist, 352*(8138), 3-6.

A corpus analysis of rhetorical strategies in the discourse of Chomsky

Keith Stuart

Universitat Politècnica de València

Abstract

This paper explores the rhetorical strategies used by Chomsky in two of his most important books (Syntactic Structures, 1957 & Aspects of The Theory of Syntax, 1965). It continues and widens the research carried out by Hoey (2001) who analysed Chomsky's rhetorical strategies but limited his study to just two passages of Chomsky's writings. One of the claims that we shall be making is that Chomsky is an expert in wrapping propositions in the form of interpersonal metaphors. I have found 264 clause complexes of this type in Aspects of The Theory of Syntax and 248 in Syntactic Structures. Some examples are given of the way Chomsky dissimulates that he is expressing an opinion through the use of the logico-semantic relationship of projection. The paper will not limit itself to these structures but analyzes a range of interpersonal meanings and their lexico-grammatical realisations. It will also suggest some reasons why Chomsky dresses up his texts in a very persuasive form of language. These reasons seem to be principally issues to do with the sociohistorical and academic context in which these two important texts were produced.

Keywords: Corpus analysis of Chomsky, interpersonal meanings, lexico-grammatical realisations

Resúmen

Este artículo explora las estrategias retóricas utilizadas por Chomsky en dos de sus libros más importantes (Syntactic Structures, 1957 y Aspects of The Theory of Syntax, 1965). El estudio sigue y amplía la investigación desarrollada por Hoey (2001), quien analizó las estrategias retóricas de Chomsky, pero limitó su estudio a dos extractos de las obras de Chomsky. Una de las afirmaciones que vamos a realizar es que Chomsky es un experto en el uso de estructuras hipotácticas en forma de metáforas interpersonales. Se han encontrado 264 oraciones complejas de este tipo en Aspects of The Theory of Syntax y 248 en Syntactic Structures. Se presentan algunos ejemplos que muestran cómo Chomsky disimula que está expresando una opinión a través de la utilización de la relación lógico-semántica de la proyección. El artículo no se limita a estas estructuras, sino que analiza una serie de significados interpersonales y sus realizaciones léxico-gramaticales. También se sugieren algunas razones por las que Chomsky disfraza sus textos con un uso muy persuasivo del lenguaje. Estas razones parecen estar principalmente relacionadas con el contexto socio-histórico y académico en el que estas dos importantes obras se produjeron.

Palabras claves: Análisis de Corpus de Chomsky, significados interpersonales, realizaciones léxico-gramaticales

1. Introduction

The objective of this paper is to describe some of the rhetorical strategies used by Chomsky in two books: *Syntactic Structures* (*SS*) and *Aspects of The Theory of Syntax* (*ATS*). I will be presenting various lexico-grammatical features realising interpersonal meanings of evaluation that make his discourse persuasive reading. In particular, I will highlight two rhetorical devices: his use of interpersonal metaphors and what I shall be calling the Chomsky Crescendo effect: lexical, syntactic and structural repetition.

Hoey (2001) carried out a similar study examining two selected passages from Chomsky's writings. His conclusion was that Chomsky uses evaluation both as a running supportive commentary on his own arguments and as a device for cowing opposition. I would agree with the former and disagree with the latter and suggest that Chomsky by embedding evaluations in projected clauses in interpersonal metaphors was being defensive. He had good reasons to be defensive as we will see later.

The paper will be organised into 5 parts: the sociohistorical and academic context within which the two books were written; data organization (how the linguistic data was organised and analysed); specific lexico-gramatical features which were examined; specific discourse features that were examined and, to conclude the paper, I will make some comments about formalisation and corpus linguistics in relation to Chomsky's work.

2. Sociohistorical and academic context

I don't want to exaggerate social factors but Chomsky is a product of his time. When Chomsky wrote *Syntactic Structures*, he was working at MIT's Research Laboratory of Electronics (RLE) as the in-house linguist in Victor Yngve's machine translation project. Both books acknowledge that the research received financial support from the US Army, the US Navy and the US Air Force. Historically, it is a time when the Cold War was at its zenith, mutual accusations of spying from the two superpowers (the US and Soviet Union) and the need for machines to translate the spoken and written texts of enemy governments and enemy agents. It was the dawn of the computer age and computers can only understand simple formal languages. This is the sociohistorical and academic context in which these two books were written. Only a simple, formal language could be processed by a computer. Chomsky had to develop a formal theory of language to be processed by a computer.

In *Syntactic Structures* and *Aspects of The Theory of Syntax*, Chomsky tries to construct a "formalized theory of linguistic structure" and places emphasis on "rigorous formulations". He defines "a grammar of the language L" as "essentially a theory of L", as well as "a device that generates all of the grammatical sequences of L and none of the ungrammatical ones". Talking about the goals of linguistic theory, he draws parallels to theories in physical sciences. He compares a finite corpus of utterances of a particular language to "observations", grammatical rules to "laws" which are stated in terms of "hypothetical constructs" such as phonemes, phrases, etc.

The fact that language is conceived as a structure implies that it can be described in an exact and formal way. The basic operations to generate the language should be as simple

as possible so a machine can be programmed to carry out these operations. Simplicity and universality of the theory is the goal, the objective.

According to Chomsky, the criteria for the "justification of grammars" are "external conditions of adequacy", "condition of generality" and "simplicity". Using the computational technique of corpus frequency, we obtain the following results for words related to *simplicity* and *complexity* in *SS* and *ATS*, which were then filtered carrying out a manual analysis:

Table 1: Frequency of words related to simplicity and complexity

	simplicity	complexity
Syntactic Structures	130	37
Aspects of The Theory of Syntax	114	89

I suspect Chomsky knew language wasn't simple and, by *ATS*, he no longer saw simplicity as a feasible goal. There is no attempt to prove or disprove his arguments. However, I thought it necessary to place Chomsky in a historical context and I strongly believe his work is the result of a particular sociohistorical and academic context.

Our objective is to show the rhetorical strategies in the discourse of Chomsky. In particular, I will highlight two rhetorical devices: his use of interpersonal metaphors and what I shall be calling the Chomsky Crescendo effect: lexical, syntactic and structural repetition. This paper is motivated by trying to come to grips with what is behind the rhetoric. I suspect that his work is the result of a particular sociohistorical and academic context and that the radicalisation of his more recent discourse is a result of being a defected member of the system.

3. Data organization & analysis

In order to analyse the data, the two texts were scanned with the permission of Chomsky and any scanning errors were cleaned up. The basic data is the following:

Table 2: Basic statistics of Syntactic Structures and Aspects of the Theory of Syntax

	SS	*ATS*
tokens (running words) in text	34,583	77,426
types (distinct words)	2957	6113
type/token ratio (TTR)	8.55	7.9
standardised TTR	34.18	38.07
sentences	1,272	2,713
mean sentence length (in words)	**27.19**	**28.54**
word length	4.79	5.12

The only thing of note is that Chomsky's sentences are rather lengthy (very crude estimations have suggested an average English sentence length to be 14.3 words). However, there is great variability which depends on knowledge domain (subject), genre and audience. There is a great deal of hypotaxis in his discourse.

The data was organized and analysed on the basis of 12 categories of lexico-grammatical realisations of interpersonal meanings of evaluation. In the table below, the different lexico-grammatical categories are summarized with some examples from the data:

Table 3: Lexico-grammatical realisations of interpersonal meanings of evaluation

1.	Comparison The **more acceptable** sentences are those that are **more likely** to be produced, **more easily** understood, **less clumsy**, and in some sense **more natural**.(ATS)
2.	Evaluative report structures (with verbs, adjectives or nouns) But in any event, **it is fairly clear that** nothing essentially new is involved here beyond the rules that generate strings and Phrase-markers. (ATS) …makes possible **the widely held belief that** there is little or no a priori structure to the system. (ATS)
3.	Words which are nearly always evaluative: best, optimal, important, interesting etc. Nevertheless, the free word order phenomenon is an **interesting** and **important** one, and much too little attention has been given to it. (ATS)
4.	Modality It **may** be **possible** in principle to develop some semantically… (SS)
5.	Evaluative adverbs (amplifiers, downtoners, emphatics) We have **barely** sketched the justification for … (SS)
6.	Adjectives with Negative prefixes This seems to me not at all an **unnatural** or **intolerable** consequence. (ATS)
7.	Repetition: incremental effect of repetition Concordance examples (67 examples in ATS) 29 Study of **descriptive or explanatory adequacy** may lead to such a conclusion; 38 **a descriptively adequate theory** to the level of **explanatory adequacy** one needs only to define an appropriate evaluation

8. Discourse markers: sentence adjuncts & sentence conjuncts **Presumably**, this is not the optimal analysis, … (ATS) **Unfortunately**, what are intended as empiricist views have generally been formulated in such an indefinite way… (ATS)
9. Short & frequent adverbs: very, far, quite, still, rather etc. It is also **quite clear** that the major goal of grammatical theory is to… (SS) We have not yet considered the following **very crucial** question: …(SS)
10. Words which might seem evaluative (but are not always): first, fundamental, new, different, high, small, strong, much, little, real, singular, alternative, similar, distinct etc. I now proceed to formulate the linguist's goals in quite **different** and independent terms…(SS) We find that the **new** form of grammar is essentially more powerful than… (SS) …there do not appear to be very **strong** reasons for denying… (SS)
11. Anaphoric (& cataphoric) nouns plus evaluation (semi-technical words): method, problem, solution, performance, approach, application, implementation, effect, etc. This is formally the **simplest solution**, and it seems intuitively correct as well. (SS) As a general method, this **approach** is **untenable**. (SS) Such a **position** is particularly **implausible** with regard to language, …(ATS) This **account** is **misleading** in one important respect. (ATS)
12. Evaluative words typical of academic discourse: significant, reliable, constant, variable, suitable, consistent, exact, efficient, applicable, limitation, etc. I think it is fair to say that a **significant** number of the basic criteria for determining constituent structure are actually transformational. (SS) …we can choose the maximally **efficient** possibility… (SS) If there is **consistent** identification, the linguists may apply…(SS)

This kind of analysis is data-driven and so the categories are more disparate than one might find in other studies of lexico-grammatical realizations of evaluation or stance in academic written discourse (see Biber, 2006). A more logically organized analysis might look like the table below, where the focus is on the various realizations of modality in the interpersonal metafunction (Halliday, 1994; Perkins, 1982).

Table 4: An outline of linguistic realizations of modality

1. **Modal auxiliaries**: may, might, shall, should, could, can, must, had to, will, would...

2. **Modal adjuncts**: allegedly, certainly, hopefully, perhaps, possibly, probably, seemingly, surely...

3. **Modal Adjectival structures**: It is sure/likely/certain/permissible/possible/ necessary/obligatory to/that...; I am sure/certain that...

4. **Modal Participial structures**: (i) Epistemic: it is believed/claimed/ considered/doubted/thought that...

5. **Modal Nominals**:

(i) Epistemic: assumption, belief, claim, certainty, consideration, doubt...

6. **Modal Lexical Verbs**: allege, appear, believe, claim, conjecture, feel, hope, require, seem, suggest, suppose, think, wish

Nevertheless, all these categories are covered in table 3 and some other not so traditional categories of evaluation emerge from our data which shows a certain advantage of using a data-driven analysis. We now proceed to investigate some specific structures of Chomsky's discourse.

4. Specific lexico-gramatical features of chomsky's rhetoric: interpersonal grammatical metaphor

In Aspects of The Theory of Syntax, there are 264 clause complexes which involve projection and are of a clause type which Halliday (1994: 354-363) has denominated interpersonal grammatical metaphor. In Syntactic Structures, there are 248 of these clause types. If we consider that there are 1,272 sentences in *SS*, then 19,5% of all sentences are formed using this kind of structure (in the case of *ATS*, we are talking about 9,73% of all 2,713 sentences).

Since interpersonal grammar is organized in two systems, mood and modality, two types of interpersonal grammatical metaphor can be distinguished. In metaphors of modality (which is what concerns us here), the grammatical variation which occurs is based on the logico-semantic relationship of projection. Whereas modal meanings are congruently realized in modal elements in the clause (i.e. modal operators, modal adjuncts), interpersonal metaphors express modal meanings outside the clause, by means of an additional projecting clause. In this way, metaphors of modality are explicit realizations of modal meanings. Speakers can express their opinions in separate clauses in two main ways.

(1) **I think** it's going to rain. (stating explicitly that this is a subjective claim/opinion)

(2) **It is obvious that** … (stating explicitly that this is an objective claim/opinion)

Chomsky's preferred option is the second form

Table 5: Some examples of Interpersonal Metaphor in Syntactic Structures

1. it thus appears that in the simplest system
2. it can easily be seen that the joint description of these two levels will be much simpler
3. it is a reasonable requirement that
4. it is also quite clear that
5. it is evident that
6. it is fair to assume that
7. it is immaterial to our discussion that
8. it is immediately evident that
9. it is important to observe that
10. it is in fact true that
11. it is intuitively obvious that
12. it is necessary that
13. it is not difficult to show that
14. it is not possible to state that
15. it is obvious that
16. it is questionable that
17. it is reasonable to expect that
18. it is strange that
19. it is sufficient that
20. it is undeniable that
21. it is unquestionable that
22. it is unreasonable to demand that
23. it is very questionable that
24. it is, of course, impossible to prove that
25. it may be possible that
26. it seems pointless that

The importance of using this kind of structure is that the author can distance himself from the opinion that he is making.

- **it seems quite clear that** no theory of linguistic structure… **(Who is it clear to?)**
- **it is unquestionable that** opposition to mixing levels, … **(Who thinks it is unquestionable?)**
- **It is quite true that** the higher levels of linguistic description... **(Who is it true to?)**

However, he is not loathed to use the first form either:

Table 6: Concordance examples of 'I think (that)…' (Syntactic Structures)

1. **I think that** we are forced to conclude that grammar is autonomous and independent of meaning…
2. **I think that** much the same thing is true of the relation between syntactic and statistical studies of…
3. However, **I think that** there are other grounds for rejecting the theory…
4. **I think that** such an approach is ill-advised, and that it can only lead to the development of ad hoc…
5. **I think** it can be shown that in each of the cases considered above, and in many other cases, …
6. **I think that** it is very questionable that this goal is attainable in any interesting way…
7. **I think that** such an account can be given…
8. Nevertheless, **I think** it is unquestionable that opposition to mixing levels…
9. **I think that** these considerations give ample justification…
10. **I think that** Fowler's objections to Harris' morphological procedures…

Looking a bit more carefully at these concordance examples, one can observe how Chomsky embeds projecting clauses within projecting clauses:

6. **I think that** (1) **it is very questionable that** (2) this goal is attainable in any interesting way…

This is clearly a criticism but Chomsky expertly embeds his negative evaluation of the goal of constructing a grammar of a language directly from the raw data.

5. Specific discourse features of Chomsky's rhetoric: Chomsky crescendo effect

In this section, we will be demonstrating what I call the Chomsky Crescendo Effect by examining some text from ATS. This rhetorical effect is realized through repetition at

various levels of discourse: lexical, syntactic and structural repetition. Examples of this kind of repetition are highlighted in the text below through underlining and enumeration.

> **1) To the extent that** a linguistic theory succeeds in selecting **2) a descriptively adequate grammar** on the basis of primary linguistic data, we can say that it meets the condition of **2) explanatory adequacy** . That is, **1) to this extent**, it offers **2) an explanation** for the intuition of the native speaker on the basis of an empirical hypothesis concerning the innate predisposition of the child to develop a certain kind of theory to deal with the evidence presented to him. Any such hypothesis can be falsified all too easily, in actual fact, by showing that it fails to **2) provide a descriptively adequate grammar** for primary linguistic data from some other language -evidently the child is not predisposed to learn one language rather than another . It is supported when it does **2) provide an adequate explanation** for some aspect of linguistic structure, an account of the way in which such knowledge might have been obtained. Clearly, it would be utopian to expect to **2) achieve explanatory adequacy** on a large scale in the present state of linguistics. Nevertheless, considerations of **2) explanatory adequacy** are often critical for advancing linguistic theory. Gross coverage of a large mass of data can often be attained by conflicting theories; for precisely this reason it is not, in itself, an achievement of any particular theoretical interest or importance. As in any other field, **4) the important problem** in linguistics is to discover a complex of data that differentiates between conflicting conceptions of linguistic structure in that one of these conflicting theories can describe these data only by ad hoc means whereas the other can explain it on the basis of some empirical assumption about the form of language. Such small-scale studies of **2) explanatory adequacy** have, in fact, provided most of the evidence that has any serious bearing on the nature of linguistic structure. Thus whether we are comparing radically different theories of grammar or trying to determine the correctness of some particular aspect of one such theory, it is questions of **2) explanatory adequacy** that must, quite often, bear the burden of **3) justification**. This remark is in no way inconsistent with the fact that **2) explanatory adequacy** on a large scale is out of reach, for the present. It simply brings out the highly tentative character of any attempt to **3) justify** an empirical claim about linguistic structure. To summarize briefly, there are two respects in which one can speak of **3) justifying a generative grammar** . On one level (that of **2) descriptive adequacy**), **3) the grammar is justified 1) to the extent that** it correctly describes its object , namely the linguistic intuition-the tacit competence-of the native speaker. In this sense, **3) the grammar is justified on external grounds** , **on grounds** of correspondence to linguistic fact . On a much deeper and hence much more rarely attainable level (that of **2) explanatory adequacy**) , **3) a grammar is justified 1) to the extent that** it is a principled **2) descriptively adequate system** , in that the linguistic theory with which it is associated selects this grammar over others, given primary linguistic data with which all are compatible . In this sense, **3) the grammar is justified on internal grounds** , **on grounds** of its relation to a linguistic theory that constitutes **2) an explanatory hypothesis** about the form of language as such . **4) The problem of 3) internal justification of 2) explanatory adequacy** is essentially **4) the problem of** constructing a theory of language acquisition, an account of the specific innate abilities that make this achievement possible. (*ATS*, p.25)

This discourse example from ATS clearly shows how Chomsky utilizes repetition to enhance his argument and persuade the reader. It is almost as if the writer is talking louder and louder, going over the same ground again and again, to convince himself and the reader. Here is another example of this kind of discourse where I focus more on structural repetition.

> Thus **it seems absurd to suppose that** then by applying transformational rules to see if it gives, finally, a well-formed sentence. But **this absurdity is simply a corollary to the deeper absurdity** of regarding the system of generative rules as a point-by-point model for the actual construction of a sentence by a speaker. Consider the simpler case of a phrase structure grammar with no transformations (for example, the grammar of a programming language, or elementary

> arithmetic, or some small part of English that might be described in these terms) . **It would clearly be absurd to suppose that** the " speaker " of such a language , in formulating an " utterance , " first selects the major categories , then the categories into which these are analyzed , and so forth , finally , at the end of the process , selecting the words or symbols that he is going to use (deciding what he is going to talk about). (*ATS*, p.139)

Recently, there have been other criticisms of Chomsky's discourse (Pullum, 2011). According to Pullum (2011), the approach advocated by *SS* springs directly out of the work of the mathematical logician Emil Post on formalizing proof. In general, Pullum's paper is highly critical of the mathematical and logical foundations of *SS*, and the coherence of its formalism. *SS* is now more than 50 years old and has had an enormous catalytic effect on linguistics so I personally find these *a posteriori* criticisms a bit too facile.

6. Conclusion

Francis *et al.* (1996) describe various patterns in the verb group, based on data from the Cobuild corpus. I have counted the verb patterns informally and have found at least 87 verb patterns in Francis's analysis. The question is: now we have the verb patterns and so what? Long lists of patterns are descriptively rich but we are often looking for simple rules. Computers understand simple rules. So, I have certain sympathy with Chomsky's attempt to simplify and universalize. This is only posible through a formal theory of language which, as yet, corpus linguistics has failed to deliver although one may argue that once we have done enough descriptive work and finished the counting we will discover what really counts. However, what it probably boils down to is that the theory was never going to be simple and, when we are dealing with language, we are in the terrain of a complex system. Therefore, Chomsky found himself trying to simplify a system that was never going to be simple and needed to resort to brilliant rhetoric to persuade himself and his readers.

References

Chomsky, N. (1957). *Syntactic Structures*. London: Mouton.

Chomsky, N. (1965). *Aspects of The Theory of Syntax*. Cambridge: M.I.T. Press.

Francis, G., Hunston, S., & Manning, E. (Eds.) (1996). *Grammar Patterns 1: Verbs*. Collins Cobuild.

Halliday, M.A.K. (1994). *An Introduction to Functional Grammar*. 2nd ed. London: Arnold.

Hoey, M. (2001). Persuasive Rhetoric in Linguistics: A Stylistic Study of some features of the language of Noam Chomsky. In Susan Hunston and Geoff Thompson eds. Evaluation in text: Authorial stance and the construction of discourses. New York: Oxford University Press.

Pullum, Geoffrey K. (forthcoming, 2011). On the mathematics of Syntactic Structures. To appear in Journal of Logic, Language and Information.

El País news reports on childhood obesity: a twelve-month corpus study[33]

Debra Westall

Universidad Politécnica de Valencia

Abstract

This research aims to explore Spanish newspaper reporting about overweight and obesity, especially that involving children and adolescents. A specific corpus was complied with 231 news items, all published by the three leading national dailies between 01/01/2008 and 31/12/2008 and extracted from the archives in the online editions of ABC *(n=88; 38.1%),* El Mundo *(n=78; 33.8%) and* El País *(n=65; 28.1%). For the present study, the 65-item* El País *sample was analyzed to determine the thematic coverage. The technical characteristics of this sample are specified, along with the key headline words, and the results obtained from the content analysis then are reported.*

Keywords: corpus design, newspaper article, content analysis, pediatric health, childhood overweight and obesity

Resumen

El objetivo de esta investigación es el de caracterizar las noticias periodísticas en España relativas al sobrepeso y la obesidad, especialmente las que centran en los niños y los adolescentes. Un corpus específico se constituyó con 231 noticias publicadas en los tres diarios nacionales de más tirada entre el 01/01/2008 y el 31/12/2008 y extraídas de las hemerotecas digitales de ABC *(n=88; 38,1%),* El Mundo *(n=78; 33,8%) y* El País *(n=65; 28,1%). Para este estudio, se analizaron las 65 noticias publicadas en* El País *con el fin de determinar los temas predominantes de la cobertura. Se describen los datos técnicos de la muestra, las palabras clave de los titulares y los resultados obtenidos del análisis del contenido realizado.*

Palabras clave: diseño de corpora, noticia periodística, análisis de contenido, salud pediátrica, sobrepeso y obesidad infantil

33 This research was supported by the Spanish Ministry of Education and Science through the R&D project "Retórica y cultura en la información periodística sobre la salud" (HUM2007-65132FILO) and FEDER funds.

1. Introduction

Given the global obesity epidemic and the media's coverage of this phenomenon over the past decade, researchers like Lawrence (2004) have begun to examine news reports on obesity through content and discourse analyses using both quantitative and qualitative indicators. Similar studies have recently been conducted in Germany (Hilbert & Ried, 2009), Norway (Malterud & Ulriksen, 2009) and Sweden (Sandberg, 2007). Yet to my knowledge, there is no comparable study available for Spain. Therefore, this research will examine Spanish written press coverage of weight-related health issues, in other words, obesity and overweight (the latter being used herein to indicate pre-obesity and obesity) and especially in regard to children (Westall, 2011).

There are many reasons for targeting childhood overweight, as opposed to other pediatric health concerns or obesity news in general. First of all, some 43 million children under five are currently overweight; most will be overweight as adults; many will be diagnosed with diabetes or cardiovascular disease, and some will die prematurely (World Health Organization, 2010: 6ss). Indeed, this was the case of a five-year old Spanish child from Murcia, whose obesity-related death was announced in the national press headlines in April of 2008, which was the first time such news was published and thus inspired the present research. Further, in 2000, approximately a quarter of the Spanish population aged 2-24 was either overweight (12.4%) or obese (13.9%) (Serra, Ribas, Aranceta, Pérez, Saavedra & Peña, 2003); by 2006, an estimated 30% of all children aged 2-9 were overweight (Branca, Nikogosian & Lobstein, 2007: 2-3), and today Spain has one of the highest prevalence rates for overweight among European youngsters (11-15 years old) (Haug, Rasmussen, Samdal, Iannotti, Kelly, Borraccino, Vereecken, Melkevik, Lazzeri, Giacchi, Ercan, Pernille, Ravens, Currie, Morgan, Ahluwalia & HBSC Obesity Writing Group, 2009).

Reports such as these are clearly alarming for global public health authorities and, when used to frame media information, are known to influence the public's understanding of the disease (Puhl & Heuer, 2009: 950ss). Boyce (2007: 203) argued that more agenda-based research is needed to determine the "role […] newspapers play in providing information on encouraging weight loss and exercise" as well as to characterize the frames used to present scientific obesity research in the printed press. Similarly, Hilbert and Ried (2009) concluded their study of obesity news in Germany as follows:

> Focusing on a comprehensive analysis of print media, aimed at providing information, is novel and needed given the potential influence of media reports on weight-related stigmatization. [...] For higher-quality national newspapers as well, it appears important to question which and how obesity-related information is presented. (Hilbert & Ried, 2009: 50)

Thus, given the current prevalence of childhood overweight in Spain together with the health consequences for those affected, it is my interest to explore how this disease is represented in the Spanish national press and to gather data on news contents, framing and contemporary obesogenic discourse in this country. In this particular study, I first describe how a specific 231-item corpus of childhood obesity news was created and

then offer details regarding the 65-item sample from *El País*, the leading Spanish daily, namely, the predominant features and most newsworthy themes covered over the twelve-month study period (2008).

2. Corpus design and methodology

This research is part of a larger study into contemporary Spanish health discourse and is based on a specific corpus of national news items published in Spain between 01/01/2008 and 31/12/2008. First, various combinations and synonyms of the key search term 'childhood obesity' were identified in the research literature in Spanish (e.g. Serra *et al*., 2003): *obesidad/sobrepeso infantil, obesidad/sobrepeso juvenil, obeso(s)/a(s), niño(s)/niña(s), adolescente(s)*. These expressions were then used to extract all pertinent news items from the online archives of *ABC, El Mundo* and *El País*. Not only are these the three leading national newspapers in Spain, but their coverage of health and medical issues is also systematically examined for the annual *Informe Quiral*; moreover, of the three dailies, *El País* had the largest circulation for period 07/2007-06/2008, averaging 579,249 copies per day and 444,290 subscriptions (OJD data, cited by Observatori de la Comunicació Científica, 2009: Metodología, p. 1).

Each item was later analyzed manually to verify its relevance, and only those results containing at least one clear reference to childhood/adolescent overweight/obesity were included in the final corpus. Thus, it was possible to examine both those news items in which overweight was the primary focus (e.g. prevalence studies), as well as those in which the key search terms were either mentioned in passing (e.g. a critic's comments on musicians' physiques) or discussed in relation to other topics (e.g. asthma, videogames). Finally, following the methodologies described in the aforementioned studies of obesity discourse, the analytical parameters were established so as to identify the technical characteristics of the news items (publication date, word count, genre and authorship), to determine the key words featured in the headlines (obesogenic and nutritional terminology) and to typify the contents of the coverage (obesity prevalence, causes, complications or prevention).

3. Results and discussion

The final 231-item corpus was created with news about childhood overweight and obesity published in *ABC* (n=88; 38.1%), *El Mundo* (n=78; 33.8%), and *El País* (n=65; 28.1%). The corpus contains approximately 135,932 words in all, with an average of 588 words per text. The distribution of the news items among the three sources and throughout the year is illustrated in Appendix 1, which also indicates the percentages corresponding to each source per month and to each month for the year.

In regard to the sample's characteristics, the findings indicated that *El País* did not tend to repeat the same news items (as did *ABC*) nor were there months with only one or no articles related to childhood overweight being published (as was the case of *El Mundo*). Second, with a total of nine items in July and November, and only two in October, the

average number of news items in *El País* was 5.4 items per month, lower than the 7.3 and 6.5 items per month for *ABC* and *El Mundo,* respectively. However, the *El País* sample contained approximately 43,675 words and averaged 672 words per text, so the articles in *El País* were somewhat longer than the 475 and 645 words per article found in *ABC* and *El Mundo*, respectively. In fact, *El País* published 12 (18.5%) stories with more than 1000 words and only 6 (9.2%) with fewer than 200. Third, the *El País* sample contained 33 (50.8%) informative news items and 23 (35.4%) feature articles, mostly special reports (19; 82.6%), as well as 9 (13.8%) opinion pieces. Finally, the names of staff reporters or journalists appeared on 51 (78.5%) of the *El País* items, and relatively few (14; 21.5%) were attributed to news agencies (EP/Europa Press, Agencias) or ELPAÍS.COM.

The *El País* headline analysis allowed for the identification of specific key obesity (19, 29.2%) and nutritional terminology (23; 35.4%). On the one hand, the key term *obesidad* was featured in twelve headlines, being used once each with the adjectives *infantil* and *mórbida,* and with the reference to *niños* in three. There were two headlines with *sobrepeso* and *obeso,* each with *niños.* Additionally, three headlines mentioned *peso*; one used the term *gordos* and another, *células adiposas,* but these five did not refer to children. On the other hand, numerous terms related to foods, nutrition and dieting appeared in the headlines: *comida (rápida, saludable), cocina (sana), adelgazar, régimen, dietas 'milagro', (sector) alimentario, alimentos (saludables), nutrición (infantil), 'Pezqueñines', manzanas, bollos, tomates, hamburguesa, las grasas 'trans'.*

The thematic analysis of the informative (33; 50.8%) and feature (23; 35.4%) articles in the sample confirmed two main frames, one social (25; 44.6%) and the other scientific (31; 55.4%). The social perspective framed not only the news of the child's death in Murcia, but also the 16 (64%) articles on public and private schemes to control or prevent obesity along with the eight (32%) articles relating obesity, youngsters and the lives of celebrities. The scientific frame was reflected in news about overweight/obesity prevalence (6; 19.4%) as well as the causes (12; 38.7%) and consequences of this disease (13; 41.9%). In the following lines, I shall focus mainly on the scientific framing which seems to be most characteristic of the *El País* news coverage and illustrate the discussion with examples from the sample.

The fact that obesity is a complex multifactorial disease was well documented in the 2008 sample, and numerous causes were identified in *El País* reports. First is the individual's "carga genética heredada", which incluyes the so-called 'thrifty' genes (*genes ahorradores*), the FTO gene (identified in 2007) and the recently-discovered MC4R (in article, also transcribed as MCR4) gene, also called the 'guilty gene' (*el gen culpable*) since it is thought to cause "ciertos tipos de obesidad, especialmente la que afecta a familias enteras […]" ("Un equipo de científicos descubre un segundo gen de la obesidad", 05/05a). Second are environmental factors including television, which was depicted as both a convenient caretaker ("la gran aliada de unos padres ocupados y cansados", in 31/10) and, in the corresponding headline, a common culprit: "La 'tele' para bebés está bajo sospecha" (31/10). Yet there are other factors which are not so well-defined but seem to contribute greatly to the epidemic. For example, reporter J. Marirrodriga pointed out: "en la cultura mediterránea siempre se ha considerado que la

gordura de los niños es síntoma de salud ("Directos del hambre a la obesidad", 11/07a), and this is further compounded by the fact that in Spain "los factores sociales (desde el horario de los padres a las presiones de la industria) son tales que no basta con un enfoque sanitario" ("La obesidad amenaza la esperanza de vida de los niños", 28/02).

El País tended to frame most news on childhood overweight as being related to calorie intake (i.e. "una dieta hiperproteica, hipergrasienta e hipercalórica") and calorie expenditure (i.e. "la poca educación física del colegio, la única actividad para muchos"), as revealed the headline to C. Arribas's feature report: "No es la hamburguesa, niños, es el deporte" (12/02). In one particular article, however, Britain's celebrity chef Jamie Oliver dismissed scientific findings regarding the roles of nutrition and exercise while affirming that "'hay muchos chicos de la City [el centro financiero de Londres] que ganan... [*sic*] bueno, solían ganar, mucho dinero que no son capaces de alimentar a sus hijos, ni con una Visa Oro'" ("El problema es que la gente no sabe cocinar", 06/11b). Most surprising, in any case, was the news about insufficient funding for physical education and athletic competition in Spain, as M.R. Bronx neatly summed up:

> La inversión pública [en la Comunidad Valenciana] para competiciones de alumnos es de 1,7 euros por niño […]. Los niños españoles son los europeos que practican menos ejercicio en horario extraescolar, según datos de la Unión Europea. En apenas dos décadas, el número de chavales españoles obesos se ha triplicado hasta superar el 16%, de acuerdo con el Ministerio de Sanidad. Pero estos datos poco parecen importar a la Generalitat a tenor de la financiación consignada al deporte escolar, a las actividades físicas y competiciones, que se celebran fuera del programa de la asignatura de Educación Física. ("A la cola en deporte escolar", 03/11a)

Regardless of the ultimate cause, serious consequences are undoubtedly associated with what C. Arribas describes as "la corona de grasa que adorna los abdómenes de cada vez más niños y niñas" ("No es la hamburguesa […]", 12/02). Based on a major study published in the *New England Journal of Medicine* in January 2008, reporter E. Rui warned readers: "Niños y niñas con sobrepeso pueden tener con 40 años patologías de 60" ("Secuelas de la obesidad infantil", 08/01) while E. Avellanada offered further details regarding these pathologies:

> Los críos están gordos. […] Se cree que en pocos años tanto la población adulta como infantil experimentará un aumento del síndrome metabólico o también denominado síndrome X, que consiste en un conjunto de dolencias como la obesidad abdominal, hipertensión y aumento de lípidos o azúcar en sangre. ("Adiós al sedentarismo", 24/04)

Moreover, *El País* news drew attention to the psycho-social consequences of childhood overweight, especially in relation to bullying:

> Acosados por estar gordos. El acoso escolar afecta por primera vez, y de forma llamativa, a los alumnos con unos kilos de más. Casi un 30% de los acosadores justifica su actitud frente al compañero porque "está gordo", y la misma percepción tienen las víctimas. Esto puede deberse, según la directora del estudio del Observatorio Estatal de la Convivencia, María José Díaz-Aguado, a un cambio de valores que presta cada vez más atención a la imagen física. ("El racismo cala en las aulas", 18/07)

Curiously, these changes in values and the consequences for our children are best detected within a social framing of the news relating weight, youngsters and celebrities. For example, actress Jennifer Love Hewitt's offers her own advice about hamburgers to body-conscious teenagers:

> 'Quiero decirles a todas que lleven bikini todo el verano y que nunca se sientan mal con su cuerpo porque llegará un momento, pasados los 25, en que comerán una hamburguesa y literalmente verán cómo se convierte en grasa de sus piernas'. ("Jennifer Love Hewitt se arrepiente de sus complejos", 21/08)

Perhaps more striking is one reporter's description of Kate Winslet's physical appearance and the reigning Hollywood standards:

> Once años después de que sus redondeces cautivaran al público en Titanic, Kate Winslet se ha transmutado en una fría y estilizada rubia platino […] la estrella resulta casi irreconocible, cual vampiresa que ha perdido varios kilos en el empeño. […] Se trata en realidad de una mujer muy guapa, pero sus medidas nunca se habían acercado hasta ahora a la peligrosa talla cero que tanto gusta en la meca del celuloide. ("Una fría y estilizada Kate Winslet", 05/11)

In conclusion, the materials gathered for the 2008 Spanish obesity news corpus have allowed for an in-depth analysis of the information offered about this major health concern in the top national newspaper while shedding light on the terminology used, the descriptions given about the causes and consequences of the epidemic and even certain indications of weight-related stigmatism, as reported by other authors (e.g. Hilbert & Ried, 2009; Sandberg, 2007). The 65-item *El País* sample has also provided the opportunity to examine a full year of news coverage to reveal certain thematic tendencies, especially with the scientific frames and, thus, has paved the way for continued research into Spanish press reporting of childhood overweight and obesity and the influence wielded by the mass media.

4. Bibliography

Boyce, T. (2007): The media and obesity. *Obesity Reviews*, *8*, 201-205.

Branca, F., Nikogosian, H., & Lobstein, T. (Eds.). (2007). *The challenge of obesity in the WHO European Region and the strategies for response - Summary.* Copenhagen: World Health Organization.

Haug, E., Rasmussen, M., Samdal, O., Iannotti, R., Kelly, C., Borraccino, A., Vereecken, C., Melkevik, O., Lazzeri, G., Giacchi, M., Ercan, O., Pernille, D., Ravens, U., Currie, C., Morgan, A., Ahluwalia, N., & HBSC Obesity Writing Group. (2009). Overweight in school-aged children and its relationship with demographic and lifestyle factors: results from the WHO-collaborative Health Behaviour in School-aged Children (HBSC) Study. *Int J Public Health, 54,* S167-S179.

Hilbert, A., & Ried, J. (2009). Obesity in Print: An Analysis of Daily Newspapers. *Obesity Facts*, *2*, 46-51. doi: 10.1159/000195697

Lawrence, R.G. (2004). Framing Obesity: The evolution of news discourse on a public health issue. *The Harvard Journal of Press/Politics, 9*(3), 56-75.

Malterud, K., & Ulriksen, K. (2010). Norwegians fear fatness more than anything else' – A qualitative study of normative newspaper messages on obesity and health. *Patient Educ Couns*, *81*(1), 47-52.

Observatori de la Comunicació Científica (UPF). (2009). *Informe Quiral 2008. Medicina, comunicación y sociedad*. Barcelona: Rubes Editorial & Fundació Vila Casas.

Puhl, R.M., & Heuer, C.A. (2009). The stigma of obesity: A review and update. *Obesity*, *17*(5), 941-964.

Sandberg, H. (2007). A matter of looks: the framing of obesity in four Swedish daily newspapers. *Communications*, *32*(4), 447-472.

Serra, L., Ribas, L., Aranceta, J., Pérez, C., Saavedra, P., & Peña, L. (2003). Obesidad infantil y juvenil en España. Resultados del Estudio EnKid (1998-2000). *Med Clin (Barc). 121*(19), 725-732.

Westall, D. (2011). Algo gordo: la obesidad infantil en la prensa española. *Estudios sobre el mensaje periodístico* (Vol. 17, in press).

World Health Organization. (2010). *Population-based prevention strategies for childhood obesity: Report of a WHO forum and technical meeting, Geneva, 15–17 December 2009*. Geneva: WHO.

Appendix 1.

Distribution of 231 news items, among three dailies in 2008, with percentages corresponding to each daily per month and each month per year.

Month/Daily	*ABC* (%, month)	*El Mundo* (%, month)	*El País* (%, month)	Total per month; (%, year)
January	8 (47.0)	2 (11.8)	7 (41.2)	17 (7.3)
February	6 (30.0)	9 (45.0)	5 (25.0)	20 (8.7)
March	11 (44.0)	11 (44.0)	3 (12.0)	25 (10.8)
April	7 (33.3)	8 (38.1)	6 (28.6)	21 (9.1)
May	11 (44.0)	7 (28.0)	7 (28.0)	25 (10.8)
June	8 (72.7)	0 (0)	3 (27.3)	11 (4.8)
July	7 (31.8)	6 (27.3)	9 (40.9)	22 (9.5)
August	5 (45.5)	1 (9.0)	5 (45.5)	11 (4.8)
September	10 (45.5)	7 (31.8)	5 (22.7)	22 (9.5)
October	5 (33.3)	8 (53.3)	2 (13.3)	15 (6.5)
November	5 (19.2)	12 (46.2)	9 (34.6)	26 (11.3)
December	5 (31.2)	7 (43.8)	4 (25.0)	16 (6.9)
Total (% for year)	88 (38.1)	78 (33.8)	65 (28.1)	231 (100%)

Gramática basada en corpus

Language documentation corpora in descriptive linguistics

Peter Bouda

Centro Interdisciplinar de Documentação Linguística e Social

This paper describes the software "Poio Analyzer" developed for descriptive linguists and language typologists. The software allows them to search and analyze interlinear data, i.e. transcriptions with morpho-syntactic annotations and translations. Poio Analyzer was developed specifically for the analysis of data from language documentation projects and is designed to support researchers who want to write descriptive grammars on previously undocumented or poorly-described languages. The corpora from language documentation provide new opportunities to the scientific community, due to their special data layout and extensive, manual annotations. This layout, together with the requirement of the descriptive linguist, also requires new kinds of analysis techniques which are implemented in our software.

Keywords: Language Documentation, Descriptive Linguistics, Language Typology, Software

En este trabajo se describe el software Poio Analyzer, desarrollado para lingüistas descriptivos y tipologistas del lenguaje. El software les permite buscar y analizar datos interlineales, es decir, transcripciones con anotaciones morfosintácticas y traducciones. El Poio Analyzer fue desarrollado específicamente para el análisis de datos de proyectos de documentación lingüística y fue diseñado para apoyar a los investigadores que quieren escribir gramáticas descriptivas sobre lenguas aún indocumentadas o poco descritas. Los corpus de documentación lingüística ofrecen nuevas oportunidades para la comunidad científica debido al diseño especial de sus datos y a las anotaciones por extenso y manuales. Esta disposición (layout), junto con el requisito del lingüista descriptivo, también requiere nuevos tipos de técnicas de análisis que se aplican en nuestro software.

Palabras clave: Documentación Lingüística, Lingüística Descriptiva, Tipología Lingüística, Software

1. Introduction

The role of corpora in the creation of descriptive grammars has gained a lot of attention in the last decades. Still, only a handful of grammars use a corpus to extract linguistic information for analysis. In recent years the usage of software tools in language documentation projects has generated a new source of linguistic data that is being used to compile descriptive grammars for lesser-used and endangered languages. The goal of this paper is to present a software solution to search and analyze annotated corpora that were created in language documentation projects. The software is called Poio Analyzer and is available for download at the website of the *Centro Interdisciplinar de Documentação Linguística e Social* (http://www.cidles.eu/ltll/poio-analyzer, licensed under GNU General Public License v3). The software is designed specifically to be used with DOBES corpora (http://www.mpi.nl/DOBES), but may be extended to other kinds of corpora at a later date.

In section 2 of this paper, I will outline some of the questions a descriptive linguist might pose to a corpus when he is in the process of writing a grammar. These questions result in a typology of searches the linguist needs to apply to a corpus, in order to extract the information about grammatical types and relations on all linguistic levels. This typology was the basis to create a list of requirements for a software tool that is currently used in two language documentation projects, the DOBES projects "Documentation of Hocank" and "Minderico - An endangered language in Portugal".

In section 3 of this paper I present technical solutions and a preliminary version of a database/concordancing software specifically designed to fulfil the functions and principles outlined in section 2. Poio Analyzer supports the Elan and Toolbox file format, two of the main software packages used in DOBES documentation projects. The data files of Elan and Toolbox typically contain transcriptions, morpho-syntactic annotations and translations, which are accessible through a search interface within the software. Search results are displayed with full interlinear data, so that context and annotation data are displayed to the user. Poio Analyzer implements the search strategies that were derived from the requirements outlined in section 3, for example successive searches on previous search results and cross-tier searches with logical operations.

2. Data and queries

2.1. *Data layout*

The basic data in DOBES projects are recordings of audio and video files. These files are transcribed and translated with the help of software tools like Elan,[34] Praat[35] and Toolbox,[36] which add those transcriptions and translations as stand-off annotations to the media files. In many projects additional levels of information are added to this.

34 http://www.lat-mpi.eu/tools/elan/
35 http://www.fon.hum.uva.nl/praat/
36 http://www.sil.org/computing/toolbox/

The levels are normally called “tiers” and contain annotations for word and morpheme segmentation and gloss annotations with morpho-syntactic descriptions. The morpho-syntactic tier contains lexical items and functional items for each morpheme. This data layout is derived from the format of so-called “interlinear glosses”, which is a common way in descriptive linguistics and language typology to cite language data. An example utterance from the Hocank [win] corpus looks like this:

(1) ref ED3024

tx *cii xųųnųjk hižą ʻųų jaagu njj hicec ʻeeja*

mo *cii xųųnų-jk hižą ʻųų jaagu njj hicec*

gl house be.small(OBJ.3SG)-DIM one do/make(SBJ.3SG) what water near

ft he made a small house by the river

dt 25/Oct/2004

The “tx” tier contains the transcribed text and “ft” is a free translation. The morpheme segmentation in “mo” (morphemes) matches the elements in “gl” (gloss), which contains the lexical and grammatical descriptions for each morpheme.

Note here that this is not an official standard: neither the segmentation characters nor the categorical labels in the morpho-syntactic tier are standardized in any way. One method for standardization of interlinear glossed text is to use the “Leipzig Glossing Rules” (Bickel, Comrie, & Haspelmath, 2008), but they have not been universally accepted by the linguistics community. However, most scientists at least use a format that is very similar to the Leipzig Glossing Rules. Within the DOBES corpora, the segmentation is described through separate annotations, so separating characters pose no problem here. Still, the categorical labels may vary from project to project. Figure 1 presents a screenshot of Elan with the example cited above. The annotations on each tier are separated by small vertical bars here.

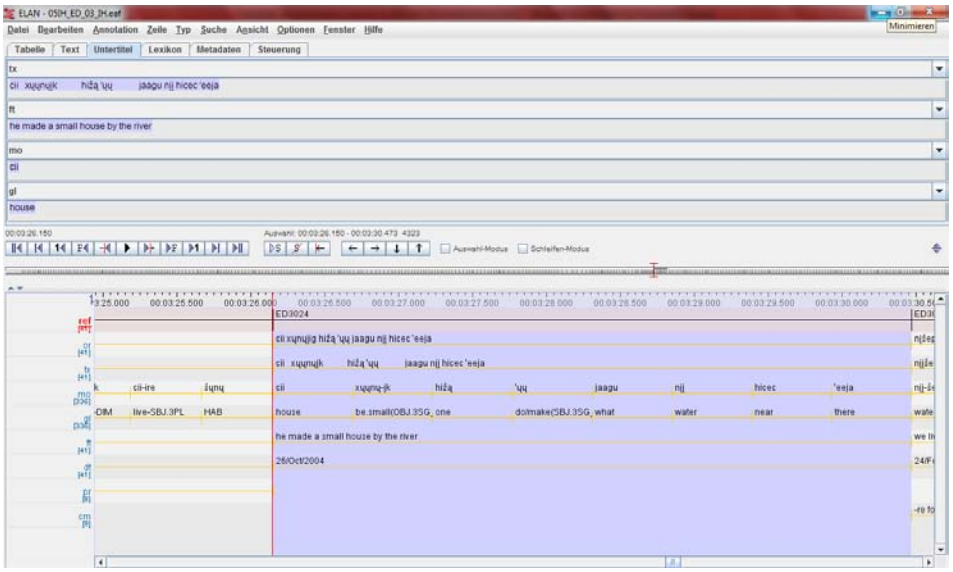


Figure 1. Interlinear example with annotations in Elan.

2.2. *Corpora in descriptive linguistics*

The DOBES corpora and its data layout provide a new kind of data source for the linguistic community. Within the project more than 50 languages were documented, each with transcriptions of a wide variety of text types and with morpho-syntactic annotations in at least a sub-part of the corpus. The goal of the projects, beside the documentation of endangered languages, is the description of the languages. For most of the languages there exists very little publically available information like (small) dictionaries, word lists and articles about individual features. Researchers are now very interested in creating extensive dictionaries and descriptive grammars for the language(s) of their documentation project. Poio Analyzer was developed as a tool to help those researchers to query and analyze their data and to make conclusions about the grammatical relations on each level (phonology, morphology and syntax) of the given language. In such a view, its implementation is heavily based on recent theoretical work about grammar writing (Ameka, Dench, & Evans, 2006) as well as on current research in computer-human-interaction about the usability of software tools (Dipper, Gotze, & Stede, 2004).

The availability of morpho-syntactic annotations and translations in corpora allows for new kinds of searches and analysis. The typology of searches has a primary division of an *onomasiological* and a *semasiological* approach to the corpus data, as maintained in the principles and requirements of grammaticography (Lehmann, & Maslova, 2004). In a semasiological view the researcher poses questions about linguistic structure, for example: What kind of morpho-phonemic processes happen in what contexts? The onomasiological approach starts from the world, and poses questions on how certain ideas are expressed in the given language, for example: How is comparison expressed?

For the first type of questions all tiers can be exploited. To find morpho-phonemic processes in Hocank, the user may search for a string on the utterance tier and its function on the gloss tier. The difference between the utterance and the word tier is giving hints at what kind of process is happening here, as shown in the following example:

(2) ref HOR068

tx *Hegų wogitekji* ***hųųroğoc*** *wa'ųąkšąną, hegų*

mo *hegų woogitek-xjį* ***ho<į-Ø->roğoc*** *wa'ų- 'ąk-šąną hegų*

gl that.way be.angry-INTS **<1E.U-3SG.A>look.at** do/be(SBJ.3SG)-POS.HOR-DECL that.way

tx *'eeja nųųgiwąkjį kirikere haa.*

mo *'eeja nųųgiwąk-jį kiri-kere haa*

gl there run-INTS arrive.back.here-go.back.there make/CAUS\1E.A

ft He was looking at me real mad and I left there running fast.

dt 25/Sep/2006

Note that it is especially important to have the full context of the search result here. The researcher is interested in comparing all elements in an utterance, in all of the tiers.

The second kind of search in the onomasiological approach is applied on the gloss and translation tier, sometimes in combination with other tiers. The availability of translations in a "major" language is what makes this approach feasible, because it is now possible to derive additional information from the translation by applying automated taggers on the translations. We are currently experimenting with morpho-syntactic, part-of-speech and semantic taggers for English (for example CLAWS[37] and USAS[38]; Garside, 1987; Wilson, & Rayson, 1993). Given those additional annotations, the researcher has at least three different options how to query for the idea of "comparison" within the Hocank corpus:

- Solution 1: find "-er"/"-est" (as in "bigg-er")
- Solution 2: search morpho-syntactic tags (comparatives)
- Solution 3: search semantic tags (for the idea of "comparison" in English)

The first solution will also list nominalized words, like "work-er", etc. The second solution may find many utterances with comparisons but may miss periphrastic constructions with "more" and "most" (in this case of course one additional search could find those). A good semantic tagging of the English translation is the best option for onomasiological queries,

37 http://ucrel.lancs.ac.uk/claws/
38 http://ucrel.lancs.ac.uk/usas

this is on-going research in our software development. The usage of existing taggers results in at least a better search result and we are currently investigating what kind of corpus queries the additional information facilitates.

In other projects, when only translations in Spanish or Russian are available, taggers for those languages can be used. Currently we are experimenting with these approaches. There is currently no option in the GUI to try out the automated tagging of the translations. This will be made available in future versions, when we come to conclusions what taggers work best with our data.

This chapter presented two examples for each of the two search types that we derived from the ideas in grammaticography. A more extensive list of search types and a full typology is available in Bouda and Helmbrecht (2011).

3. Development process and technical solutions

Initially, the development of the Poio tools started within the DOBES project "Minderico". It became clear during the project that certain things were either not possible with Elan or Toolbox, the two common tools in language documentation projects (for example complex searches with hits presented in full interlinear context), or very difficult to carry out (morpho-syntactic annotations to transcriptions). At first, we developed a project specific tool to help the scientists within the project to annotate and analyze their data. Later, the development process was more formalized and scientists from other projects were involved to help with the creation of new ideas and features. From a theoretical viewpoint the development process consisted and consists of two parts:

1) The review of existing tools, their advantages and drawbacks (mainly Elan, in our case).

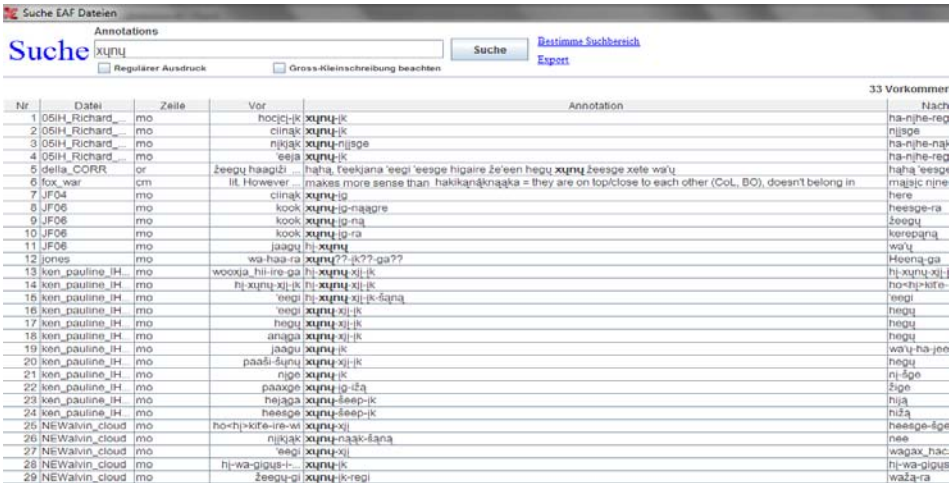
2) The formalization of requirements and their specification within project management framework (with ideas from Extreme Programming, in our case).

We describe both parts in the following two sections. The last section describes the most important requirements that were derived from both approaches.

3.1. Review of Elan

Elan is developed by the Max-Planck-Institute in Nijmegen, Netherlands, and is now the main tool to transcribe and annotate linguistic data in the DOBES language documentation projects. Elan especially focuses on audio and video transcriptions, where data is stored time-aligned with media files. Through hierarchical tiers the user can also add non-time-aligned information like morpho-syntactic annotations, which are normally not directly connected with timing information of the media file but with the transcription or a word segmentation tier in between. Search and analysis capabilities were only added later to the software, which led to the current situation where this functionality is scattered in three different places within the GUI.

First, there is a simple search that just searches all annotations for a given (sub)string (see screenshot in Figure 2). The results are presented in tabular format that lists the surrounding of the annotation and the position within the file (timing and tier name). The user can jump directly to the position in the file by double-clicking on the search result.


Suche EAF Dateien

Annotations

Suche xuɲu Suche Bestimme Suchbereich Export

Regulärer Ausdruck Gross-Kleinschreibung beachten

33 Vorkommer

Nr	Datei	Zeile	Vor	Annotation	Nach
1	05IH_Richard_...	mo	hocịcị-ịk	**xuɲu**-ịk	ha-nịhe-reg
2	05IH_Richard_...	mo	ciinąk	**xuɲu**-ịk	nịisge
3	05IH_Richard_...	mo	nịkịąk	**xuɲu**-nịisge	ha-nịhe-naą
4	05IH_Richard_...	mo	'eeja	**xuɲu**-ịk	ha-nịhe-reg
5	della_CORR	or	žeegụ haagiži ...	haha, 'eekjana 'eegi 'eesge higaire že'een hegụ **xuɲu** žeesge xete wa'ụ	haha 'eesge
6	fox_war	cm	lit. However ...	makes more sense than hakikąnąknąąka = they are on top/close to each other (CoL, BO), doesn't belong in	maịsịc nịne
7	JF04	mo	ciinąk	**xuɲu**-ịg	here
8	JF06	mo	kook	**xuɲu**-ịg-naągre	heesge-ra
9	JF06	mo	kook	**xuɲu**-ịg-na	žeegụ
10	JF06	mo	kook	**xuɲu**-ịg-ra	kerepąnią
11	JF06	mo	jaagụ	hị-**xuɲu**	wa'ụ
12	jones	mo	wa-haa-ra	**xuɲu**??-ịk??-ga??	Heeną-ga
13	ken_pauline_IH...	mo	wooxja_hi-ire-ga	hị-**xuɲu**-xịi-ịk	hị-xuɲu-xịi-ị
14	ken_pauline_IH...	mo	hị-xuɲu-xịi-ịk	hị-**xuɲu**-xịi-ịk	no<hị>kitẹ-
15	ken_pauline_IH...	mo	'eegi	hị-**xuɲu**-xịi-ịk-ɓąną	'eegi
16	ken_pauline_IH...	mo	'eegi	**xuɲu**-xịi-ịk	hegụ
17	ken_pauline_IH...	mo	hegụ	**xuɲu**-xịi-ịk	hegụ
18	ken_pauline_IH...	mo	anąga	**xuɲu**-xịi-ịk	hegụ
19	ken_pauline_IH...	mo	jaagu	**xuɲu**-ịk	wa'ụ-ha-jee
20	ken_pauline_IH...	mo	paaši-šuɲu	**xuɲu**-xịi-ịk	hegụ
21	ken_pauline_IH...	mo	nịge	**xuɲu**-ịk	nị-šge
22	ken_pauline_IH...	mo	paaxge	**xuɲu**-ịg-iža	žige
23	ken_pauline_IH...	mo	hejagą	**xuɲu**-šeep-ịk	hiją
24	ken_pauline_IH...	mo	heesge	**xuɲu**-šeep-ịk	hiža
25	NEWalvin_cloud	mo	ho<hị>kitẹ-ire-wi	**xuɲu**-xịi	heesge-šge
26	NEWalvin_cloud	mo	nịikịąk	**xuɲu**-naąk-ɓąną	nee
27	NEWalvin_cloud	mo	'eegi	**xuɲu**-xịi	wagax_hac
28	NEWalvin_cloud	mo	hị-wa-gigụs-i-...	**xuɲu**-ịk	hị-wa-gigus
29	NEWalvin_cloud	mo	žeegụ-gi	**xuɲu**-ịk-regi	waža-ra

Figure 2. Simple search in Elan.

Figure 3 shows the screenshot of a complex search, which allows the user to search on different tiers and connect the search string through several operations. For example, the user can search for overlaps of search strings on two different tiers. The search result is displayed in a list on the lower part of the window. Only the tiers that were searched are displayed. Again, the user can double click on any hit and Elan will open the file at the position were the search string was found.

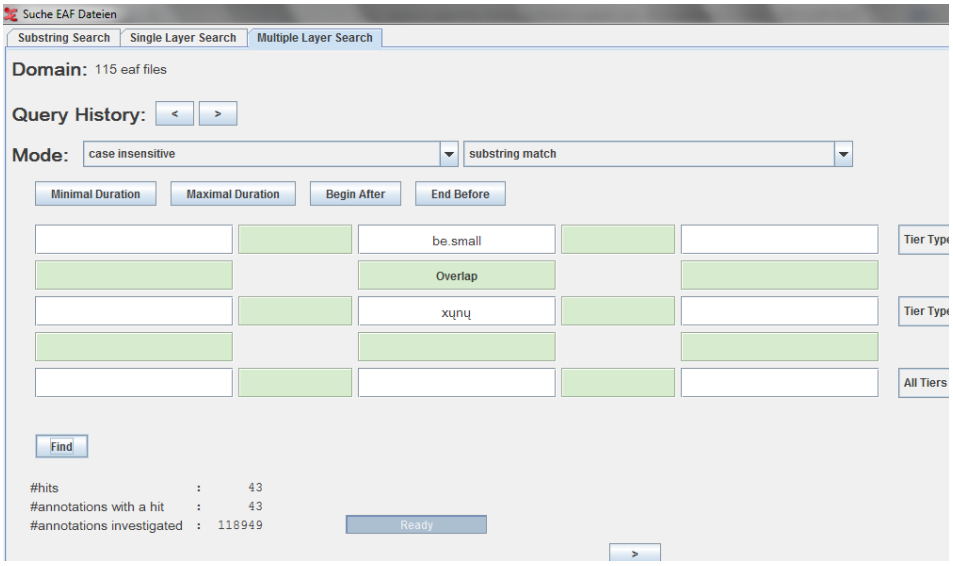


Figure 3. Complex search in Elan.

The third possibility to analyze linguistic data in Elan files is the export facility, which is not shown here. The user can export annotations from a given tier and search the exported files in other software (in a text editor, or more sophisticated corpus analysis tools). In that way it is possible to export lists of words, morphemes, glosses, etc. and, for example, sort them in an Excel table to find common prefixes and identify prefixes and lexical items.

All the described capabilities of Elan have one major drawback: the search results are presented out-of-context. The user has to open the file to see the full interlinear context of the search hit and has to switch between several windows to compare the results. Both search types contain at least one word left and right of the hit, but this is not really important for our users. All scientists we asked agreed that they want to have the result as a full interlinear text with the whole utterance and all its tiers.

Nonetheless, the complex search is an elaborated tool that allows analysis of the data by providing several options to combine search terms on different tiers. A new software tool should at least mimic this behavior and provide search options similar to those in Elan's complex search.

3.2. Project management and development process

The project management and development process were influenced by ideas from the Extreme Programming framework (Wolf, Rook, & Lippert, 2005) from the beginning. The most important principles in our case were the interaction of developers and scientists to derive the specification and frequent software and prototype releases in short development cycles. The main tool to derive the software requirements and

specifications were the so-called user stories, by which users of linguistic software describe how certain tasks and workflows should be carried out in hypothetical software. A user story should be as short and precise as possible. We formulated the user stories in joint work with the developer and the scientists, and broke up longer user stories in shorter ones. A typical user story that was derived from the review of Elan together with a user of Elan was:

"All search results are displayed in full interlinear context, i.e. the hit is displayed within the utterance and all of its tiers."

Whenever there were unclear requirements we implemented a prototype to make things clearer. Often, the presentation of a prototype generated new ideas for user stories that were than formulated and refined.

The next step was to rate the user stories, so that more important features were developed earlier within the implementation phase. For this, we made a list of all the user stories and sent them to several scientists of DOBES projects. Each scientist should rate each story on a scale from one to ten, how important he weights that story. In the end, we calculated the average weight for each story. We started the implementation of features by going through the sorted list of weighted features, implementing one feature after the other. The first version of the software, that is now available, implements only a small subset of all the features in our user stories. But still the software turned out to be useful for scientists in at least two of the DOBES projects, as the most important features are already available.

3.3. Most wanted features

The process described in the last two sections led to list of requirements. A priority was assigned to each requirement, according to the weights we got back from the scientists. The most important requirements for software to analyze data from language documentation were the following (in random order):

The first three of these requirements are implemented in the current version of Poio Analyzer. The rest of the requirements are work in progress, the current focus lies on the "search for word lists" feature. The last two requirements represent user stories which still need more elaboration. Especially the last point, the "search in translations", is currently being implemented in a prototype to be able to discuss this feature with scientists in more depth. In those cases a development process that consists of development cycles with user stories and prototypes helps developers and users to focus on the features that are necessary in everyday usage of the software, without the overhead of implementing functionality that is not used in the end. Extreme Programming provides such a process and has been successfully used to develop our Poio tools. Figure 4 shows the current UI of Poio Analyzer.

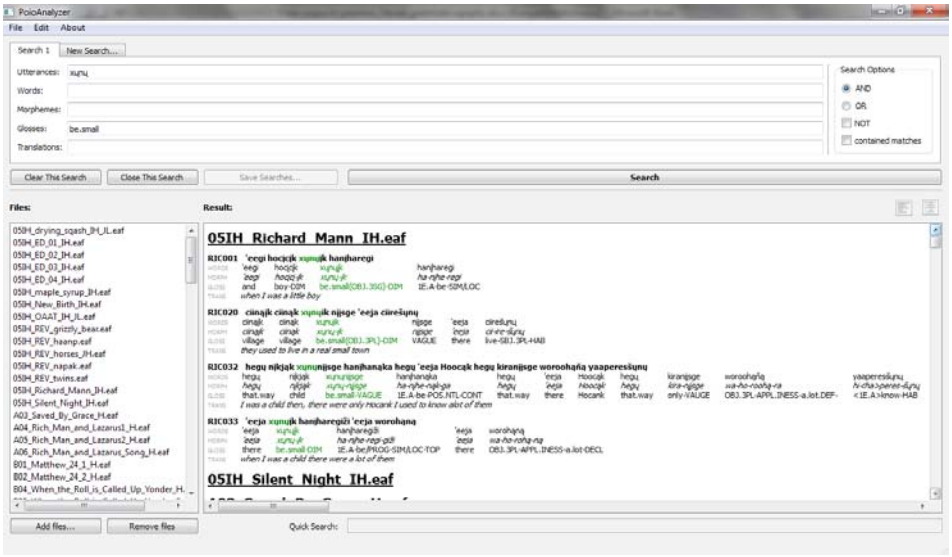


Figure 4. UI of Poio Analyzer, with search terms.

Summary

Morpho-syntactic annotations and translations in corpora from language documentation projects lead to new insights on poorly documented or undocumented languages. To be able to gain profit from these new opportunities it is crucial that general linguists have access to modern technology implemented as software tools. Poio Analyzer will make it easy for any user to access and analyze the data. As shown, current tools used in language documentation focus on editing and annotating the data. After the documentation and archiving phase it became clear the scientists also want to analyze their data. Poio Analyzer has its focus on analysis of interlinear data from the beginning, and was developed with the help and feedback of several scientists working with real data in different projects. This, together with the review of existing software tools, led to a first version of the software that focuses on the basic needs of descriptive linguists that want to query their corpus to write a grammar about the languages they investigate. Poio Analyzer is being actively developed, and requirements derived from user stories and software reviews are implemented step-by-step to improve the software's quality and to add new features that the scientific community requested.

References

Ameka, F. & Dench, A. & Evans, N. (Eds.). (2006). *Catching Language: The Standing Challenge of Grammar Writing.* Berlin: Mouton de Gruyter.

Bickel, B. & Comrie, B. & Haspelmath, M. (2008). *The Leipzig Glossing Rules. Conventions for Interlinear Morpheme by Morpheme Glosses.* Retrieved 2011-05-01 from http://www.eva.mpg.de/lingua/resources/glossing-rules.php

Bouda, P. & Helmbrecht, J. (2011). From corpus to grammar: how DOBES corpora can be exploited for descriptive linguistics. Paper presented at the *Workshop Electronic Grammaticography at the 2nd International Conference on Language Documentation and Conservation,* Hawaii, USA, February 11-13

Dipper, S. & Gotze, M. & Stede, M. (2004). Simple Annotation Tools for Complex Annotation Tasks: an Evaluation. *Proceedings of the LREC Workshop in XML-based Richly Annotated Corpora,* 54-62

Garside, R. (1987). The CLAWS Word-tagging System. In R. Garside, G. Leech & G. Sampson (Eds.), *The Computational Analysis of English: A Corpus-based Approach.* London: Longman.

Lehman, C. & Maslova, E. (2004). Grammaticography. In C. Lehamnn, G. Booij, J. Mugdan & S. Skopeteas (Eds.), *Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung.* 2. Halbband (pp. 1857-1882). Berlin & New York: W. de Gruyter.

Wilson, A. & Rayson, P. (1993). Automatic Content Analysis of Spoken Discourse. In C. Souter & E. Atwell (Eds.), *Corpus Based Computational Linguistics* (pp. 215-226). Amsterdam: Rodopi.

Wolf, H. & Rook, S. & Lippert, M. (2005). eXtreme Programing: Eine Einführung mit Empfehlungen und Erfahrungen aus der Praxis. Heidelberg: Dpunkt Verlag.

Possessor NPs and referential choice in English business prose (a corpus research)

Mariya Khudyakova

Moscow State University

The choice of an appropriate referential expression (definite description, proper name or pronoun) depends on multiple factors. This paper focuses on how the possessor position of a referential expression and its antecedent affect referential choice. other factors, such as syntactical role, form and definiteness of the antecedent, and animacy of the referent are considered. The study is based on a subcorpus of the specially designed RefRhet corpus.

Key words: corpus research, referential choice, anaphora, possessive pronouns

La elección de una expresión referencial conveniente (descripción definitiva, nombre propio o pronombre) depende de varios factores. En esta papel se aclara el problema de cómo la posición del poseedor de una expresión referencial y su antecedente influye en la elección referencial. Se examina también otros factores como el papel sintáctico, forma y definitividad del antecedente, el referente animado o no animado. La investigación se funda en un subcorpus del corpus especial RefRhet.

Palabras claves: investigación de corpus, elección referencial, anáfora, pronombres posesivos

1. Introduction

While producing discourse, a speaker constantly decides what referential expression to use to name a referent. Choosing an appropriate form of a language expression – full descriptive NP, proper name, pronoun, etc.—to refer to an object, person or abstract entity – is called referential choice. Referential choice is a complex cognitive process. For decades linguists have been making different models of referential choice, involving different syntactic, semantic, pragmatic and other factors.

Those factors can be divided into five groups: properties of the antecedent[39], of the referential expression and the referent, the nature of the relation between the antecedent and the anaphor, the genre of the text. Among the properties of the referential expression and its antecedent there are such factors as the syntactical role of the expression (Arnold, 2008; Kibrik, 2003), its semantic role (Rose, 2007), its phrase type (Kibrik, 1997). The properties of the referent, that can affect referential choice, are the animacy of the referent (Greenbacker & McKoy, 2009; Dahl & Fraurud, 1996) and the semantic properties of the referent (e.g. sortal classes in (Strube & Wolters, 2000). The relation between the referential expression and its antecedent is usually expressed in terms of distance, e.g. linear distance in sentences (Greenbacker & McKoy, 2009) or paragraphs (Kibrik, 1997), or rhetorical distance (Kibrik, 1997).

This paper focuses mainly on the role of a specific syntactical position of the referential expression—the possessor position. The two major questions are: 1) How does the referential choice (between full NPs and pronouns)in the possessor position happen? 2) Does the possessor position of the antecedent influence the referential choice of the anaphor?

A practical meaning of these tasks can be illustrated by the example. The language expressions 1, 2, 3 and 4 refer to the same object, that is, are coreferent.

> Tony pulls a tape measure across the front [of what was once a stately Victorian home]1. A deep trench now runs along [its]2 north wall, exposed when [the house]3 lurched two feet off [its]4 foundation during last week's earthquake.

The questions are the following: 1) Is there any difference in the possibility of pronominalization of the referential expressions in the possessor position (2 and 4) and non-possessor position (3)? 2) Does the possessor position of the NP2 affect the referential choice for the NP3?

2. Terminology

There is a certain inconsistence in the terminology dealing with possessor positions. The usual term for the pronouns referring to the possessor (*his*, *my*, etc) is "possessive pronouns". But also the term "possessive" is used for NPs referring to the possessor + object (*his car*, *John's house*, etc) (Willemse, 2009; Storto, 2007; Barker, 2000). In this paper the decision was made to name s-genitive and of-genitive full noun phrases and pronouns, which refer to the possessor, "possessor full NPs" and "possessor pronouns"

39 Antecedent and anaphor are coreferent expressions, the antecedent being the closest one in the previous context to the anaphor.

respectively. The non-possessor pronouns and full NPs are called actant pronouns and actant full NPs.

3. RefRhet corpus

The research is based on the specially annotated RefRhet corpus which consists of 385 Wall Street Journal articles (Kibrik, Dobrov, Zalmanov, Linnik and Loukachevitch, 2010) The RefRhet corpus is based on the English-language corpus RST Discourse Treebank, created under the direction of Daniel Marcu (http://www.isi.edu/~marcu/discourse/Corpora.html), see (Carlson, Marcu, Okurowski and 2003). The corpus contains 176 383 words.

Referential annotation was added to RST Discourse Treebank, and as a result the RefRhet corpus emerged. Referential annotation was performed with the help of a so-called annotation scheme, see (Krasavina & Chiarcos, 2007). The annotation scheme employed contains a set of annotated parameters, or factors.

An element that undergoes annotation, called markable, is a text constituent that can serve as a referential expression. Coreference relations are posited between markables. In addition, each markable contains a number of annotated features (grammatical role, animacy, etc.) that can affect referential choice.

Since all of the annotations are performed manually, a certain number of mistakes is inevitable. In order to exclude such mistakes the decision has been made to annotate each text twice and then compare these annotations automatically. Such comparison results in a list of markables that either appear only in one of the annotations, or have different feature values in the two annotations. Subsequently, annotators from a different group choose the correct analysis out of the two available.

The present-day stage of the RefRhet corpus is as follows: 157 texts are annotated twice, 193 texts are annotated once, and 25 texts are not yet annotated.

For the research a subcorpus of 31 text was chosen. These are the texts that had been annotated twice, and also the procedure of the comparison and correction of the annotations was performed. The subcorpus contains 3453 markables. Since the current annotation scheme suggests the annotation of possessor pronouns, but not of possessor full NPs, the cases of s-genitive and of-genitive were annotated in the subcorpus. In order to exclude the cases of the reflexive possessor pronouns, that are rather syntactical than the result of the referential choice (Bach & Partee, 1980), the possessor pronouns, whose antecedents were in the same clause, were not taken into consideration.

The correlation between different factors and the referential choice was elicited with the help of log-linear analysis, which is used for establishing the correlation of two or more factors.

4. Results

There are 3092 definite NPs in the subcorpus of RefRhet, 85% of which are full NPs, and only 15% are pronouns. NPs in the possessor position present 19% of the chosen

markables. The distribution of the types of referential expressions in possessor and actant position are in Table 1.

Table 1. The distribution of full NPs and pronouns in the possessor and actant position in the subcorpus.

	Possessor position		Actant position		total	
Full NPs	259	8%	2253	73%	2512	81%
pronouns	213	7%	367	12%	580	19%
total	472	15%	2620	84%	3092	100%

As can be seen from Table 1, possessors are more likely to be pronominalized than actants.

Also there is a strong correlation between the animacy of the referent, referential choice and the possessorness of the referential expression. Animate possessors are pronominalized in 80% cases while inanimate possessors and actant NPs are more likely to be full NPs. The most predictable is the referential choice for inanimate referents in actant positions.

The most interesting cases are when possessors and actants demonstrate contrary tendencies to be pronominalized depending on some properties of the antecedents, for example, the definiteness of the antecedent. As can be seen in Table 2, possessors are more likely to be expressed as pronouns after indefinite NPs, and after definite antecedent the numbers of possessor pronouns and full NPs are almost equal, while pronominalized actants have a contrary tendency: they are more likely to be full after definite antecedents.

Table 2. the distribution of full NPs and pronouns in the possessor and actant position in the subcorpus and the definiteness of their antecedents.

antecedent	anaphor											
	possessor						actant					
	Full NPs		Pronouns		**total**		Full NPs		Pronouns		**total**	
Definite NPs	174	47%	195	53%	**369**	**100%**	912	73%	344	27%	**1256**	**100%**
Indefinite NPs	5	29%	12	71%	**17**	**100%**	73	60%	48	40%	**121**	**100%**
total	**386**						**1377**					

There is also a correlation between the form of the antecedent and referential choice in possessor / actant position. Possessors after full NPs are pronominalized less often than after pronouns. The pronoun form of the antecedent has a strong effect on the form of the possessor anaphor—such anaphors are pronouns in ¾ of the cases. Actants have a contrary tendency: they are more likely to be full NPs, than pronouns, especially after full NP antecedents.

The correlation between the syntactical role of the antecedent and the form and possessorness of the anaphor is also statistically significant. Actant NPs demonstrate a tendency to be full NPs after all types of antecedents, while possessor NPs are pronouns after 74% of subject antecedents, 52% of direct object antecedents and 40% of other antecedents.

There is no significant correlation between the possessorness of the antecedent and referential choice.

The research has shown that such factor as possessor / actant position of the antecedent and the anaphor affect referential choice. This feature will be added to the annotation scheme of RefRhet.

References

Arnold, J. (2008). Reference Production: Production-internal and Addressee-oriented Processes. Language and Cognitive Processes. Retrieved from http://www.unc.edu/~jarnold/pages/publications.html

Bach, E. & Partee, B. (1980). Anaphora and semantic structure. In B. Partee (ed.), *Compositionality in Formal Semantics - Selected Papers by Barbara H. Partee*. Blackwell.

Barker, Ch. 2000. Definite Possessives and Discourse Novelty. In *Theoretical Linguistics Volume 26 (3)*. De Gruyter

Carlson, L., Marcu, D. and Okurowski, M. (2003). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In J. van Kuppevelt and R. Smith (eds.) *Current directions in discourse and dialogue*. Dordrecht: Kluwer, 2003. Pp. 85–112.

Chiarcos Ch. & Krasavina O. (2005). Annotation Guidelines. PoCoS — Potsdam Conference Scheme. Draft.

Dahl, Ö. & Fraurud, K. (1996). Animacy in Grammar and Discourse. In Th. Fretheim & J. Gundel (eds.) *Reference and Referent Accessibility*. Amsterdam/Philadelphia: John Benjamins.

Greenbacker, Ch. & McCoy, K. (2009). Feature Selection for Reference Generation as Informed by Psycholinguistic Research. In *Proceedings of the 2009 Workshop on Production of Referring Expressions (PRE-CogSci 2009)*, Amsterdam.

Kibrik, A. (1997). Modelirovanie mnogofaktornogo protsessa: model referentsialnogo sredstva v russkom yazyke. *Vestnik MGU, 1997.4*, 94-105.

Kibrik, A. (2003) Analiz diskursa v kognitivnoy perspective. PhD thesis.

Kibrik, A., Dobrov, G., Zalmanov, D., Linnik, A. and Loukachevitch, N. (2010). Referencial'nyj vybor kak mnogofaktornyj verojatnostnyj process [Referential choice as a multi-factor probabilistic process]. In Aleksandr E. Kibrik (ed.), *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2010)*. Bekasovo, Moscow region. Moscow: RGGU, 173–181.

Rose, R. (2007). Pronoun Resolution and The Influence of Syntactic and Semantic Information on Discourse Prominence. In Branco, A. (ed.), *Anaphora: Analysis, Algorithms and Applications* (pp. 28-43). Berlin: Springer-Verlag.

STORTO, G. (2007). On the structure of indefinite possessives. In B. Jackson and T. Matthews (eds.), *Proceedings of Semantics and Linguistics Theory X*. Ithaca, NY: CLC. 2007.

STRUBE, M. & WOLTERS, M. (2000). A Probabilistic Genre-Independent Model of Pronominalization. In *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics, April 29-May 04, 2000, Seattle, Washington.*

WILLEMSE, P. (2007). Direct and indirect anaphora and the possessee referent of possessive NPs in English. In A. Branco, T. McEnery, R. Mitkov and F. Silva (eds.) *Proceedings of DAARC 2007 (6th Anaphora and Anaphora Resolution Colloquium)*. Porto: CLUP.

Lexicología y lexicografía basadas en corpus

Corpus paralelos alineados: Segmentación textual con fines lexicográficos

Bernadette Borosi

UNED

RESUMEN

En nuestra comunicación presentamos las ideas básicas de un método recopilatorio de unidades en corpus especializado de la temática de medio ambiente, para fines lexicográficos.

Establecidas a priori, conforme a la Teoría Funcional de Lexicografía, las necesidades de los usuarios potenciales de nuestro diccionario bilingüe español-húngaro proyectado, la aplicación del método terminológico sistemático nos facilitará la identificación de los nudos cognitivos y sus interrelaciones en el discurso especializado, posibilitando asimismo el registro sistemático de las unidades y cadenas textuales en las distintas estructuras lexicográficas. La visualización previa de los datos comparativos permitirá la modelización de la extracción de datos y su presentación en una interfaz interactiva.

Palabras clave: lingüística de corpus, corpus paralelo alineado, lexicografía bilingüe, lexicografía especializada, Teoría Funcional de Lexicografía, terminología, húngaro

ABSTRACT

In this paper, based on a specialised parallel corpus in the area of environment, we present the fundamental ideas of a methodology proposal for the compilation of units with lexicographical aims.

Established a priori, according to the Function Theory of Lexicography, the needs of the potential users of our proyected spanish-hungarian dictionary, by applying the sistematic terminological methodology we identify the cognitive nodes and their interrelations in the specialised discourse, thus enabling the systematic inclusion of the units and textual chains in the different lexicographical structures. The previous visualisation of comparative data provides the framework for the modeling of data extraction and their presentation in an interactive interface.

Keywords: corpus linguistics, aligned parallel corpus, bilingual lexicography, specialised lexicography, Function Theory of Lexicography, terminology, hungarian

1. Introducción: la lexicografía en la sociedad de la información

El avance de las tecnologías de la información y las comunicaciones está transformando, de forma palpable, constante y progresiva, el panorama lexicográfico, tanto en su vertiente teórica como práctica. Al mismo tiempo, los productos lexicográficos de nueva generación accesibles en Internet cada vez concentran mayor interés pedagógico, convirtiéndose en herramientas de información léxica complementarias, en ocasiones sustitutorias, de metodologías didácticas, principalmente en la enseñanza de lenguas extranjeras. Las actuales expectativas y exigencias de los usuarios potenciales de estas nuevas herramientas de consulta bilingües, apremian a los lexicógrafos a sugerir fórmulas magistrales que perfeccionen su contenido, la estructuración de datos o el acceso a la información buscada.

Indudablemente, la calidad del diccionario estará relacionada con el grado de satisfacción del usuario quien, impulsado por unas necesidades derivadas de unas situaciones extralexicográficas concretas (Tarp, 2008: 41), acude al diccionario en busca de respuestas. En este sentido, la proyección de una obra lexicográfica no se puede desvincular de los usuarios potenciales. Más aún, todas las fases de su elaboración, desde la recopilación del léxico hasta la visualización de datos en una interfaz interactiva, deberán supeditarse *a priori* a un enfoque orientado al usuario.

El presente artículo versa sobre las ideas fundamentales de un método recopilatorio de unidades a partir de textos paralelos alineados, basado en la segmentación textual sistemática en función de las necesidades de los usuarios potenciales, para un proyecto de diccionario de aprendizaje bilingüe húngaro-español en línea, de la temática de medio ambiente.

1.1 La Teoría Funcional de la Lexicografía (TFL)

¿Quién se identificará como nuestro *usuario potencial*? ¿Qué *situación* le llevará a consultar nuestro diccionario? ¿Se pueden prever sus *necesidades*?

Tarp estudia las características e interrelaciones de estos tres elementos extralexicográficos que, junto con el elemento de *asistencia*, de naturaleza intralexicográfica, constituyen las cuatro categorías lexicográficas de su teoría lexicográfica orientada hacia el usuario (Tarp, 2008: 41-44). En el marco de dicha teoría, denominada *Teoría Funcional de la Lexicografía* (TFL), Tarp define la función lexicográfica como "la satisfacción de los tipos específicos de necesidad lexicográficamente relevantes que puedan surgir en un tipo específico de usuario potencial en un tipo específico de situación extralingüística (Tarp, 2008: 81)[40].

Partiendo de este núcleo conceptual, Tarp relaciona las características del usuario con factores como su lengua materna (LM), el dominio de lenguas extranjeras (LE), el dominio de disciplinas especializadas o sus habilidades generales de consulta lexicográfica; hace una distinción entre necesidades primarias (relacionadas con la función) y secundarias (relacionadas con el uso); y diferencia entre situaciones comunicativas (relativas a la

40 "A lexicographical function is the satisfaction of the specific types of lexicographically relevant need that may arise in a specific type of potential user in a specific type of extra-lexicographical situation" (Tarp, 2008: 81).

recepción, producción, traducción y revisión de textos en la LM y LE) y situaciones cognitivas (sistemática y puntual). Cualquier dato que se incluya en un diccionario se hará en función de estos factores.

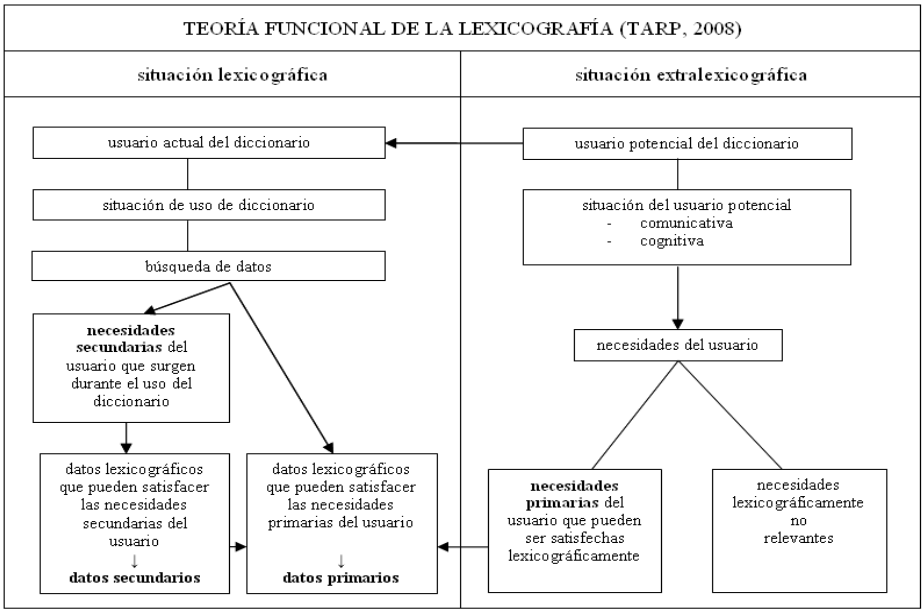


Figura 1. Usuario, situaciones y necesidades (esquema adaptado de Tarp)

1.2 La práctica lexicográfica y la terminología

Cuando estamos proyectando un diccionario especializado, inevitablemente, estamos tratando con unidades léxicas específicas de un ámbito de conocimiento. Estas unidades constituyen un objeto de estudio común en la terminología, siendo ésta un campo de conocimiento con larga trayectoria en el análisis de las unidades terminológicas (UT) a partir de corpus y, por tanto, en consonancia con las recomendaciones de Bergenholtz[41], conviene examinar las posibles aportaciones que la metodología sistemática que se aplica en los trabajos terminológicos pudiera tener en la lexicografía especializada.

En los estudios recientes se considera que

> la terminología es una materia, de carácter interdisciplinar, integrada por fundamentos procedentes de las ciencias del lenguaje, de las ciencias de la cognición y de las ciencias sociales. Estos tres fundamentos inspiran a su vez la poliedricidad de la unidad terminológica, que, en consecuencia, es al mismo tiempo una unidad lingüística, una unidad cognitiva y una unidad sociocultural. […] como consecuencia de esa interdisciplinariedad de triple base, la práctica terminológica es también tridimensional (Cabré, 2005: 70).

41 "[…] we consider serious terminological work an absolute prerequisite for high-quality specialised dictionaries. In this respect, specialised lexicography may benefit from terminology, and it is in this light that we view terminology, or at least areas of terminology, as an integral part of specialised lexicography in a wider sense." (Bergenholtz y Tarp, 1995: 11)

Este método analítico interdisciplinar tiene una doble función: descriptiva y prescriptiva. Mientras que el trabajo prescriptivo está vinculado a la terminografía o "terminología aplicada a la recopilación de términos y a la confección de diccionarios" (Cabré, 2005: 71), que "lleva directamente a la normalización (en el sentido de estandarización) de los términos propios de un determinado dominio especializado" (Cabré, 2005: 29),

> "en un *trabajo descriptivo*, la terminología se entiende como una actividad de recopilación e ilustración de las formas detectadas en el discurso especializado. Es el propio discurso el que proporciona al terminólogo la información cognitiva necesaria sobre el ámbito de conocimiento, y es también el discurso el que le provee de unidades para expresar este conocimiento. El resultado de un trabajo de este tipo es un listado amplio de *unidades de conocimiento* de distintos grados de lexicalización (por tanto, incluyendo unidades terminológicas, fraseológicas y contextos específicos) (Cabré, 2000: 38),

donde "cada unidad terminológica corresponde a un *nudo cognitivo* dentro de un campo de especialidad, y el conjunto de dichos nudos, conectados por relaciones específicas (causa-efecto, todo-parte, contigüidad, anterioridad-posterioridad, etc.), constituye la representación conceptual de dicha especialidad" (Cabré, 2000: 37). Si esto es así, se puede deducir que el trabajo terminológico no se concibe sin el discurso especializado (DE) como fuente de cualquier investigación, dado que "los documentos son el único testimonio del uso de un término en su ámbito de especialidad y una muestra de sus características gramaticales y semánticas" (Cabré, 2000: 34).

Volviendo, en este punto, a nuestro propósito principal de elaborar un diccionario especializado de aprendizaje bilingüe, en la sociedad de la información caracterizada por la perentoria extensión del conocimiento especializado que genera las situaciones comunicativas y cognitivas –referidas por Tarp– que derivan en unas necesidades lexicográficas, observamos que los intereses de la lexicografía especializada convergen con varios componentes de la terminología descriptiva:

1. la lexicografía especializada, de sólidas bases lingüísticas, requiere un método sistemático de recopilación de unidades que podría apoyarse en el trabajo terminológico,
2. la lexicografía especializada tiene que basarse en corpus especializado para dar cuenta del uso real de los términos, y
3. la lexicografía especializada debe prestar atención a las interrelaciones cognitivas de los términos en el DE.

2. Segmentación textual en corpus especializado bilingüe

2.1 Las fuentes documentales y los usuarios potenciales

En nuestro proyecto lexicográfico, pensado para distintos grupos de usuarios, estamos utilizando varios tipos de fuentes para la extracción de información lingüística y cognitiva (libros de texto, artículos especializados, monografías, información accesible en Internet, etc.). Cabe notar que los usuarios potenciales, en función de su "cualificación lingüística y

enciclopédica" (Tarp, 2008: 54-55), nos obligarán a recurrir a una(s) fuente(s) frente a otra(s), es decir, no se podrá aprovechar cualquier tipo de fuente para cualquier tipo de usuario.

Nuestra fuente principal la constituye la legislación europea, por un lado, por incluir generalmente la definición de la terminología y, por otro, por la posibilidad añadida de localizar con bastante facilidad equivalentes fiables. Estos textos se caracterizan por su alto grado de especialización, esto es, alto nivel de abstracción, mostrando alta densidad terminológica. Consecuentemente, la intersección entre el discurso general (DG) y el discurso especializado (DE) se reduce al mínimo.

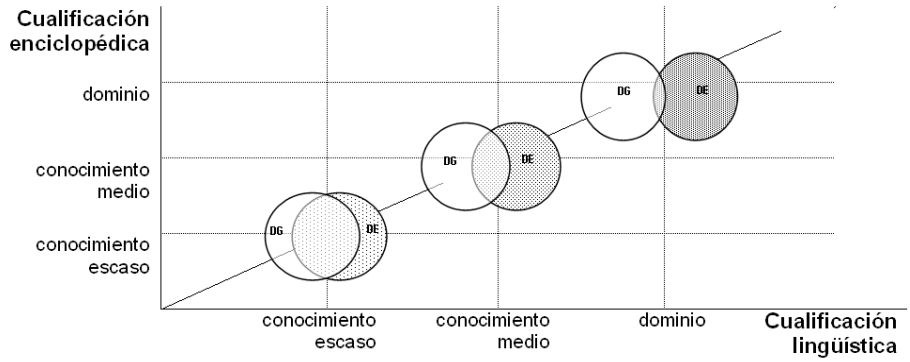


Figura 2. Relación entre la densidad terminológica de la fuente y la cualificación lingüística y enciclopédica del usuario

De este modo, los datos extraídos serán aptos sólo para aquellos usuarios con cierto dominio del ámbito de conocimiento, y que, en nuestro caso, lo constituyen estudiantes universitarios hungarófonos de carreras técnicas, relacionadas principalmente con la ingeniería ambiental, en los últimos años de sus estudios, con la perspectiva de desarrollar también actividades profesionales como mediador comunicativo (traductor, revisor, etc.). En este sentido, tienen un elevado grado de dominio del conocimiento especializado en su lengua materna y están familiarizados con la estructuración conceptual y lingüística del DE.

2.2 Los usuarios potenciales y sus necesidades lexicográficamente relevantes

De acuerdo con la TFL, los usuarios nos plantearán dos tipos de necesidades de información: 1) necesidades primarias que les conducen a consultar una herramienta lexicográfica y que están relacionadas con las funciones lexicográficas y 2) necesidades secundarias relacionadas con el uso del diccionario. Las necesidades primarias se cubrirán con datos primarios y las secundarias con datos secundarios:

NECESIDAD	SOLUCIÓN
necesidad primaria -relativa a la función-	dato primario
necesidad secundaria -relativa al uso-	dato secundario

Figura 3. Tipos de necesidades

En el supuesto concreto de traducción especializada de la lengua materna (LM = húngaro) a la lengua extranjera (LE = español), estamos ante una situación comunicativa relativa, por un lado, a la recepción en la LM y, por otro, a la producción en la LE, asociadas eventualmente con una situación cognitiva puntual, relacionada, en la primera fase, con el conocimiento enciclopédico de la disciplina específica y, en la segunda, con el conocimiento lingüístico productivo de la LE. Así, siguiendo las pautas de Tarp (Tarp, 2008: 146-166), las necesidades previstas para este caso concreto se podrían esquematizar de la siguiente manera:

	RECEPCIÓN en LM (hu) (relativa a la traducción L1→L2)	PRODUCCIÓN en LE (es) (relativa a la traducción L1→L2)
DATOS PRIMARIOS (función)	▪ lema L1 - significado (L1) - [nota pragmática] - equivalente (L2) ▪ expresiones y colocaciones - significado (L1) - [nota pragmática] - equivalente (L2)	▪ lema L2 - ortografía - clase de palabra - género - flexión morfológica - formas compuestas - propiedades sintácticas ▪ colocaciones L2 ▪ expresiones L2 ▪ sinónimos, antónimos, etc. L2
DATOS SECUNDARIOS (uso)	▪ lema L1 - ortografía - variantes ortográficas - formas irregulares como lema - formas compuestas como lema - clase de palabra	▪ lema L2 - ortografía - clase de palabra - género ▪ variantes ortográficas ▪ significado lemas L2 ▪ significado colocaciones L2 ▪ significado expresiones L2

Figura 4. Necesidades relativas a la función y al uso en relación con la traducción de la LM a la LE (esquema adaptado de Tarp)

La segmentación textual de la fuente estará orientada a la extracción de información relativa a estos datos determinados de antemano.

2.3 Delimitación de cadenas textuales en función de las necesidades de los usuarios

Partiendo de la visualización bilingüe en línea de nuestra fuente legislativa[42], elaboramos –con el fin de localizar las correspondencias entre los dos idiomas con mayor facilidad– un corpus paralelo (bitexto) alineado desde la versión española.

42 Directiva 92/43/CEE del Consejo, de 21 de mayo de 1992, relativa a la conservación de los hábitats naturales y de la fauna y flora silvestres.

<L 37> Artículo 1	1. cikk
<L 38> A efectos de la presente Directiva, se entenderá por:	Ezen irányelv alkalmazásában:
<L 39> a) "conservación": un conjunto de medidas necesarias <L 40> para mantener o restablecer los hábitats naturales y <L 41> las poblaciones de especies de fauna y de flora <L 42> silvestres en un estado favorable con arreglo a las <L 43> letras e) e i);	a) védelem: valamennyi, a vadon élő állat- és növényfajok természetes élőhelyei és populációi e) és i) pontban meghatározott kedvező védettségi helyzetének fenntartásához, illetve helyreállításához szükséges intézkedés;
<L 44> b) "hábitats naturales": zonas terrestres o acuáticas <L 45> diferenciadas por sus características geográficas, <L 46> abióticas y bióticas, tanto si son enteramente <L 47> naturales como seminaturales;	b) természetes élőhely: jellegzetes földrajzi, abiotikus és biotikus tényezők alapján elhatárolható, természetes állapotában megőrzött vagy természetszerű szárazföldi illetve vízi terület;
<L 48> c) "tipos de hábitats naturales de interés comunitario": <L 49> los que, en el territorio a que se refiere el artículo 2:	c) közösségi jelentőségű természetes élőhely: olyan élőhely, amelyet a 2. cikkben meghatározott területen belül:
<L 50> i) se encuentran amenazados de desaparición en su <L 51> área de distribución natural;	i. természetes kiterjedésében az eltűnés veszélye fenyeget;
<L 52> o bien	vagy
<L 53> ii) presentan un área de distribución natural reducida <L 54> a causa de su regresión o debido a su área intrínsecamente <L 55> restringida;	ii. visszaszorulása illetve eleve korlátozott területe következtében természetes kiterjedése csekély;

Figura 5. Corpus paralelo alineado (bitexto de referencia) con codificación de líneas

Pero, ¿cómo comenzar la segmentación?

Al tratarse de dos idiomas sumamente diferentes desde el punto de vista morfosintáctico, resultaría bastante infructuoso empezar el análisis textual en el plano lingüístico. ¿Y si, apoyándonos en los parámetros terminológicos, nos situamos en el plano cognitivo y partiendo de los nudos de conocimiento, inequívocamente identificables en los dos idiomas, examinamos las estructuras lingüísticas que expresan y relacionan dichos nudos?

Sin la necesidad de elaborar el mapa conceptual del ámbito concreto (que en el método terminológico sí adquiere gran relevancia), con un simple listado de frecuencia conseguimos reconocer los nudos cognitivos principales cuyos contextos se podrían obtener cómodamente en un listado de concordancias en formato KWIC.

territorio europeo de los Estados miembros, los	Habitats	naturales siguen degradandose y que un numero creciente
habida cuenta de que los	Habitats	y las especies amenazadas forman parte del patrimonio natural
pesan sobre determinados tipos de	Habitats	naturales y sobre determinadas especies
el restablecimiento o el mantenimiento de los	Habitats	naturales y de las especies de interes comunitario
Para el mantenimiento o la supervivencia de un tipo de	Habitat	natural prioritario o de una especie prioritaria
Medidas destinadas a fomentar la conservacion de los	Habitats	naturales prioritarios y de las especies prioritarias de interes comunitario
de la distribucion desigual de tales	Habitats	y especies en la Comunidad y, por otra, de
sistema de vigilancia del estado de conservacion de los	Habitats	naturales y de las especies mencionadas
mantener o restablecer los	Habitats	naturales y las poblaciones de especies de fauna y de flora silvestres
tipos de	Habitats	naturales de interes comunitario
tipos de	Habitats	naturales prioritarios
tipos de	Habitats	naturales amenazados de desaparicion presentes en el territorio contemplado
El "estado de conservacion" de un	Habitat	natural se considerara "favorable" cuando
constituyendo a largo plazo un elemento vital de los	Habitats	naturales a los que pertenezca
exista y probablemente siga existiendo un	Habitat	de extension suficiente para mantener sus poblaciones a largo plazo
mantener o restablecer un tipo de	Habitat	natural de los que se citan en el Anexo I
garantizar la biodiversidad mediante la conservacion de los	Habitats	naturales y de la fauna y flora silvestres
estado de conservacion favorable, de los	Habitats	naturales y de las especies silvestres
unir a garantizar la biodiversidad mediante la conservacion de los	Habitats	naturales y de la fauna y flora silvestres

Figura 6. Listado de concordancias en formato KWIC

A la par de identificar las equivalencias correspondientes, iremos recopilando las UT y analizando todas las estructuras vinculadas con cada nudo cognitivo, conforme a los datos primarios y secundarios que precisamos registrar en nuestro diccionario. Es decir, nos interesa inventariar no sólo los términos específicos del DE con sus características gramaticales, sino también aquellas estructuras semántico-funcionales que reflejan las relaciones lógicas y conceptuales con otros nudos. En este sentido, los términos en cuanto unidades lingüísticas adquieren un componente cognitivo complementario, convirtiéndose desde el punto de vista analítico en unidades lingüístico-cognitivas que constituirán nuestras *unidades lexicográficas* (ULX) que se registrarán en el diccionario.

Así, de acuerdo con los datos obtenidos, en nuestra fuente el término *hábitat* se identifica como nudo cognitivo principal con 78 apariciones, presentando una bifurcación en dos nudos cohiponímicos, *hábitat natural* y *hábitat de una especie*, respectivamente. En torno a estos tres núcleos conceptuales se estructura el contenido del texto, donde los componentes relacionales pueden aparecer de forma explícita o implícita, y a cualquier distancia dentro del entramado conceptual del DE. Por ejemplo, el Anexo I de la Directiva agrupa, de forma explícita, los tipos de *hábitats naturales de interés comunitario*, constituyéndose todos ellos subcategorías conceptuales de *hábitat*, y siendo lingüísticamente cohipónimos de la UT *hábitat natural*, por lo que será recomendable su registro lexicográfico entre los datos primarios como remisiones del lema. Igualmente, hallamos cadenas textuales que se interrelacionan de forma implícita, como, por ejemplo, en el fragmento "los hábitats y las especies amenazadas forman parte del patrimonio natural de la Comunidad" se advierte la relación meronímica entre *hábitats* y *patrimonio natural*, por lo que se aconseja también la inclusión de éste entre los datos primarios como término relacionado. El análisis en fragmentos textuales más extensos nos evidencia la concatenación de UT mediante distintas relaciones conceptuales, como ocurre, por ejemplo, en el siguiente fragmento:

> Se crea una red ecológica europea coherente de zonas especiales de conservación, denominada «Natura 2000». Dicha red, compuesta por los lugares que alberguen tipos de hábitats naturales que figuran en el Anexo I y de hábitats de especies que figuran en el Anexo II, deberá garantizar el mantenimiento o, en su caso, el restablecimiento, en un estado de conservación favorable, de los tipos de hábitats naturales y de los hábitats de las especies de que se trate en su área de distribución natural.

Observamos que los hábitats naturales integran las zonas especiales de conservación que conforman una red ecológica europea denominada «Natura 2000». Estas nuevas relaciones semánticas de meronimia nos conducirán al encadenamiento de remisiones lexicográficas de doble dirección (hábitat natural ↔ zona especial de conservación ↔ red ecológica europea ↔ Natura 2000) que permitirá el acceso a toda la cadena en cualquier punto y en cualquier dirección, mientras que las funciones o finalidad de la red ecológica se incorporarán entre los datos primarios del término relacionado (estado de conservación favorable) como colocaciones (mantener ~, restablecer ~, garantizar el mantenimiento del ~, garantizar el restablecimiento del ~). Por otro lado, el fragmento ilustra el uso de las variantes denominativas *hábitats de especies* y *hábitats de las especies* frente a *hábitat de una especie*, formas que figurarán como datos secundarios, con los contextos de uso correspondientes. En todo caso, las UT se registrarán con uno de los contextos en los que aparecen que servirá también para ilustrar sus características gramaticales.

Asimismo, siempre que sea posible, se aprovechará el propio texto legislativo para recabar la definición de las UT -como dato secundario de las remisiones-, que nos podrán evidenciar también nuevos nexos conceptuales implícitos. Por ejemplo, en "«hábitats naturales»: zonas terrestres o acuáticas diferenciadas por sus características geográficas, abióticas y bióticas, tanto si son enteramente naturales como seminaturales" o en "«hábitat de una especie»: medio definido por factores abióticos y bióticos específicos donde vive la especie en una de las fases de su ciclo biológico", las unidades *zona terrestre*, *zona acuática*, *(factor) abiótico*, *(factor) biótico*, *seminatural* o *ciclo biológico* no se podrán desvincular de las UT a las que caracterizan y, al constituir además vocabulario definidor, su presencia resultará indispensable en el diccionario.

Naturalmente, las equivalencias de las UT en húngaro mantendrán las mismas relaciones semántico-funcionales con los distintos nudos cognitivos que las UT en español. Lo que variará (o podrá variar) es la estructura lingüística mediante la cual se expresará dicha relación conceptual. Por ejemplo:

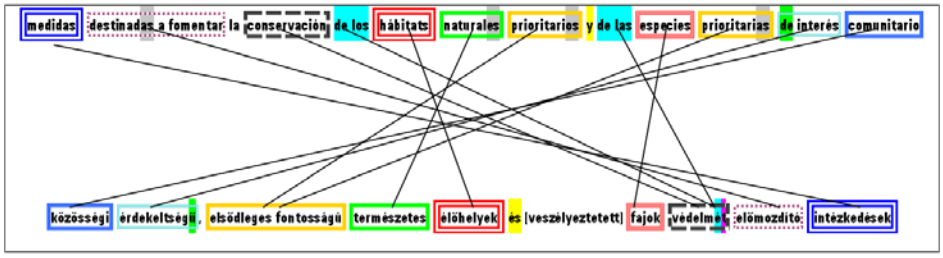


Figura 7. Estructuras lingüísticas comparativas español-húngaro

Las posibles diferencias lingüísticas determinarán la longitud de las cadenas textuales que se requerirá registrar en el diccionario con el objeto de apoyar la producción de textos en la LE. Algunos de los aspectos a tener en cuenta serán la sistemática inversión del orden de palabras, la ausencia de género, la acumulación de afijos, la doble conjugación verbal, o la flexión casual condicionada por regímenes y recciones en húngaro.

3. Visualización de datos

En la base de datos bilingüe que alimentará el diccionario demo, disponemos de varios formularios para grabar los datos, así como de distintas consultas y plantillas de informes para visualizar los datos grabados. Algunas de las plantillas nos permiten listar las unidades en función de sus relaciones conceptuales, y otras muestran los distintos apartados de la microestructura.

área de distribución natural	természetes kiterjedés
[en su] área de distribución natural	természetes kiterjedés[ükön]
estado de conservación	védettségi állapot
estado de conservación favorable	kedvező védettségi állapot
garantizar el estado de conservación favorable	biztosít kedvező védettségi állapotot
garantizar el mantenimiento de el estado de conservación favorable	biztosítja a kedvező védettségi állapot fenntartását
garantizar el restablecimiento de el estado de conservación favorable	biztosítja a kedvező védettségi állapot helyreállítását
mantener el estado de conservación favorable	fenntart védettségi állapotot
restablecer el estado de conservación favorable	helyreállít védettségi állapotot
hábitat[s] de especies	fajok élőhelye[i]
hábitat[s] natural[es]	természetes élőhely[ek]
tipo[s] de hábitats naturales	természetes élőhelytípusok
Natura 2000	Natura 2000
red ecológica	ökológiai hálózat
red ecológica europea	európai ökológiai hálózat
red ecológica europea coherente	egységes európai ökológiai hálózat
red ecológica europea coherente de zonas especiales de conservación (DEF)	különleges természetmegőrzési területek egységes európai ökológiai hálózata (DEF)
zona[s] especial[es] de conservación	különleges természetmegőrzési terület[ek]

Figura 8. Visualización demo de relaciones conceptuales de Natura 2000

4. Conclusiones

La lexicografía electrónica (e-lexicography) está ganando cada vez más terreno a la lexicografía impresa. Las tecnologías de la información y las comunicaciones aportan el medio idóneo para la elaboración de herramientas lexicográficas de calidad, no obstante, las ventajas tecnológicas sólo podrán ser aprovechadas aplicando unos principios teóricos apropiados y una metodología sistemática de trabajo analítico y recopilatorio. Partiendo de las necesidades de los usuarios potenciales estudiadas en profundidad en la TFL, hemos revisado la conveniencia de utilizar en la lexicografía especializada el método terminológico sistemático, basado en corpus especializado. El análisis evidencia que identificando los nudos de conocimiento en el discurso especializado y sus interrelaciones semántico-funcionales, no sólo se obtiene una visión más completa sobre el entramado conceptual del ámbito de conocimiento, sino la segmentación textual conceptual propicia la *recopilación estructurada* de unidades y cadenas textuales semánticamente autónomas (monolexías y polilexías), cuyos nexos determinarán su ubicación en las estructuras lexicográficas. Por otro lado, el método recopilatorio sistemático, basado en la *pertinencia* y *relevancia*, además de ampliar el abanico de unidades registradas y enriquecer el contenido con la necesaria *contextualización*, permite definir con mayor exactitud las rutas de acceso a la información que el usuario podrá necesitar en una situación concreta de consulta, que comportará una extracción más eficiente de datos.

Igualmente, el método tiene especial importancia en la lexicografía bilingüe, pues proporciona información lingüística contrastiva y da cuenta de las divergencias lingüísticas en función de las necesidades definidas. En este sentido, su aportación podría resultar relevante para reforzar el aprendizaje de lenguas extranjeras.

Ciertamente, la lexicografía especializada no se concibe sin corpus especializado que constituye tanto la fuente de contextualización como un medio de análisis polifuncional. Por otra parte, las TIC tienen cada vez mayor implicación en la confección de herramientas lexicográficas, por participar en todas las fases de su elaboración, desde la recopilación de unidades hasta la modelización de estructuras de presentación lexicográfica en una interfaz interactiva, convirtiéndose en parte integrante de la herramienta de consulta proyectada.

Referencias bibliográficas

Bergenholtz, H. y Tarp, S. (Eds.) (1995), *Manual of Specialised Lexicography. The preparation of Specialised Dictionaries.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Cabré, M.T., Gonzalo, C. y García, V. (2000): *Documentación, terminología y traducción*. Madrid, Editorial Síntesis.

Cabré, M. T. (2005): *La terminología: Representación y comunicación*. Barcelona, Universitat Pompeu Fabra, IULA.

Tarp, S., (2008): *Lexicography in the Borderland between Knowledge and Non-Knowledge. General Lexicographical Theory with Particular Focus on Learner's Lexicography*. Tübingen: Max Niemeyer Verlag.

Sense and syntax of *speak* and *talk*

Garikoitz Knörr

Universitat de València

Keith Stuart

Universitat Politècnica de València

Abstract

The paper presents a corpus analysis of the speech act verbs speak and talk. First, we analyze the frequency, both absolute and relative, of these verbs in the corpus consulted (BNC and COCA). We then go on to explain the difficulties in the semantic characterization of these verba dicendi. After some observations about their behavior in certain derivational processes of word formation, we analyze two syntactic environments common to speak and talk: in combination with the prepositions 'to' and 'with', and speak / talk + about. At this point, we highlight the apparent preference of talk to be used with the progressive aspect, which would support the existence of some correlation between sense and syntax.

Keywords: Corpus analysis of 'speak' and 'talk', sense and syntax, word formation

Resúmen

El artículo presenta un análisis de corpus de los verbos speak y talk. En primer lugar, se analiza la frecuencia, tanto absoluta como relativa, de estos verbos en los corpus consultados (BNC y COCA). A continuación, se explican las dificultades en la caracterización semántica de estos verba dicendi. Tras algunas observaciones sobre su comportamiento en determinados procesos derivativos para la formación de nuevas palabras, finalmente se analizan dos contextos sintácticos habituales de speak y talk: en combinación con to y with, y en la construcción speak/talk about. En este punto se hace hincapié en la aparente preferencia de talk por el aspecto progresivo, lo que avalaría la existencia de cierta correlación entre significado y sintaxis.

Palabras claves: Análisis de Corpus de 'speak' y 'talk', significado y sintaxis, procesos derivativos en la formación de palabras

1. Introduction

The objective is to analyse two verbs (*speak* and *talk*) based on the data provided by the BNC[43] and the COCA[44].

Speak and *talk* are two speech act verbs, i.e. verbs of communication, which are typically studied together with some other *verba dicendi,* particularly with *say* and *tell* (cf, e.g., Dirven *et al.* 1982). However, *speak* and *talk* share enough features to be analysed separately. For example, Dirven *et al.* (1982: 10) explain in their in-depth analysis of *speak* that "in contrast with *say* and *tell,* which focus on the speaker and the message […], *speak* —together with e.g. *talk, converse, chat* and *gossip*— puts the speaker and the communicative event itself in a central position". In the same work, they refer to speak and talk as being "sometimes interchangeable" (ibid., p. 11). As far as *talk* is concerned, Dirven *et al.* (ibid., p. 42) explain the difference between the verbs *talk* and *speak* in these terms: "The difference —however small— between *talk* and *speak* is that *talk about* normally denotes a discourse topic, and that *speak about* is more neutral in this respect".

2. Speak and talk in the BNC and the COCA

Assuming, thus, that *speak* and *talk* are in fact very similar verbs, sometimes –if not often— interchangeable, an obvious starting point is to compare their relative frequency in the BNC and the COCA. *Speak* and *talk* are two high-frequency verbs: *talk* ranks 168 and *speak* 336 in the frequency list from the COCA list of all lemmas. Their frequency is more notable in the list of verbs, where they rank 40 and 74 respectively.[45]

Table 1. Relative frequency of speak and talk in the BNC and the COCA

[Speak]	[Talk]
BNC 24.908	29.417 BNC
COCA 123.758	239.978 COCA

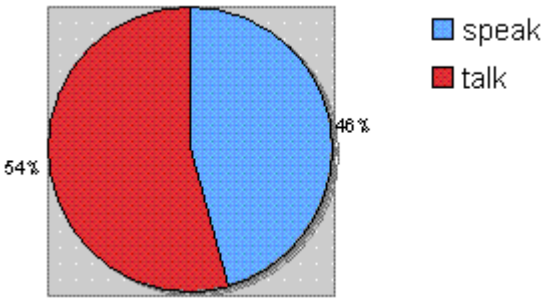


Figure 1. Speak and talk in the BNC

43 Davies, Mark. (2004-) BYU-BNC: The British National Corpus. Available online at http://corpus.byu.edu/bnc.

44 Davies, Mark. (2008-) The Corpus of Contemporary American English (COCA): 425 million words, 1990-present. Available online at http://www.americancorpus.org.

45 Davies, Mark. (2011) Word frequency data from the Corpus of Contemporary American English (COCA). Downloaded from http://www.wordfrequency.info on April 03, 2011.

Talk (as a verb) is a bit more frequent than *speak* in the BNC, whereas in the COCA *talk* is almost twice as frequent as *speak,* which constitutes a relevant difference. It could be interesting to elucidate if it is a merely dialectal difference, or if it might be linked to the fact that the COCA includes more recent data, so a diachronic analysis might reveal that there has been some kind of development in the usage of *talk*, something already suggested by Carl Buck as early as 1915:

> Eng. talk, though the notion of informal, familiar speech is dominant, and even a depreciatory sense evident in certain phrases, is also used without any such feeling, and colloquially it is a growing rival of speak. The child «learns to talk», one may «talk French», and «he talked well» or «what did he talk on?» may refer to the most dignified and formal address (Buck, 1915: 5).

3. Meaning — What dictionaries say

Buck (1915: 2) states that "in most of the modern languages [...] there is *par excellence* one verb of 'speaking' and one of 'saying', as Fr. *parler* and *dire".* In most Romance languages the situation is similar, and in some Germanic languages as well — although it is not the case of German, for example, where we find *sprechen* and *sagen*, but also *reden*, which rivals in some cases with *sprechen* and in other cases with *sagen*.

> Informants being asked about the basic verbs of communication in their language and being forced to limit their choice to two verbs, may easily turn up with such pairs as *say* and *speak*, *zeggen* and *spreken*, *sagen* and *sprechen*, *dire* et *parler*. Still this is not reflected in the frequency of occurrence of these verbs (Lehmann, 1977: 102).

The problem is that in English there are two verbs, two different lemmas used to convey the meaning of Fr. *parler*, Sp. *hablar*, etc., namely *speak* and *talk*. Consequently, this is what we find, for example, in the *Collins Spanish-English, English-Spanish dictionary* when we look up *speak, talk,* and *hablar* respectively*:*

> **speak** 2 vi (a) (gen) hablar
>
> **talk** 2 vi (gen) hablar
>
> **hablar** to speak, talk

If we look at monolingual dictionaries of English, the situation is very similar. For example, the CLDE gives "to talk to someone" as a definition of *speak,* and in the CALDE we find *talk* defined as "to speak to someone". Obviously, this kind of circularity can be confusing to English learners.

> **speak**: to talk to someone *(CLDE)*
>
> **talk**: to speak to someone *(CALDE)*

Something similar happens in other monolingual dictionaries of English:

Table 2. Dictionary definitions of *speak* and *talk*

	speak	*talk*
OED	a. To utter or pronounce words or articulate sounds; to use or exercise the faculty of speech; to express one's thoughts by words.	a. To exercise the faculty of speech; to speak, utter words, say things [...]
LDOCE5	1 to talk to someone about something 2 to use your voice to produce words	1 to say things to someone as part of a conversation 3 to produce words and express thoughts, opinions, ideas etc

In view of this problem, many dictionaries are offering some additional guidelines. For example, *Oxford Study* includes a usage note under the entry *hablar*, which begins by saying: «**To speak** y **to talk** tienen prácticamente el mismo significado, aunque **to speak** es el término más general». So *speak* is described as being "the most general term". That explanation, however, is of little help for ESL students: it is true that it is felt to be the general *parler*-term, but that is not reflected in the actual use.

The instructions and the example sentences are clear and helpful in most cases. However, there seems to be a contradiction between this:

> When you mention what language someone uses, always use **speak**
>
> *She speaks* (=knows how to use) *French and Spanish.*
>
> We spoke in German at first, then English.

and the following subentry found in the entry for *talk* in the same dictionary:

> Talk (in) French/German etc
>
> They started talking in Spanish.

Again, this kind of thing might result in causing confusion in English learners.

4. Morphology

Although this paper focuses on *speak* and *talk* as verbs, it is also interesting to look at their productivity as nominal suffixes. When it comes to derivation, to the creation of new word forms, *talk* and *speak* seem to tend to overlap again, even if they are no longer verbs (cf. Vincent 2001).

If we perform the search **speak* in the BNC, we get about 70 matches, including word forms like *double-speak / doublespeak* (12), *newspeak* (10), *adspeak* (4), *Eurospeak* /

Euro-speak (6); some other forms occur twice in the corpus, such as *crosspeak* (shouldn't it be 'cross-speak'?), *IBMspeak, malespeak, police-speak* and *rockspeak;* and finally we have a long list of nonce words, including *Saddamspeak, soccer-speak, White-House-speak, computerspeak,* etc.

The COCA, on the other hand, yields over 400 matches, with *doublespeak / double-speak* (71), followed by *coachspeak / coach-speak* (44) (of which there is no trace in the BNC), and *newspeak / new-speak* (26). Curiously enough, the next most frequent form is *sisterspeak* (22) (which turns out to be the name of a section in the *Ebony* magazine, where all 22 examples are taken from, so one always has to take into consideration corpus design), followed by *techspeak / tech-speak* (14), *teenspeak / teen-speak* (14), *cyberspeak* (9) and others. Again, most of the results are nonce words including, for example, *yuppiespeak, pentagonspeak* (also *pentagon-speak),* and *petrolspeak.*

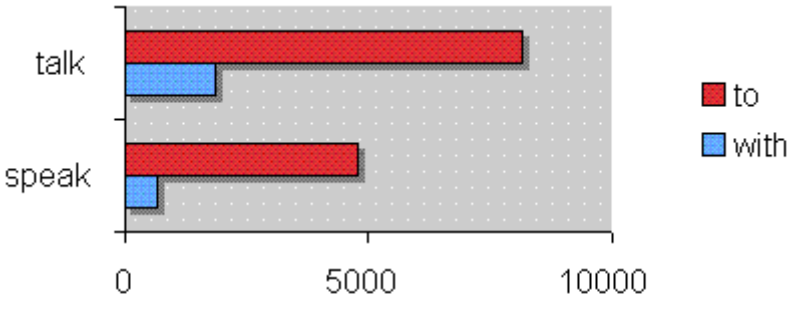
Two main conclusions can be drawn. As for the constructions «X talk» and «X(-)speak», we observe that *talk* and *-speak* behave similarly once more, and not only in what refers to the semantics of the words but also in the syntax. The nominalization of these two speech acts is part of a general trend in the English language to package information in noun groups and to take advantage of the syntactic resource of compounding (here, in the form of compound nouns). Note they are both used as suffixes rather than prefixes. Although slight shifts in meaning may result, most of the resulting compounds *(techspeak* vs *tech-talk),* and the co-existence of *doublespeak* and *double-talk* show they are practically synonyms and it is likely with the evolution of the language one of the two forms will die out.

5. Syntax

A noteworthy particularity about *speak* and *talk* is that apart from sharing semantic features they are syntactically interchangeable, unlike *say* and *tell*.

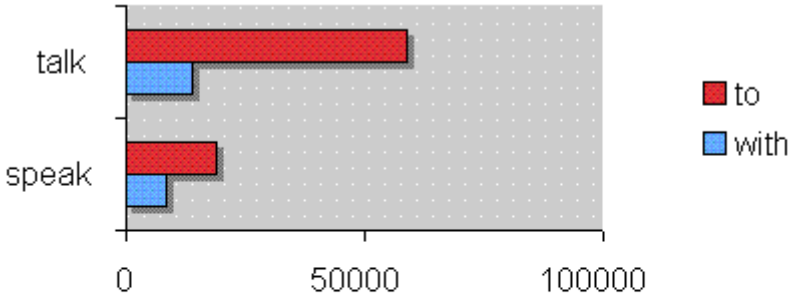
5.1. Speak *and* talk *in combination with* to *and* with

We were interested in finding out how *speak* and *talk* combine with two of the most common prepositions, *to* and *with*.



	speak	talk
to	4801	8116
with	663	1859

Figure 2. *Speak* and *talk* + *to*/*with* in the BNC



	speak	talk
to	19346	58708
with	8630	13858

Figure 3. *Speak* and *talk* + *to/with* in the COCA

If we look at the matches in the BNC (Fig. 2) and the COCA (Fig. 3), we see that *to* is clearly the most frequent in both corpora, both in combination with *speak* and with *talk*, although *with* seems to be a bit more frequent in the COCA than in the BNC, in particular in combination with *speak*.

Although this is only a quantitative analysis, a closer look reveals that there is little or no relationship between *to* and one-directionality on the one hand and *with* and bidirectionality

on the other. In fact, the data show that *to* is simply the unmarked preposition to be used in combination with both *speak* and *talk.*

5.2. Speak *and* talk *in combination with* about

The BNC reveals that *talking about* something is much more common than *speaking about* something (see Table 2), and the same results are duplicated in the COCA. One could argue that this result is to be expected, since we have already seen (cf. section 2 above) that *talk* is more frequent than *speak* in both corpora.

Table 3. *Speak/talk about* in the BNC & COCA

BNC

		TOT
1	TALKING ABOUT	5338
2	TALK ABOUT	3988
3	TALKED ABOUT	1409
4	TALKS ABOUT	414
5	TALKIN' ABOUT	19
	TOTAL	11168

		TOT
1	SPOKE ABOUT	188
2	SPEAK ABOUT	147
3	SPEAKING ABOUT	97
4	SPOKEN ABOUT	87
5	SPEAKS ABOUT	17
6	SPEAKIN' ABOUT	1
	TOTAL	537

COCA

		TOT
1	TALKING ABOUT	48440
2	TALK ABOUT	46824
3	TALKED ABOUT	14958
4	TALKS ABOUT	4762
5	TALKIN ABOUT	23
	TOTAL	115007

		TOT
1	SPOKE ABOUT	1200
2	SPEAK ABOUT	1125
3	SPEAKING ABOUT	753
4	SPOKEN ABOUT	314
5	SPEAKS ABOUT	235
	TOTAL	3627

However, note that the difference between *speak* and *talk* in this case is strongly related with the frequency of use of the different verb tenses: *talking* in the first place, followed by *talk, talked,* and *talks*, whereas we have *speak spoke, speak, speaking* (and spoken) -- and that order is is to be found both corpora. In other words, the progressive aspect is much more likely to be used with *talk* than it is with *speak*; whereas the past tense is more likely to be used with *speak*. *Talking about* something is 3.8 times more common than *talked about* in the BNC and 3.2 times more common in the COCA (the figures are stable across the two corpora). However, *spoke about* something is 1.9 times more common than *speaking about* in the BNC and 1.6 times more common in the COCA (in other words, nearly twice as common).

In the view of this, the following question arises: could it be that the choice of the tense motivates (or at least is somehow related to) the choice of the verb? Or does each of the verbs inherently carry information about the aspect? *Talking* may be considered semantically as more of an ongoing process (more interactive perhaps). The data clearly shows that the progressive aspect is much more likely to be used with *talk*. You are

speaking *about* something approximately 0.00019% of the time according to the COCA; whereas you are ***talking*** *about* something 0.012%. They both seem to be very low percentages but the important thing here is the relative difference between them because we are abstracting data from a very large corpus (425 million words). For every time a native speaker has the choice between using *talking about* and *speaking about* he or she are 63.7 times more likely to use the former. In other words, in a network of choices about language the unmarked choice is *talking about.*

6. Conclusions and further research

We have broadly surveyed some of the difficulties found in the semantic characterization of *speak* and *talk.* Dictionaries do not always succeed at providing all the information needed to determine which one to choose where Romance languages would use *parler, hablar,* etc. A diachronic analysis might reveal to what extent *talk* is gaining ground on *speak* in environments historically reserved for the latter. On the other hand, we have seen that *to* is the unmarked preposition in the *talk/speak + to/with someone* environment. Finally, we have noted the apparently close relationship between sense and syntax in the *talk/speak about* sequence.

References

Buck, C. D. (1915). Words of Speaking and Saying in the Indo-European Languages: First Paper. *The American Journal of Philology* Vol. 36, No. 1, 1-18

Davies, M. (2011). Word frequency data from the Corpus of Contemporary American English (COCA). Downloaded from http://www.wordfrequency.info on April 03, 2011.

Davies, M. (2004-). BYU-BNC: The British National Corpus. Available online at http://corpus.byu.edu/bnc.

Davies, M. (2008-). The Corpus of Contemporary American English (COCA): 425 million words, 1990-present. Available online at http://www.americancorpus.org.

Dirven, R., Goossens, L., Putseys, Y. and Vorlat, E. (1982). *The scene of linguistic action and its perspectivization by SPEAK, TALK, SAY and TELL*. Amsterdam: Benjamins.

Lehmann, D. (1977). A confrontation of say, speak, talk, tell with possible German counterparts. *Papers and Studies in Contrastive Linguistics* 6: 99-109.

Vincent, J. (2001). Talk-speak: gioco e ideologia nei logonimi inglesi, in C. Vallini, ed., *Le Parole per le parole - I logonimi nelle lingue e nei metalinguaggi*. Roma: Il Calamaio, 701-738.

Is automatic production of dictionary entries in the first Slovene online dictionary of abbreviations *Slovarček krajšav* possible?

Mojca Kompara

Department of Comparative and General Linguistics, Faculty of Arts, University of Ljubljana, Slovenia

Abstract

The possibility of automatic production of dictionary entries in the first Slovene online dictionary of abbreviations Slovarček krajšav in Termania software is discussed in this paper. In the first step, a demonstration algorithm has been used which focuses on the automatic recognition of abbreviations and abbreviation's expansions. The upgraded algorithm is used on a Slovene corpus of over 60 million words.

The acquired data is manually cleaned; good pairs are verified and used for production of the first Slovene abbreviations' dictionary. Simple entries are produced entirely automatically, complex, "semi" automatically. In simple and complex entries we are focusing on the automatic production of nominative Slovene structures of abbreviation's expansions out of non nominative. The main problem in complex entries are encyclopaedic data and translations for now included manually.

The algorithm for automatic recognition is the link between the electronic text and the "semi" automatically produced dictionary of abbreviations (Kompara, 2009).

Key words: abbreviations, dictionaries, "semi" automatic, automatic.

Resumen

En este articulo se trata de la posibilidad de producción automática de las entradas del diccionario en el primer diccionario esloveno de abreviaturas Slovarček krajšav utilizando software Termania. En primer paso, un algoritmo de demostración se ha utilizado que se centra en el reconocimiento automático de abreviaturas y de las expansiones. El algoritmo actualizado se utiliza para analizar un corpus esloveno de más de 60 millones de palabras.

Los datos obtenidos se limpian manualmente; pares buenos son verificados y utilizados para la producción del primer diccionario esloveno de abreviaturas. Entradas simples son producidas enteramente automáticamente, complejas, en forma "semi" automática. En las entradas simples y complejas nos centramos en la producción automática de estructuras nominativas eslovenas de expansiones con estructuras no nominativas. El problema principal en las entradas complejas son datos enciclopédicos y traducciones que por ahora tienen que ser incluidos de forma manual.

El algoritmo de reconocimiento automático forma el vínculo entre el texto electrónico y la producción "semi" automática del diccionario de abreviaturas (Kompara, 2009).

Palabras clave: abreviaturas, diccionarios, "semi" automático, automático.

Introduction

The scope of this article is to present the possibility of automatic production of dictionary entries in the first Slovene online dictionary of abbreviations *Slovarček krajšav* (Kompara, 2006) in *Termania* (Amebis, 2010) software. The paper presents the newly build Slovene software for dictionary production *Termania* and the possibility of automatic production of abbreviations' dictionary entries.

Why automatic approach?

In Slovene we still do not have a contemporary dictionary of abbreviations, although the very first one *Kratice, Mala izdaja* (Župančič, 1948) was written in 1948. Unfortunately is it too old to include new abbreviations and there are just few copies of it available to the public. Fortunately we have an online dictionary of abbreviations *Slovarček krajšav,* containing over 5,000 entries and providing translations of all foreign abbreviations. The dictionary was published in 2006 as my undergraduate university project on the web page of the *Fran Ramovš Institute of Slovenian Language*. The web page is freely available for public in Slovene and English language. It took me three years to collect the abbreviations, the abbreviations' expansions, and the translation, to add qualifiers, encyclopaedic data and to correct it. In order to optimise the process in my future work I used the automatic approach for recognition of abbreviations and abbreviations' expansions from electronic texts.

Automatic approach of recognition

The pioneer in automatic recognition of abbreviations and abbreviation's expansion is Taghva (1998). Automatic recognition was dealt also by Yeast (1999), Larkey, Ogilvie, Price and Tamilio (2000), Schwartz and Hearst (2003), Park and Byrd (2001), Chang (2002) but they all deal just with recognition in English texts. The first multilingual approach was made by Zahariev (2004). His approach is considered special due to the fact that he is not limiting just to one language recognitions, he recognises abbreviation-expansion pairs also in Chinese, Japanese and other exotic languages. The preparation of the algorithm had many phases but gave the possibility to come across the abbreviation-expansion pairs suitable for the automatic production of the dictionary of abbreviations. In the last stage of the development the number of letters in the abbreviations was extended from five to ten and both left and right context were observed. The abbreviation-expansion pairs were recognised using the lexical method of recognition. All four types of pattern: *(abbreviation) expansion, (expansion) abbreviation, abbreviation (expansion), expansion (abbreviation)* were used. After the newly established rules for recognition a demo version of the algorithm was produced. The system called *MKstrings* is composed of two windows, in the first one we add text rich in abbreviations, after clicking *Click here to process data* in the second window abbreviations and expansions occur as can be seen from figure 1.

The algorithm is not taking into consideration abbreviations with no expansions and abbreviations with numbers. The algorithm was improved using randomly selected texts

Figure 1. Demo version of the algorithm

rich in abbreviations (from the website *24ur.com*). Problems occurred mainly in examples containing the abbreviation *e.g. RS* in the expansion and non capitalised abbreviation *e.g. DARS* in *Družba za avtoceste v RS (Dars)*. But also after taking into consideration this step the problem was still not solved. Prepositions *za* and *v*, represented a problem too, because at the present stage the algorithm was able to consider just one preposition in the expansion. Problems occurred also in some copy-pasted examples e.g. *Urada za varstvo konkurence (UVK)*, recognised when retyped. An interesting issue are also patterns composed of a foreign abbreviation and a Slovene expansion, e.g. *Združenje evropskih avtomobilskih proizvajalcev (ACEA)*. Such patterns were not observed in the present article and will be recognised in the future. The modified and improved software is able to filter larger amounts of data. A larger corpus composed of 60 million words (newspaper *Delo* from 2005 to 2009) was used. The algorithm filtered the corpus in 30 minutes and gave 5,820 abbreviation-expansion pairs. The obtained pairs were manually revised and verified using *Google*. Among the revised and verified pairs 4% of false pairs occurred, e.g. *PO predstavljenih podatkih o, NA na vse argumente, IN in novincev*. The precision of the algorithm is 96%. Among the revised and verified pairs were also genuine abbreviations not matching with the right expansions, e.g. *HIV virusom i hepatitisom in virusom* because the expansions were missing. Another problem was the occurrence of the same abbreviation-expansion pair. The algorithm took into consideration all the possible expansions of one abbreviation. For the abbreviation *MNZ*, three expansions were genuine, *Ministrstvo za notranje zadeve*, *Medočinska nogometna zveza* and *Muzej novejše zgodovine*. After the exclusion of all false pairs, verification and revision of 2,665 genuine abbreviations-expansion pairs occurred. Among the Slovene analysed texts the algorithm recognised also some foreign abbreviations. Among the foreign some problems occurred in misrecognition of parts of expansions, e.g. *FEE for Environmental Education*, where *Foundation* is missing. To discover whether the algorithm is universal[46] some English and Italian texts randomly selected and available online were used. Providing suitable stop list for foreign languages e.g. combinations of prepositions and articles (*preposizione articoltata*) for Italian and bearing in mind the typology of a foreign language we can talk about the universality of the rules for recognition.

44 Such approach was dealt also by Zahariev (2004).

4. Dictionary production in Termania software

The acquired data was used for automatic production of the dictionary of abbreviations. For entry editing *Termania* software is used. *Termania* is a free on-line dictionary portal with integrated dictionary browsing and editing tools developed by *Amebis* software company from Kamnik, Slovenia, in cooperation with *Trojina*, Institute for Applied Slovene Studies. It provides an interface for dictionary browsing and a simple but reasonably versatile on-line dictionary editing tool. The portal is intended for general public users with no specialized computer or lexicographic knowledge, but with an interest to share terminological or general language knowledge, either by offering translations in a bilingual or multilingual environment or providing definitions in a monolingual context. The portal is intended to serve as the central terminology data and opinion exchange node for Slovene terminology. The access is free of charge but registration is required.

4.1 Types of dictionary entries

Dictionary entries in the first automatically built Slovene online dictionary of abbreviations[47] are divided into simple and complex. An example of simple entry is visible from figure 2.

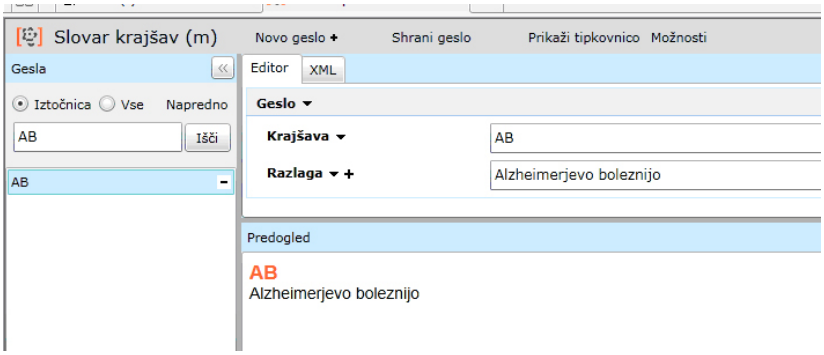


Figure 2. Example of simple dictionary entry

In figure 2 the editing tool (available just in Slovene language) is shown. On the left side you enter the desired abbreviation and press Search (*Išči*), on the upper part of the right side of the window you can see the editing tool. In this section one can correct its entry. The layout of the corrected article ready for publication is visible in the lower part of the right side of the editing tool. The dictionary article in figure 2 is a raw article that has to be appropriately edited. Simple entries are produced entirely automatically but manual verification and revision is still needed. When talking about simple entries we bear in mind mainly Slovene entries, covering just the abbreviation, the language qualifier and the expansion. As seen above the abbreviation-expansion pairs are recognised automatically by the algorithm for recognition, language qualifiers are added automatically too. In

47 Still in the process of editing and not accessible for public.

editing simple Slovene entries a problem occurred in expansions. In Slovene language nominative and non nominative structures are present and non nominative have to turn to nominative, as seen in example (1).

(1) **AB**

Alzheimerjevo boleznijo (non nominative structure)

Alzheimerjeva bolezen (nominative structure)

In Slovene we use the analyser module from *Presis* machine translation software to translate the texts to *Presis Interlingua*. For this task, the analyser module was changed in order to accept just noun phrases in various cases instead of both sentences and noun phrases. The result in *Interlingua* is checked to find out if the noun phrase is nominative. If it is not, the *Interlingua* result is changed to nominative and sent to the generator module of *Presis* to translate *Interlingua* back to Slovene. Furthermore the number of noun phrase is also checked since nominative dual or plural forms can be the same as non-nominative singular forms (*Alzheimerjeve bolezni*). But unfortunately such approach works just for Slovene language. The edited example of the simple entry where the language qualifier is added and the non nominative expansion is changed to nominative can be seen from figure 3.

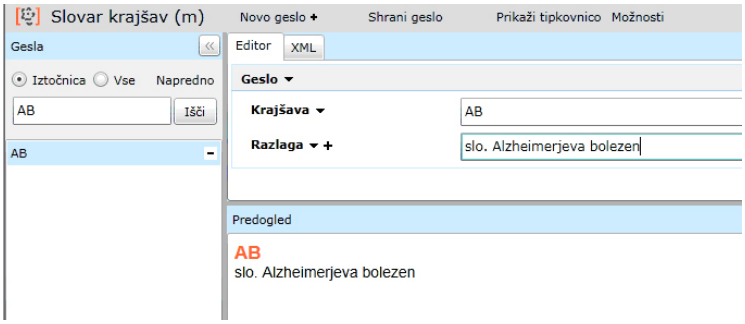


Figure 3. Example of edited simple dictionary entry

Such approach is used also in complex entries. Complex entries are considered those having Slovene and foreign expansion for one abbreviation. An example of complex entry is visible from figure 4.

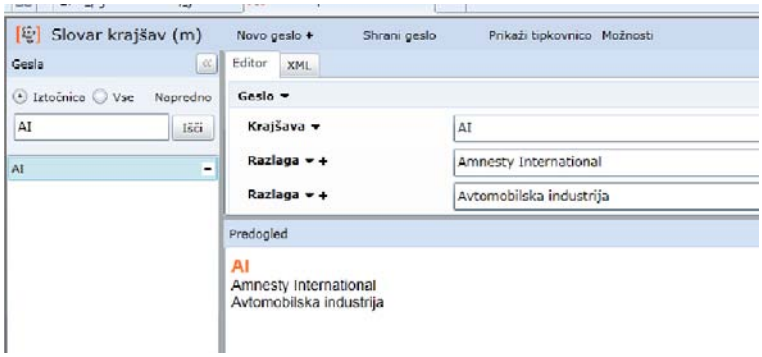


Figure 4. Example of complex dictionary entry

Automatic conversion of non nominative expansions to nominative is used also in complex entries. Language qualifiers are also added automatically for each foreign language. The edited complex entry is seen from figure 5.

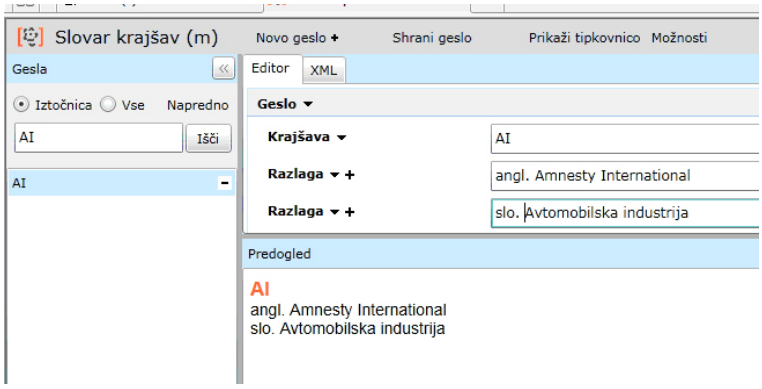


Figure 5. Example of edited complex dictionary entry

5. Under development – not yet automatised

The main problem in complex entries are encyclopaedic data and translations for now included manually, but in the future automatically. In terms of encyclopaedic data we have to define which abbreviation needs additional information in the dictionary article, as seen in example (2). Nowadays *AJ* is a well-known abbreviation and from this perspective encyclopaedic data is not needed but in just few years it's usage can be reduced and with it also it's actual meaning (not seen from the expansion). In such examples we ask ourselves where and when encyclopaedic data is really necessary.

(2) **AJ**

arab. Al Jazeera

Arabic news and current affairs TV channel

As a good example, I used types of entries with encyclopaedic data from the Italian dictionary of abbreviations *DidiSi* (Righini, 2001) where encyclopaedic data is consistently included and is not too short or too long and bears in mind the users needs. The main problem in providing encyclopaedic data is also the selection of reference. In online dictionaries a cross-reference could be used from the expansion to an encyclopaedia. In some cases encyclopaedic data could be omitted also using field specific qualifiers, e.g. (*mus*.) for *music*. Such approach is more applicable for paper dictionaries. The next problem are the translations that are divided in two subgroups. In the first, as seen from example (3) we have an entirely foreign abbreviation-expansion pair translated (for now) manually into Slovene. In the dictionary article the encyclopedic data is added at the end of the dictionary article. In example (3) we are looking of a translation in the surrounding text or we try to translate it using some translation tools.

(3) **AMS**

angl. American Mathematical Society

Ameriško matematično društvo

Ameriško društvo poklicnih matematikov posvečeno zanimanju za raziskovanje in učenje matematike.

The other subgroup comprises patterns composed of a foreign abbreviation and a Slovene expansion, e.g. *Združenje evropskih avtomobilskih proizvajalcev (ACEA)*. Such patterns were not observed in the present article but will be recognised in the future. In this particular case we already have the translation next to the abbreviation and we are looking for the original meaning, the expansion. Such pair cannot be recognised using the lexical approach, but statistical approach will have to be used instead.

6. Conclusion

Abbreviations became part of our everyday life and are produced on a daily base. Abbreviations are not something new or a fashionable way of communication; they were used even by Cicero (Kompara, 2005). The paper describes the automatic recognition of abbreviations and abbreviation expansions in electronic text and the possibility of automatic dictionary production of dictionary entries in the first Slovene online dictionary of abbreviations. The demo version of the algorithm is presented as well as the automatised approach for non nominative expansions and language qualifiers. Simple dictionary entries are produced entirely automatically but manual revision is still

needed. On the other hand complex entries represent a bigger issue. Non nominative expansion are turned nominative using automatic approach, language qualifiers are also added automatically, the main problem remain encyclopaedic data and translations. Some ideas for automatic insertion of them is also discussed in the papers. Entirely automatic production of dictionary entries in the first Slovene online dictionary of abbreviations *Slovarček krajšav* in *Termania* software is possible for simple entries but complex are still under development and are produced (for now) semi automatically. Automatic approach is for sure useful for the lexicographer but cannot replace one.

References

Amebis (1995). *Termania.* [online] Avaliable at: < http://www.termania.net/> [Accessed 16 February 2011].

Chang, J. T. (2002). Creating an Online Dictionary of Abbreviations from MEDLINE. *Journal of American Medical Informatics Association (JAMIA), IX(VI),* 612-620.

Google. [online] Avaliable at: <http://www.google.si/> [Accessed 16 February 2011].

Kompara, M. (2005). *Slovensko-italijanski glosar krajšav*. B.A. Ljubljana: University of Ljubljana, Faculty of Arts, Department of Translation.

Kompara, M. (2006). *Slovarček krajšav*. [online] Avaliable at: <http://bos.zrc-sazu.si/kratice.html> [Accessed 16 February 2011].

Kompara, M. (2009). Prepoznavanje krajšav v slovenskih elektronskih besedilih. *Jezikoslovni zapiski 15, 1-2,* 95-112.

Kompara, M. (2010). Automatic recognition of abbreviations in electronic texts. *Interlingüística 21,* 654-661.

Larkey, L. S., Ogilvie, P., Price, M. A., and Tamilio, B. (2000). Acrophile: An Automated Acronym Extractor and Server. In *Proceedings of the fifth ACM conference on Digital libraries*. Retrieved from http://www.sciweavers.org/publications/acrophile-automated-acronym-extractor-and-server.

Park, Y., & Byrd, R. J. (2001). Hybrid TextMin-ing for Finding Abbreviations and Their Definitions. *IMB Thomas J. Watson Research Center*, 167-170.

Righini, E. (2001). *Dizionario di Sigle Abbreviazioni e Simboli*. Bologna: Zanichelli.

Schwartz, A. S., & Hearst. M. A. (2003). A simple algorithm for identifying abbreviation definitions in biomedical texts. *Proceedings of the Pacific Symposium on Biocomputing*. Retrieved from http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.2481.

Taghva, K. & Gilbreth, J. (1998). Recognizing acronyms and their definitions. *IJDAR* I(IV), 191-198.

Zahariev, M. (2004). *A (Acronyms)*. Ph. D. Ottawa: School of Computing Science, Simon Fraser University.

ŽUPANČIČ, J. (1948). *Kratice, Mala izdaja*. Ljubljana: DZS.

YEATES, S., 1999. Automatic extraction of acronyms from text. *Proceedings of the Third New Zealand Computer Science Research Students' Conference*. Retrieved from http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.1738.

24UR.COM. [online] Avaliable at: <http://24ur.com/> [Accessed 16 February 2011].

Combined approach to modern lexicographic tools: the case of the first Slovene dictionary of tourism terminology

Mojca Kompara, Ana Beguš and Elena Sverko
Faculty of Humanities, University of Primorska

Abstract

The paper presents the newly built Slovene Dictionary of Tourism Terminology, compiled on the basis of the Multilingual Corpus of Tourist Texts (Mikolič, Beguš, Dukič, Koderman, 2008). The Corpus includes 27 million words, mostly in Slovene, but also in English and Italian, thus representing a bigger multilingual LSP corpus for Slovene language (Mikolič et al., 2008).

The Dictionary of Tourism Terminology is being compiled using a newly designed software Termania (Amebis, 2010), which provides a flexible and user-friendly interface for editing dictionary entries. The dictionary currently consists of 2,000 terms. Automatic and manual approach were combined in the editing process.

The results show that automatic approach is useful and helpful for lexicographers but cannot replace them. The Dictionary of Tourism Terminology represents a good example of a corpus-based LSP dictionary in the electronic format, which represents an important trend of future development in the field of electronic lexicography.

Key words: dictionary, tourism, automation

Resumen

El artículo presenta el nuevo Diccionario esloveno de Terminología Turística, elaborado utilizando la base del Corpus multilingüe de textos turísticos (Mikolič, Beguš, Dukič, Koderman, 2008). El Corpus incluye 27 millones de palabras, sobre todo en esloveno, pero también en inglés e italiano, lo que representa un corpus más grande multilingüe LSP de lengua eslovena (Mikolič et al., 2008).

El Diccionario de Terminología Turística está siendo compilado con un nuevo software Termania (Amebis, 2010), que proporciona una interfaz flexible y fácil de utilizar para editar las entradas del diccionario. El diccionario se compone ctualmente de 2.000 términos. En la compilación del diccionario se combinan los métodos automáticos y manuales.

Los resultados muestran que el método automático es útil para el lexicógrafo, pero no lo puede reemplazar. Sin embargo, el Diccionario de Terminología Turística representa un buen ejemplo de un diccionario LSP de formato electrónico basado en corpus, lo que es una importante tendencia del desarrollo futuro en el campo de la lexicografía electrónica.

Palabras clave: diccionario, turismo, automático?

1. Introduction

The paper presents the newly built Slovene *Dictionary of Tourism Terminology*, compiled on the basis of the *Multilingual Corpus of Tourist Texts*[48] (Mikolič *et al*., 2008) and using the *Termania*[49] editing software. The compilation of terminological dictionaries represents one of the fundamental tasks of the language politics of every linguistic community, as stated in Mikolič, Beguš (2010). Although several terminological dictionaries (mainly in printed format) were compiled in the 1990s, there is still no contemporary explanatory dictionary of tourism available for Slovene. The only reliable explanatory sources therefore remain foreign dictionaries of tourism, which, however, do not cover specific Slovene tourism-related terminology. Furthermore, tourism as a mass phenomenon represents a composite activity that covers very diverse social and economic fields and is intensely present in the contemporary globalisation processes; therefore its terminology is highly relevant and rapidly developing. It is also one of the key areas where Slovene language and discourse enters into contact with other languages and discourses. For these reasons, the production of a contemporary dictionary of tourism seems essential.

The *Dictionary of Tourism Terminology* was created in the frame of an applied research project, funded by the *National Research Agency* (ARRS). It responds to the terminological inconsistencies and lack of Slovene terminology in the field of tourism, as perceived by experts working in the field of tourism. The aim of the project was therefore to gather, on the basis of the analyses of corpus data which show actual language use, tourist terminology and to arrange it in dictionary (book and electronic) format, which is more familiar to users, as well as to point – through the connection between the *Dictionary* and the *Corpus* – at the importance and advantages of contemporary electronic language resources, also for languages for special purposes. The work on the project was carried out in more phases. Before beginning the work on the editing of dictionary entries in the *Dictionary of Tourism Terminology*, some basic considerations had to be taken into account. One of the basic determinants for every dictionary is its target audience. The target audience for the *Dictionary of Tourism Terminology* were Slovene experts in the field of tourism who felt they needed appropriate terminological resources in order to carry out their work at home and abroad. As already mentioned, this specific target group was also the group that inspired the project in the first place. Secondly, the broader target audience were translators, language and technical editors and other authors of tourist texts. The *Dictionary* also wished to serve as a reference source on tourist terminology to the general public interested in this field. The *Dictionary of Tourism Terminology*, compiled on the basis of corpus data which shows authentic language use, therefore represents an important foundation for the entire tourism industry, ranging from tourist agencies, accommodations and other tourist service providers as well as cultural associations,

48 The basic research project *Multilingual Corpus of Tourist Texts – Information Source and Analytical Basis of Slovene Natural and Cultural Heritage* aimed to develop theoretical bases for a multilingual corpus of professional language and to draw up a Slovene-Italian-English corpus of tourist texts, and then to conduct an analysis of these texts based on theoretical starting points of intercultural pragmatics, translation theory, critical discourse analysis and terminological field by using a completed corpus. The corpus currently includes 27 million words in Slovene, English and Italian and is freely available at http://jt.upr.si/turisticnikorpus.

49 http://www.termania.net/.

economic organisations, as well as for educational and research institutions working in the field of tourism studies.

The *Dictionary of Tourism Terminology* will be available in book and printed form, but the electronic format of it was favoured. If at first, potential users were unmotivated to use the electronic format, due to its presumed technical complicatedness, this format is becoming increasingly familiar with the spread of electronic language resources on the Internet, and its advantages, such as ease of browsing, hipper linking, inclusion of video material, is clearly evident.

In order to identify terminological candidates for the *Corpus*, lists of words and collocations (monograms, bigrams and trigrams) were generated automatically from the *Corpus* and sorted by frequency. The *Corpus* was morphosyntactically annotated and lemmatised with the *TOTALE* software (Erjavec, 2006). The tags and lemmas were used for the selection of monograms on the basis of the lemmas, since they represent headwords for the *Dictionary of Tourism Terminology*. For bigrams and trigrams, lemmas were not useful; however, tags were used for eliminating impossible/improbable combinations of bigrams and trigrams. These lists, which were considerably long, were first cleaned automatically, and later also manually. The lists were automatically parsed with English, Italian and German lexicons, in order to eliminate all foreign words[50]. In the next phase, a special set of words to be eliminated automatically was prepared, such as certain connectors, various types of pronouns, prepositions, auxiliary verbs and adverbs, interjections and modal expressions. The 'cleared' lists were then manually checked for terms, using *Corpus* analyses, word sketches, and consultation with experts in the field of Tourism (*Slovene Tourist Organisation as project partner, Faculty of Tourism Studies of the University of Primorska*).

2. Compilation of dictionary entries – Automatic insertion of data into the Termania interface

A dictionary is a system that needs to be built, not a text that needs to be written (Humar, 2004). "Creating a dictionary involves making decisions: big decisions at the planning stage and – as the project goes forward – smaller ones on a day-to-day basis." (Atkins and Rundell, 2008). This was the leading attempt of our project. Suitable terminological candidates from the *Corpus* were inserted automatically into the *Termania* editing software, where they were edited. *Termania* is a free on-line dictionary portal with integrated dictionary browsing and editing tools developed by *Amebis* software company from Kamnik, Slovenia, in cooperation with *Trojina*, Institute for Applied Slovene Studies. It provides an interface for dictionary browsing and a simple but reasonably versatile on-line dictionary editing tool. The portal is intended for general public users with no specialized computer or lexicographic knowledge, but with an interest to share terminological or general language knowledge, either by offering translations in a bilingual or multilingual environment or providing definitions in a monolingual context.

50 Loanwords, appearing in Slovene tourist discourse, such as 'last minute' or 'all-inclusive' were searched for independently in a separate list of English monograms, bigrams and trigrams.

The portal is intended to serve as the central terminology data and opinion exchange node for Slovene terminology. The access is free of charge, but requires registration (Krek, 2011). In the pre-editing phase, certain data was entered automatically in the interface in order to facilitate the lexicographers' work. Automatically inserted were headword, language qualifier, word class, field-specific qualifiers, example of use and translation into English.

3. Types of entries

Entries were classified as empty and partially completed. Empty entries were those not inserted automatically, such as all new entries not present on the existing list of candidate, but still considered relevant and necessary for the dictionary, e.g. synonyms, cross-references etc. An example of an empty entry is visible in Figure 1.

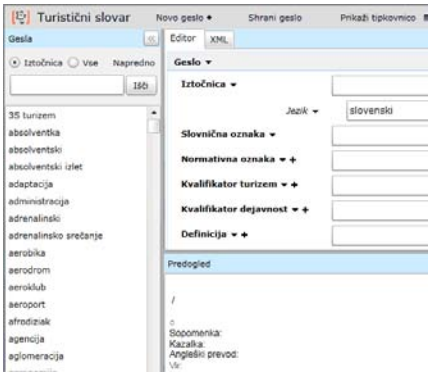


Figure 1. Example of empty entry

The editing software is available only in Slovene and, as seen in Figure 1, consists of the *Search tool* (*'Išči'*) on the left side of the window. The upper part of the right side of the window shows the empty fields to be completed or edited, and below is the *Preview* (*Predogled*). As seen in Figure 1, the *Preview* section shows empty fields for *synonym* (*sinonim*), *cross-reference* (*kazalka*), *English translation* (*Angleški prevod*) and *Source* (*Vir*). All the other entries were partially completed. An example of a partially completed entry is visible in Figure 2.

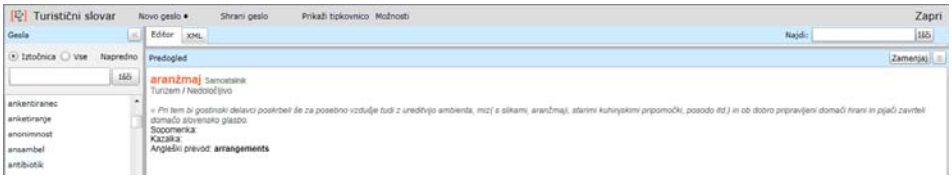


Figure 2. Example of partially completed entry

Figure 2 shows an example of a partially completed entry where all the data is inserted automatically, but the entry still requires manual editing before being published. In the *Preview section* we can notice the *headword* (*aranžmaj*), the *word class* (*samostalnik*), the *field-specific qualifiers* (*Turizem/Nedoločljivo*) and an *example of use* extracted automatically from the *Corpus* (*Pri tem bi gostinski delavci poskrbeli še za posebno vzdušje tudi z ureditvijo ambienta, miz (s slikami, aranžmaji, starimi kuhinjskimi pripomočki, posodo itd.) in ob dobro pripravljeni domači hrani in pijači zavrteli domačo slovensko glasbo*). The translation into English (*arrangements*) is given last and as seen from the example in many cases was not appropriate and for that reason it has to be changed.

4. Editing procedure in Termania interface

Tourism represents a large area covering many subfields, such as sea tourism, snow tourism, luxury tourism etc. Due to the fact that all subfields differ among each other, lists of suitable references were prepared in advance for each subfield. Such lists were subject to correction as new dictionaries or sources were added during the editing phase. Since ours was the first *Dictionary of Tourism Terminology* for Slovene and other specific terminological glossaries and/or dictionaries for Slovene are still not frequent enough, we used mainly foreign references. A supporting tool for editing entries in *Termania* was *Sketch Engine,*[51] an online tool designed for anyone wanting to research how words behave. It is a *Corpus Query System* incorporating word sketches, one-page, automatic, corpus-derived summary of a word's grammatical and collocational behaviour (Sketch Engine, 2011). It gives access to large corpora (30M-10B words) for 42 languages and gives the possibility to build your own corpus.

The *Multilingual Corpus of Tourist Texts* was inserted into *Sketch Engine*, which allows searching for concordances or word sketches. Suitable definitions were usually checked in the concordances. Concordances were used also when the automatically obtained examples of usage had to be changed with more suitable. Figure 3 shows the word sketch

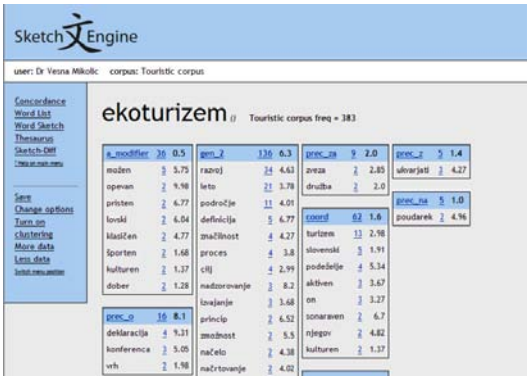


Figure 3. Word Sketch for the term "ekoturizem"

51 http://www.sketchengine.co.uk/

for "*ekoturizem*" (ecotourism) and the word's grammatical and collocational behaviour. With the help of such sketches, suitable (tourism-oriented) collocations were manually entered into the editing mask. The main problem in *Word Sketch* is the fact that it works properly just for monograms and not for bigrams or trigrams.

5. Manual approach

The final editing process was entirely manual. The partially edited entries were revised and checked for suitability. Added were new fields, one or more definitions, collocations using *word sketches*, synonyms and cross-references. Every cross-reference was added entirely manually to the dictionary. Sometimes examples of usage were replaced with more suitable ones. References were added at the end of the dictionary entry. Definitions represented the main problem in the editing process. A complex issue were also collocations, synonyms and cross-references; for the simple reason that we had to stop at a certain point. We set the number of suitable collocations, synonyms and cross-references to maximum five per entry. All the definitions were tourism oriented, and in many cases entirely new. When writing definitions we avoided too long, too short and encyclopaedic definitions. We focused on the needs of the users. Translations, inserted automatically were checked for suitability and replaced if not suitable.

6. Conclusion

In the present article we showed the process of compiling dictionary entries for the *Dictionary of Tourism Terminology*, which represents the first terminological dictionary of tourism for Slovene. The automated approach was used as an aid to lexicographers in the initial phases of the editing process. The results of the project show that the automatic approach in compiling LSP dictionaries is useful and helpful for the lexicographers, but cannot replace them, since some stages in the editing of dictionary entries cannot be fully automated. The first Slovene *Dictionary of Tourism Terminology* represents one of the first terminological dictionaries that was compiled using the *Termania* editing software, which has proven to be very useful in both the automatic and manual phase and thus represents an important contribution to the field and a future trend in computational lexicography.

References

Amebis (2010). *Termania.* Avaliable at: < http://www.termania.net/> [Accessed 16 February 2011].

Atkins, B. T., & Rundell, M. (2008) *The Oxford Guide to Practical Lexicography*. Oxford, New York: Oxford University Press.

Erjavec, T. (2006). Multilingual tokenisation, tagging, and lemmatisation with totale. Paper presented at the 9th INTEX/NOOJ Conference, Belgrade, Serbia, June 1-3.

Humar, M. (2004) Stanje in vloga slovenske terminologije in terminografije. Terminologija v času globalizacije. *Zbornik prispevkov s simpozija Terminologija v času globalizacije, Ljubljana, 5.-6. junij 2003*, 20–21.

Krek, S. (2011). Termania – Free On-Line Dictionary Portal; Retrieved from: http://www.simonkrek.si/Objave/5Krek_SD011.pdf

Mikolič, V., Beguš, A. (2010). Identifikacija terminov za turistični terminološki slovar. *Ann, Ser. hist. sociol., 20, 1*, 233-240.

Mikolič, V., Beguš, A., Dukič, D. & Koderman, M. (2008) Vpliv namembnosti korpusa na označevanje besedilnega gradiva za "Večjezični korpus turističnih besedil". *Zbornik Šeste konference Jezikovne tehnologije, 16. do 17. oktober 2008: zbornik 11. mednarodne multikonference Informacijska družba*, 60-64.

Sketch Engine (2010). Avaliable at: < http://www.sketchengine.co.uk/> [Accessed 16 February 2011].

Turistični korpus (2010). Avaliable at: < http://jt.upr.si/turisticnikorpus> [Accessed 16 February 2011].

La compilación de DiCoEnviro en español

María Teresa Ortego Antón
Departamento de Lengua Española – Universidad de Valladolid

Los componentes del OLST, conscientes de que los diccionarios especializados de medio ambiente no satisfacen las necesidades de los usuarios, se propusieron crear un diccionario electrónico en inglés, francés y español, DiCoEnviro, Dictionnaire fondamental de l'environnement, que sigue los principios de la lexicología combinatoria y explicatoria.

En el presente artículo explicamos la metodología seguida para la elaboración de las entradas en español a partir del análisis y extracción de datos de un corpus sobre medio ambiente compilado previamente: la selección de términos, la elección de los contextos, la descripción de la estructura actancial y de las relaciones léxicas y la vinculación de equivalentes.

Palabras clave: corpus, DiCoEnviro, lexicología combinatoria y explicativa, medio ambiente, término.

Members from OLST, being aware that specialized dictionaries about environment do not satisfy users' needs, decided to develop an e-dictionary in English, French and Spanish entitled DiCoEnviro, Dictionnaire fondamental de l'environnement, which follows the principles of Explanatory Combinatorial Lexicology.

In this paper, we explain the methodology to create the Spanish entries by analyzing and extracting data from an environmental corpus compiled previously: the headword selection, the choice of contexts, the description of actantial structure and lexical relations and the linking of equivalents.

Key words: corpus, DiCoEnviro, explanatory and combinatory lexicology, environment, term.

1. Introducción

El dominio del medio ambiente cada vez cobra más importancia como consecuencia de la diferenciación por actividades socioeconómicas y los expertos necesitan traductores que permitan que el conocimiento traspase las barreras lingüísticas. Los traductores, a su vez, necesitan herramientas fiables como los diccionarios especializados. Sin embargo, la literatura lexicográfica española muestra la insatisfacción de los traductores con este tipo de herramientas (Gutiérrez Rodilla, 1998; García Palacios, 2002; Pastor y Alcina, 2010), entre otros) pero recientemente el panorama ha cambiado gracias a los avances de la informática y a la aplicación de la investigación en lingüística de corpus a la compilación de diccionarios.

Desde el OLST (Université de Montréal – Canadá), el equipo ÉCLECTIK, liderado por la profesora L'Homme (2007), consciente de la necesidad de un diccionario especializado sobre medio ambiente, siguiendo el ejemplo de DiCoInfo, se propuso crear un diccionario electrónico titulado DiCoEnviro, *Dictionnaire fondamental de l'environnement*, que sigue los principios de la lexicología combinatoria y explicativa (Mel'čuk et al. 1984-1999, 2007).

En este artículo pretendemos describir la metodología seguida para elaborar las fichas terminológicas en español de DicoEnviro a partir del análisis y extracción de datos de un corpus sobre medio ambiente, que comenzó en septiembre de 2010. Por el momento, el número de fichas en español asciende a 146, aunque aproximadamente otras 50 están en construcción y no son visibles en línea.

2. Características del corpus

El corpus se compone de 85 archivos cuya extensión supera 1,5 millones de palabras (1.571.274 palabras) sobre el dominio del medio ambiente y el subdominio del cambio climático. Dado que el dominio del medio ambiente es muy vasto y heterogéneo, se optó por centrase únicamente en el subdominio del cambio climático.

La composición del mencionado corpus corrió a cargo de la investigadora Sarah Iveth Carreño Cruz en el marco de su memoria fin de máster.

Durante la selección de los textos que forman el corpus, se intentó incluir textos representativos de las variedades de español. De ahí que la procedencia geográfica de los textos sea variada: México, Argentina, Chile, Ecuador, España, El Salvador, Perú y Venezuela.

Las fuentes de las que proceden los textos que forman el corpus son organismos internacionales, que publicaron informes en línea.

En lo que concierne a la antigüedad de los textos, la mayoría fueron publicados entre 2000 y 2008, aunque una pequeña proporción fue publicada entre 1994 y 1999.

Para denominar los textos que forman el corpus, se ha incluido el nombre del organismo o su acrónimo y las iniciales del país de referencia.

La información que contiene el corpus es de gran relevancia porque permite extraer los términos, los contextos, la información gramatical, los actantes y las relaciones léxicas, semánticas, morfológicas y sintagmáticas.

3. Elección de los términos

La primera decisión es seleccionar los términos que formarán parte de la macroestructura de DiCoEnviro. A partir del corpus, utilizamos un extractor automático de términos: TermoStat Web 3.0 (Drouin, 2003). Este extractor genera una lista de candidatos a término basándose en criterios de especifidad. Cada candidato a término recibe una puntuación según la frecuencia en el corpus analizado y la frecuencia en otro corpus pretratado denominado corpus de referencia. Para español, los cinco primeros resultados son:

Tabla 1. Listado de los diez primeros candidatos a término extraídos con TermoStat 3.0

CANDIDATO A TÉRMINO	FRECUENCIA	ESPECIFIDAD	VARIANTES ORTOGRÁFICAS	MATRIZ
climático	3503	266.2	*climático* *climáticos* *climáticas*	Adjetivo
emisión	4657	264.47	*emisión* *emisiones*	Sustantivo
cambio	4535	180.64	*cambio* *cambios*	Sustantivo
carbono	1465	174.56	*carbono*	Sustantivo
gas	2086	171.08	*gas* *gases*	Sustantivo

Una vez que disponemos de la lista de candidatos a término, verificamos si cumplen cuatro parámetros (L'Homme, 2008: 90-91):

- Si denotan una entidad ligada a un dominio, por ejemplo *carbono* es un tipo de gas.
- Si presentan actantes de naturaleza especializada, por ejemplo *absorber*: un ecosistema absorbe radiación de la atmósfera.
- Si los vínculos morfológicos van acompañados de vínculos semánticos, por ejemplo dos unidades léxicas que pertenecen a una misma familia: *clima* y *climático*.
- Si existen otros vínculos paradigmáticos, por ejemplo dos términos que son sinónimos o antónimos, por ejemplo *absorción* y *emisión.*

Una vez seleccionados los términos, se pasa a la siguiente etapa, la redacción de la ficha terminológica.

4. La redacción de la ficha terminológica

En esta fase nos ayudamos de TextSTAT[52], un analizador de concordancias automático gratuito que busca en el corpus todas las ocurrencias de un término, con el contexto anterior y posterior. En nuestro caso, elegimos ver en pantalla 75 palabras antes y después del término escogido.

Una vez que tenemos todos los contextos en los que aparece el término, comenzamos con lo que propiamente se puede denominar la redacción de la ficha terminológica. Utilizamos una plantilla en formato XML y el programa Oxygen. La plantilla de ficha en blanco es común para todas las lenguas y similares a las utilizadas par DiCoInfo (L'Homme, 2008, 2009). Antes de comenzar con el análisis de datos, completamos la información de gestión: el autor (TOA) y la fecha (aaaa-mm-dd).

Continuamos escogiendo la forma de lematización del término, la más frecuente de los contextos y siguiendo estas normas:

- Los sustantivos se lematizan en singular.
- Los adjetivos se lematizan en masculino singular.
- Los verbos se lematizan en infinitivo.

Después, describimos las diferentes partes que componen la redacción de la ficha terminológica.

4.1. Significado y contexto

De la observación detallada de todas las ocurrencias del término en el corpus con el analizador de concordancias, distinguimos las diferentes acepciones (1, 2, etc.) y los significados próximos (1a, 1b, 1c). Ejemplo:

- "contaminante": contaminante$_1$ (sustantivo) y contaminante$_2$ (adjetivo);
- "río": río$_{1a}$ (tipo de ecosistema) y río$_{1b}$ (corriente de agua dulce).

Para cada significado diferente, intentamos extraer 20 contextos que sean representativos e ilustren los actantes y la fraseología con la que se utiliza la unidad léxica y la fuente de la que proceden. Se ordenan alfabéticamente según la denominación del texto del corpus del que proceden. En la versión en línea únicamente se pueden obtener los tres primeros contextos, por lo tanto estos contextos deben ilustrar con precisión los actantes y las posibles relaciones. Ejemplo:

> **ecosistema $_1$,** n.m.
>
> Contextos
>
> *Otras causas directas de la pérdida del capital natural son la contaminación de los ecosistemas, así como la introducción de especies invasoras y la variabilidad climática.* (Fuente: 3A_COMUN_MX)

52 http://neon.niederlandistik.fu-berlin.de/textstat/

Con un gran nivel de certeza se puede asegurar que el cambio climático hará que parte de los ecosistemas acuáticos continentales españoles pasen de ser permanentes a estacionales (Fuente: 4A_COMUN_UNFCC_ES)

Las temperaturas y humedad inusualmente altas parecen estar afectando a los ecosistemas boscosos. (Fuente: CC_AMERICA_LATINA)

4.2. La estructura actancial

Una vez que hemos delimitado las acepciones y hemos seleccionado los contextos, nos centramos en la estructura actancial, que L'Homme (2008: 94) describe como "*la structure actancielle décrit les participants essentiels pour décrire le sens d'un terme. Il ne s'agit pas d'une définition à proprement parler, mais cette structure fournit déjà des éléments importants sur le sens des termes*"[53].

La estructura actancial describe los participantes involucrados en el significado del término: los agentes, los pacientes, el destino, la superficie, la fuerza natural, etc.

A partir de nuestro conocimiento terminológico y con la ayuda de los contextos exponemos la estructura actancial. Por ejemplo, para absorber:

absorber $_1$, v. tr.

Estructura actancial: absorber: Destino {ecosistema 1, gas 1} ~ Paciente {radiación 1} de Fuente {atmósfera 1}

Una vez determinada la estructura y las realizaciones que servirán de ejemplo, rastreamos el corpus en busca de posibles unidades léxicas que actúen como Destino, Paciente y Fuente de absorber y las incluimos en orden alfabético. Si alguna de ellas fuera objeto de una ficha del diccionario, la vinculamos, de ahí que el navegador la muestre en azul y permita la referencia cruzada con otras fichas de DiCoEnviro en español.

absorber $_1$, v. tr.

Estructura actancial: absorber: Destino {ecosistema 1, gas 1} ~ Paciente {radiación 1} de Fuente {atmósfera 1}

Relaciones lingüísticas de los actantes

destino
aerosol $_1$, árbol, biosfera $_1$, bosque $_1$, ecosistema $_1$, gas $_1$, gas de efecto invernadero, mar, océano $_1$, planta, suelo, superficie, sustancia, Tierra, vapor, vegetal
paciente
calor $_1$, dióxido de carbono, energía $_1$, gas $_1$, radiación $_1$
fuente
atmósfera $_1$, sol, Tierra

53 "La estructura actancial describe los participantes esenciales para describir el significado de un término. No se trata de una definición en sentido estricto, sino que dicha estructura proporciona elementos importantes sobre el significado de los términos".

Destacamos que en el caso de términos de significado cercano, los actantes coinciden, como podemos comprobar entre absorber y absorción:

absorción $_1$, n. f.

Estructura actancial: absorción: ~ de Paciente {radiación 1} por Destino {sumidero}

Relaciones lingüísticas de los actantes

paciente
calor $_1$, carbono 1, dióxido de carbono, energía $_1$, gas $_1$, gas de efecto invernadero, radiación $_1$
destino
árbol, bosque $_1$, ecosistema $_1$, capa de ozono, planeta, sumidero, superficie

Una vez que hemos completado la estructura actancial, pasamos a la siguiente etapa, las relaciones léxicas.

4.3. Las relaciones léxicas

La penúltima etapa de la redacción de los artículos es confeccionar la lista de relaciones léxicas, compuesta por relaciones paradigmáticas y sintagmáticas. Para codificarlas se utilizan dos niveles: el primer nivel, despojado de metalenguaje técnico, va dirigido al usuario que desea acceder a la información sobre los términos y su combinatoria; y el segundo nivel, dirigido a los lingüistas, lexicógrafos, terminólogos y traductores. Existe un tercer nivel que no aparece en la web, que se refiere a la función léxica siguiendo las pautas de la lexicología y la semántica léxica (Polguère, 2008).

A su vez, esta categoría se compone de cinco subcategorías: significados relacionados, opuestos, otras categorías gramaticales y derivados, tipos de, combinaciones y otros.

4.3.1. Significados relacionados

En primer lugar, identificamos las variantes, sinónimos, cuasisinónimos y los significados relacionados. Si existen sinónimos cuyo grado de sinonimia es total o variantes, se describen en una rúbrica denominada “sinónimos”. Ejemplo:

carbono $_1$, n. m.

Sinónimo(s): C

El resto de relaciones de sinonimia se describen en el apartado “Significados relacionados”. A continuación exponemos varios casos de relaciones entre conceptos:

Tabla 2: Relaciones de sinonimia y de significados relacionados

Explicación - término típico	Explicación - rol actancial	Término relacionado
conservación	Sinónimo	preservación
erosionar	significado relacionado	desertificar
hielo	significado relacionado	glaciar

4.3.2. Opuestos

La siguiente etapa es la identificación de las relaciones de antonimia, oposición, etc.

Tabla 3. Relaciones de antonimia y oposición

Explicación - término típico	Explicación - Rol actancial	Término relacionado
conservar 1	Antónimo	dañar
capturar 1	Antónimo	emitir 2.1
disminución	Opuesto	subida
inundación	Contrastivo	sequía

4.3.3. Otras categorías gramaticales y derivados

Continuamos completando las relaciones morfológicas, es decir, términos que pertenecen a la misma familia. Para encontrar estas relaciones, se utiliza la búsqueda truncada. Por ejemplo, buscando atmósfer* obtenemos como resultado atmósfera y atmosférico.

4.3.4. Tipos de

Esta categoría se utiliza principalmente para describir los tipos de un determinado sustantivo encontrados en el corpus. Por ejemplo, los tipos de ecosistema:

ecosistema $_1$, n. m.

Relaciones léxicas

Roles actanciales

Explicación - término típico	**Explicación - rol actancial**	**Término relacionado**
Tipos de		
Tipo de e.	Tipo de «palabra clave»	río $_1$
Tipo de e.	Tipo de «palabra clave»	manglar $_1$
Que trata de un lugar específico	Que trata de un lugar específico	~ marino
Que trata de un lugar específico	Que trata de un lugar específico	~ terrestre

4.3.5. Combinaciones

Seguimos completando la información sobre la fraseología de los términos. Como apunta L'Homme (2008: 78-79), muy pocos diccionarios especializados describen las combinaciones léxicas típicas. Sin embargo, la información fraseológica es muy útil para el usuario en tareas de producción de L1 a L2 o en traducción inversa porque se describen los verbos con los que se utilizan los términos sustantivo, por ejemplo para bosque: aprovechamiento de un ~, deforestación de un ~, sostenibilidad de un ~, etc. Ejemplo:

ecosistema $_{1}$, n. m.

Relaciones léxicas

Roles actanciales

Explicación - término típico	**Explicación - rol actancial**	Término relacionado
Combinaciones		
Un e. se vuelve diferente	Una «palabra clave» se vuelve diferente	cambiar$_{1a}$
Sustantivo para algo o alguien que produce una variación en un e.	Sustantivo para algo o alguien que produce una variación en una «palabra clave»	cambio$_{2}$ en un ~
Algo o alguien constituye un riesgo potencial para la riqueza de un e.	Algo o alguien constituye un riesgo potencial para la riqueza de una «palabra clave»	amenazar$_{1}$ un ~
-> SUSTANTIVO	-> SUSTANTIVO	amenaza$_{1}$ para un ~
Sustantivo para la e. que se vuelve peor	Sustantivo para la «palabra clave» que se vuelve peor	degradación$_{1a}$ de un ~
Algo o alguien causa que un e. se vuelva peor	Algo o alguien causa que una «palabra clave» se vuelva peor	afectar$_{1}$ a un ~
Algo o alguien conserva un e. en su estado actual	Algo o alguien conserva una «palabra clave» en su estado actual	conservar$_{1}$ un ~
-> SUSTANTIVO	-> SUSTANTIVO	conservación$_{1}$ de un ~
Sustantivo para algo o alguien que conserva un e. en su estado actual	Sustantivo para algo o alguien que conserva una «palabra clave» en su estado actual	preservación$_{1}$ de un ~

4.3.6. Otros

El último apartado de las relaciones léxicas da cabida a todas las relaciones que no encajan en ninguno de los apartados anteriores pero que pueden resultar interesantes para los usuarios. Ejemplo:

atmósfera $_1$, n. f.

Relaciones léxicas

Roles actanciales

Explicación - término típico	**Explicación - rol actancial**	Término relacionado
Otros		
División	División	baja ~
División	División	estratosfera$_1$
División	División	troposfera$_1$

4.4. Los equivalentes

La última etapa en la confección de las entradas de DiCoEnviro en español es la búsqueda de equivalentes. Para vincular una unidad léxica con otra en inglés o francés, observamos si los significados coinciden y si los actantes en una y otra lengua son similares. Si coinciden, los relacionamos entre ellos. Por ejemplo, "ecosistema" está vinculado con el equivalente *ecosystem* en inglés y *écosystème* en francés.

Una vez completada la ficha, se revisa y se cambia el estatus de 3 a 2, con el fin de que se pueda consultar en línea. La consulta en línea puede realizarse mediante el buscador o mediante un índice alfabético de palabras. El resultado completo de una ficha puede consultarse en el apéndice.

5. Conclusión

Como podemos apreciar, DiCoEnviro, diccionario fundamental del medio ambiente, es un diccionario en fase de construcción que pretende describir los términos del medio ambiente (subdominio cambio climático). La novedad de esta obra reside en la descripción de las relaciones léxicas, semánticas y sintagmáticas a partir de la información de un corpus siguiendo los principios de la lexicología combinatoria y explicatoria y la vinculación de equivalentes en tres lenguas: inglés, francés y español.

Sin embargo, está todavía en fase de construcción. Entre las tareas pendientes, resaltamos la inclusión de más términos, la anotación de contextos y la definición de los significados. Además, por ahora los términos incluidos son univerbales, pero sería necesario incluir los términos pluriverbales, como por ejemplo "dióxido de carbono" y ampliar la cobertura a otras áreas del medio ambiente.

Su estatus "en construcción" podría ser una restricción para los usuarios, pero aún sin ser una obra acabada, esta herramienta es muy útil: es fiable (basado en corpus), accesible (disponible en Internet) e incluye información sobre las relaciones léxicas, semánticas y sintagmáticas, que no es fácil encontrar en otros diccionarios de tipología similar.

6. Agradecimientos

Al equipo ÉCLECTIK del OLST (Université de Montréal – Canadá), en especial, a su directora, la prof. Marie-Claude L'Homme, por ofrecerme la oportunidad de trabajar en el proyecto de redacción de las entradas en español de DiCoEnviro (la lista completa de investigadores puede consultarse en http://olst.ling.umontreal.ca/dicoenviro/equipe-es.html/); y a la Universidad de Valladolid por la Ayuda para Estancias Breves, sin su apoyo económico no habría sido posible la colaboración en el mencionado proyecto.

7. Apéndice

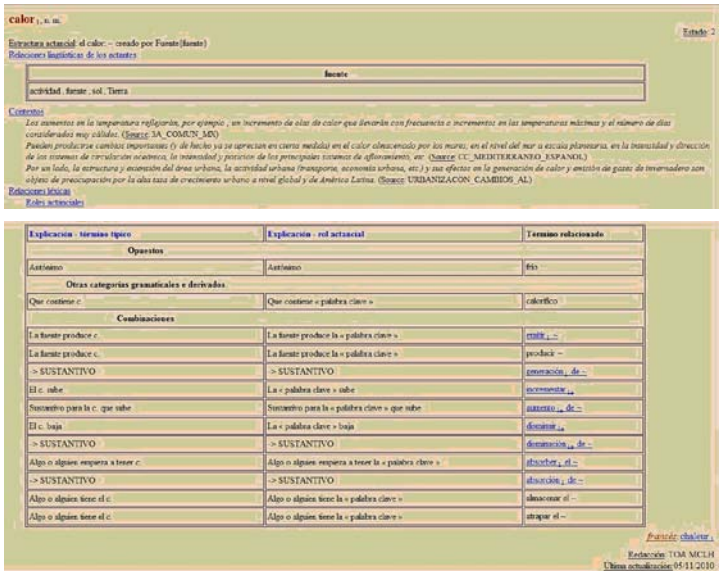


Figura A1.

8. Referencias bibliográficas

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, *9* (1), 99-117.

García Palacios, J. (2002). El artículo lexicográfico en el diccionario de especialidad. En I. Ahumada (2002). *Diccionario y lenguas de especialidad. V seminario de lexicografía hispánica. Jaén, 21 al 23 de noviembre de 2001*. Jaén: Universidad de Jaén.

L'Homme, M.C. (2007). Using Explanatory and Combinatorial Lexicology to Describe Terms. En L. Wanner (Ed.). *Selected Lexical and Grammatical Topics in the Meaning-Text Theory. In Honour of Igor Mel'cuk* (pp. 11-50). Amsterdam/Philadelphia: John Benjamins.

L'Homme, M.C. (2008). Le DiCoInfo. Méthodologie pour une nouvelle génération de dictionnaires spécialisés. *Traduire, 217*, 78-103.

L'Homme, M.C. (2009). *DiCoInfo: Dictionnaire fondamental de l'informatique et de l'Internet.* Disponible en http://olst.ling.umontreal.ca/dicoinfo/manuel-DiCoInfo.pdf [Fecha de consulta: 21 de marzo de 2011].

L'Homme, M.C. y Laneville, M.E. (2010). *DiCoEnviro: El diccionario fundamental del medio ambiente.* Disponible en http://olst.ling.umontreal.ca/dicoenviro/Dicoenviro_manual_Es.pdf [Fecha de consulta: 21 de enero de 2011].

Mel'čuk, I. y otros (1984-1999). *Dictionnaire explicatif et combinatoire du français contemporain*. Recherches lexico-sémantiques I-IV. Montréal: Les Presses de l'Université de Montréal.

Mel'čuk, I. y Polguère, A. (2007). *Lexique actif du français*. Bruxelles: Duculot.

Pastor, V. y Alcina, A. (2010). Search Techniques in Electronic Dictionaries: A Classification for Translators, *International Journal of Lexicography 23* (3), 307-354.

Polguère, A. (2008). *Lexicologie et sémantique lexicale. Notions fondamentales*. Montréal: Les Presses de l'Université de Montréal.

Rodilla, B. M. (1998). *La ciencia empieza en la palabra*. Barcelona: Península.

Análisis cuantitativo del uso real de los verbos pronominales estrictos del castellano utilizando un corpus diacrónico (Google Books)[54]

Irene Renau, Rogelio Nazar

Institut Universitari de Lingüística Aplicada

Universitat Pompeu Fabra, Barcelona

Resumen

Este trabajo presenta un análisis cuantitativo del uso de los verbos pronominales estrictos del castellano empleando un corpus diacrónico de grandes dimensiones (una parte de Google Books). Los verbos pronominales estrictos, comunes en todas las lenguas románicas, son aquellos que no pueden prescindir del pronombre que los acompaña, como en «Se ha fugado de nuevo». Tras recoger los lemas verbales pronominales de tres diccionarios generales de lengua española, se buscan en el corpus mencionado. Los resultados apuntan a que muchos verbos marcados en las fuentes lexicográficas como pronominales inherentes tienen variedad de estructuras transitivas e intransitivas en el uso. También se observa que muchas de las formas no se encuentran, lo que revela una distancia de la representación de estos verbos en los diccionarios, con respecto a los datos empíricos.

Palabras clave: *corpus diacrónico, Google Books, metalexicografía, verbos pronominales estrictos*

Quantitative Analysis of the Real Use of Spanish Strict Pronominal Verbs Using a Diachronic Corpus (Google Books)

Summary

*This paper presents a quantitative analysis of the use of Spanish strict pronominal verbs using a very large diachronic corpus (a part of Google Books). Strict pronominal verbs, common in all Romance languages, are those that cannot be used without the pronoun, as in "*Se ha fugado *de nuevo" (= 'S/He has escaped again'). After collecting the verbal pronominal lemmas of three Spanish general dictionaries, we looked for them in the mentioned corpus. The results show that a lot of verbs which appear in the lexicographical sources as strict pronominals have variety of transitive and intransitive structures in the real use. Also, it can be observed that many of the forms are not found, which reveals a divergence in the representation of these verbs in dictionaries, regarding the empirical data.*

Key words: *diachronic corpus, Google Books, strict pronominal verbs, theoretical lexicography*

54 Este trabajo ha recibido una subvención de los siguientes proyectos del MEC: «Agrupación semántica y relaciones lexicológicas en el diccionario», dir.ª: J. DeCesaris (HUM2009-07588/FILO); APLE: «Procesos de actualización del léxico del español a partir de la prensa», período 2010-2012, dir.ª: M. T. Cabré (FFI2009-12188-C05-01/FILO). También ha recibido subvención de la Fundación Comillas en relación con el proyecto de *Diccionario de aprendizaje del español como lengua extranjera.*

1. Introducción y objetivos

Este trabajo presenta un análisis cuantitativo del uso de los verbos pronominales estrictos del castellano empleando un corpus diacrónico de grandes dimensiones (una parte de Google Books). Los verbos pronominales estrictos (también llamados *puros*, *inherentes*, *intrínsecos* o de otros modos), comunes en todas las lenguas románicas, son aquellos que no pueden prescindir del pronombre que los acompaña:

> Si de algo puedo *jactarme* es de haber trabajado con intensidad.
>
> Se ha vuelto a escapar. *Se ha fugado* de nuevo.
>
> *Se personó* en casa de los Hernández.

Los verbos *jactarse, fugarse* o *personarse* de estos ejemplos no entran en el español actual en construcciones sin clítico, como **Él jacta de haber trabajado con intensidad, *Ha fugado de nuevo* o **El problema personó al hombre en casa de los Hernández.*

Se parte de la evidencia de que, en varios diccionarios actuales, los verbos pronominales estrictos han sido lematizados en algunos casos en su forma pronominal (*jactarse*) y en otros en su forma no pronominal (*jactar*). La finalidad del trabajo es estudiar estos verbos en un corpus diacrónico voluminoso que permita averiguar a qué se deben dichas divergencias, y que permita contrastar la información de las fuentes lexicográficas con datos reales de uso. Así pues, el objetivo que nos proponemos es a la vez lexicológico y metalexicográfico.

El artículo tendrá la siguiente estructura: en primer lugar, se expondrá la noción de verbo pronominal estricto tal como se ha descrito en las gramáticas, y también su representación en tres diccionarios generales de lengua española; a continuación, se establecerán las hipótesis de las que partirá el análisis de corpus; se describirá seguidamente este análisis, con sus resultados; el trabajo se cerrará con las conclusiones y trabajo futuro.

2. Los verbos pronominales estrictos en los estudios gramaticales

Los verbos pronominales —es decir, aquellos que van acompañados de pronombre clítico—[55] existen desde los inicios de las lenguas románicas como una de las consecuencias del paso del sistema de casos latino al nuevo sistema flexivo, y han sido estudiados ya en el comienzo de la tradición gramatical española (Nebrija menciona ya en su *Gramática castellana,* de 1492, el uso del pronombre reflexivo *se* como recurso del castellano para suplir la voz media o impersonal). En concreto, aparecen ya lemas solo pronominales en el *Diccionario de Autoridades* (1726-1739), por ejemplo, *abstenerse*, del latín *abstinere,* o *dignarse*, de *dignari.*

Pese a su documentación temprana, los verbos que no pueden prescindir del pronombre no han sido específicamente muy estudiados, posiblemente por considerarse que son el

55 Un estudio más amplio de este tipo de verbos pronominales y de otros se encuentra en la tesis doctoral de Renau (en preparación), de la que parte el presente estudio.

final de un proceso de gramaticalización que prácticamente ha lexicalizado al pronombre, convirtiéndolo en un rasgo morfológico del verbo y, por tanto, sin características sintácticas o semánticas propias. Un estudio más detallado de estas unidades léxicas muestra sin embargo su diversidad y especificidad[56]. De todos modos, se presta atención a los verbos pronominales estrictos en Sánchez López (2002: 96-102), Otero (1999: 1465-1472) y la *Nueva gramática de la lengua española,* RAE (2010: 3102-3103). En estos estudios, se explica, entre otras pocas cuestiones, que acostumbran a ser inacusativos y que suelen llevar complemento de régimen (*jactarse de, atreverse a, obstinarse en...*).

La variación en estos verbos «invariables» puede ser diacrónica (*arrogar, atrever, desvivir* y otros son formas antiguas del castellano, mientras que la lengua actuar solo admite *arrogarse, atreverse, desvivirse,* etc.; véase la figura) y diatópica (como en este ejemplo de Cuba: «Al presentarse los refuerzos por él solicitados, *adentró* al ejército más allá de lo deseado»[57]; en España y otros países se emplea solo *adentrarse*).También puede haber explotaciones expresivas[58] que conviertan un verbo empleado solo con pronombre en uno con estructura no pronominal: en un caso como «No creo que se suicidase. [...] *Le suicidaron*»[59], el verbo es explotado sintácticamente para sostener un cambio semántico (‘matar a alguien simulando un suicidio’).

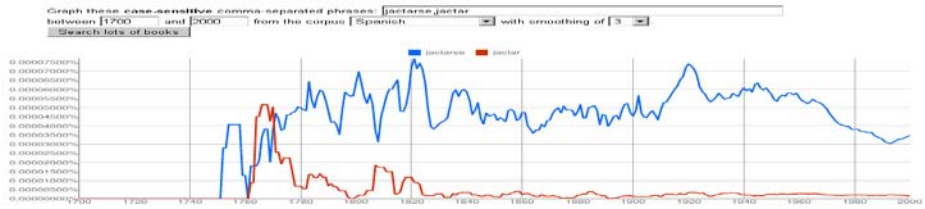


Figura. Gráfico ofrecido por Google Books N-gram Viewer (http://ngrams.googlelabs.com) que compara el uso de la forma jactar y jactarse desde el siglo XVIII hasta la actualidad.

3. Representación lexicográfica

Con el fin de observar la representación lexicográfica de los verbos pronominales estrictos, se tomaron tres diccionarios generales del español actual que tuvieran versión en CD-ROM y opciones de búsqueda avanzada: el *Diccionario de la lengua española* (DRAE), de la Real Academia Española; el *Diccionario de uso del español* (DUE), de María Moliner; y el *Diccionario de uso del español de América y España* (DUEAE), dirigido por Paz Battaner.

Se buscaron todos los lemas verbales que terminaran en *-se*, y se extrajeron manual o automáticamente. Los tres listados se compararon mediante un procedimiento también

56 Queda fuera del alcance del presente estudio profundizar en esta cuestión, que ha sido abordada para los verbos pronominales estrictos del catalán por Lorente (2010 y en prensa).

57 Ejemplo del Corpus del Español, de M. Davies (http://www.corpusdelespanol.org).

58 Para la noción de explotación frente a norma en el uso de la lengua, nos basamos en la Theory of Norms and Exploitations de Hanks (2004 y en prensa)..

59 Tomado del Spanish Web Corpus, disponible en Sketch Engine (http://the.sketchengine.co.uk).

automático, y se obtuvo un listado de 1.002 lemas en total. La tabla 1 muestra el número de lemas pronominales de cada diccionario empleado en comparación con el total de lemas verbales. Se observa que constituyen una parte ínfima del lemario verbal.

Tabla 1. Número de lemas pronominales en relación con el total de lemas verbales de los tres diccionarios empleados para el análisis.

	Lemas verbales	Lemas verbales prnls.	%
DRAE	12.017	760	0,70
DUE	10.928	587	0,59
DUEAE	6.653	335	0,55

En la tabla 2 se muestran los cruces entre obras.

Tabla 2. Número de lemas pronominales en el DRAE, el DUE y el DUEAE, y coincidencias de estos entre las tres obras.

	n	%
DRAE	760	75,85
DUE	587	58,58
DUEAE	335	33,43
DRAE + DUE + DUEAE	135	13,47
DRAE + DUE + no DUEAE	350	34,93
DRAE + DUEAE + no DUE	39	3,89
DUE + DUEAE + no DRAE	21	2,10
DRAE + no DUE + no DUEAE	236	23,55
DUE + no DRAE + no DUEAE	81	8,08
DUEAE + no DRAE + no DUE	140,00	13,97
DRAE o DUE o DUEAE	1.002	100

Puede observarse que las coincidencias mayores se observan entre DRAE y DUEAE. Por otro lado, es muy reducido el número de lemas coincidentes (n = 135, 13,47 %). En tercer lugar, un dato que no muestra la tabla es que, de los 1.002 lemas, 267 (26,65 %) tienen un equivalente no pronominal en alguna de las otras dos obras; por ejemplo, *jactarse* (DUEAE) se representa como *jactar* en DRAE y DUE.

4. Hipótesis

La variación observada en las tres fuentes lexicográficas consultadas motiva nuestro estudio de corpus. Las sospechas acerca de los distintos criterios puramente lexicográficos como origen de las diferencias tienen indicios evidentes: por ejemplo, el DRAE y el DUE contienen lemas anticuados (marcadas con la abreviatura *ant.*), mientras que el DUEAE solo refleja el uso actual. Pero nos preguntamos si tal diversidad es solo de origen lexicográfico, o podría haber diferencias lingüísticas reales que el corpus pudiera indicar. Se realizó una pequeña búsqueda manual preliminar en distintos corpus tanto sincrónicos como diacrónicos[60], y se vio que muchos verbos no aparecían ni una vez en ninguno de ellos, o aparecían tan pocos casos que los datos no eran concluyentes. Se decidió entonces emplear un corpus más grande.

Teniendo en cuenta los datos buscados en los diccionarios y esta pequeña exploración manual, se formularon las dos siguientes hipótesis:

- Hipótesis 1: La gran variación en la representación lexicográfica es reflejo de variación en el uso. Existen verbos que no son realmente pronominales estrictos pese a estar representados de este modo en los diccionarios.
- Hipótesis 2: Verbos considerados estrictamente pronominales son más frecuentes en forma de participio o tienen concomitancias con la estructura *estar + participio.*

Se explica a continuación el análisis de corpus realizado para contestar estas dos hipótesis.

5. Análisis de corpus

5.1. Descripción del corpus empleado

Se empleó un corpus de enegramas (de 1 a 5 enegramas) extraído de Google Books y puesto recientemente a libre disposición de la comunidad científica (Michel, Shen, Aiden, Veres, Gray, Google Books Team, et al., 2011). Este corpus está formado aproximadamente por 5 millones de libros publicados en varios idiomas desde 1500

60 Se buscaron los 23 primeros verbos del listado alfabetizado en el Corpus del Español (cit.); el Corpus Diacrónico del Español, CORDE (RAE), http://corpus.rae.es/cordenet.html; el Spanish Web Corpus (cit.); y el Corpus de Referencia del Español Actual, CREA (RAE), http://corpus.rae.es/creanet.html.

hasta 2008 y posteriormente digitalizados, y las unidades léxicas que se muestran en los enegramas han de haber aparecido al menos 40 veces. La parte de lengua española consta de 45.000 millones de palabras, lo que convierte a este corpus en el más grande jamás compilado para este idioma[61]. En concreto, para nuestros objetivos, un período tan amplio (1500-2008) salva la diferencia de atención de los tres diccionarios en cuanto a datos históricos.

5.2. Metodología

El corpus de Google es en realidad un índice de enegramas, es decir, una serie de tablas con secuencias de hasta cinco palabras, con indicación de frecuencia de aparición en cada año.

Este corpus no está lematizado, por lo tanto, los verbos tuvieron que buscarse por cada una de sus formas por medio de un procedimiento automatizado. Para obtener las formas de los lemas verbales se empleó la base de datos morfológica OSLIN (Janssen 2005). De los enegramas recuperados con las formas verbales, se contabilizaron como pronominales aquellos en los aparecía un pronombre enclítico o en concordancia con la forma verbal y como no pronominales el resto de los enegramas, excepto las formas en participio, que se separaron en una clase distinta. Se ignoraron casos considerados ambiguos, como aquellos en los que existe una coincidencia formal con otra categoría gramatical (como el sustantivo en *persona, fuga, cautela, hernia,* etc.). Se ignoraron también las combinaciones con verbos de soporte, los casos en los que el verbo analizado aparecía en posición inicial en el enegrama (excepto aquellos con enclíticos) y aquello en los aparecía un segundo verbo antes del verbo analizado. Por conveniencia práctica, se dividió el corpus en tres períodos: 1500-1870, 1871-1970 y 1971-2008, y los valores se expresaron en frecuencia relativa para compensar las diferencias en tamaño de las partes. El resultado del proceso es una matriz con los 1.002 verbos que registra su distribución de frecuencia en el tiempo en las clases *pronominal, no pronominal* y *participio*. En todos los casos, se calculó la significación estadística de las diferencias entre pronominal y no pronominal en el último período (test binomial), porque se trata de la franja que más interesa a nuestro estudio, y se ordenaron los verbos dividiendo su frecuencia como pronominales por su frecuencia como no pronominales.

6. Resultados

Los resultados de los tres períodos en que se dividieron los datos se resumen en la tabla 3.

En ella se muestra el número de verbos que tienen formas pronominales, no pronominales o en participio en los rangos porcentuales indicados en la primera columna. Por ejemplo, en el primer período (1500-1870) hay 68 verbos que tienen un 90-100% de formas pronominales; hay 100 verbos que tienen un 90-100% de formas no pronominales; y hay 26 verbos que tienen un 90-100% de formas en participio. Así, los valores de cada celda de la misma fila no se refieren a los mismos verbos.

61 Para lengua inglesa se están empezando a compilar corpus de grandes dimensiones, como el de Pomikálev, Rychly y Kilgarrif (2009). Por el momento, en español se dispone, entre otros, del Corpus del Español, de 100.000.000 de palabras; el CREA, de 160.000.000; y el CORDE, de 250.000.000.

Tabla 3. Número y porcentaje de formas pronominales, no pronominales y participios de los tres períodos en que se dividió el análisis.

	1500-1870			1871-1970			1971-2008		
Intervalos porcentuales	**Prnls.**	**No prnls.**	**Part.**	**Prnls.**	**No prnls.**	**Part.**	**Prnls.**	**No prnls.**	**Part.**
90-100 %	68	100	26	105	148	37	116	153	37
80-89 %	27	35	9	24	20	4	20	20	7
70-79 %	19	17	12	22	17	11	16	18	9
60-69 %	11	11	9	13	13	11	12	24	14
50-59 %	21	19	12	15	17	14	8	11	13
40-49 %	14	13	11	14	15	10	10	7	10
30-39 %	10	7	11	11	12	12	25	13	11
20-29 %	11	19	11	17	23	22	18	16	19
10-19 %	18	19	17	16	22	28	16	20	21
0,1-9 %	23	45	48	33	51	51	31	59	59
0,00 (sí OSLIN / No GBooks)	611	548	667	563	495	633	561	492	633
No OSLIN	169	169	169	169	169	169	169	169	169

6.1. Porcentaje de verbos pronominales estrictos

En los tres períodos se observa que la mayoría de los verbos admiten con frecuencia usos no pronominales. Parece normal que, tal como se ha indicado en el apartado 2, haya un cierto porcentaje de desviación con respecto a la norma general por motivos expresivos o por errores. Pero más allá de un 20% (porcentaje que se toma como umbral arbitrario) de casos de desviación por motivos expresivos (o errores de uso u otras circunstancias), se considera que no puede hablarse de forma tajante de «verbo pronominal estricto» (al menos si se contempla un corpus diacrónico de unos 500 años) en muchos de los casos objeto de estudio. Se muestran en la tabla 4 los mismos datos que los de la tabla 3, separando los datos en dos franjas porcentuales según este 20 % indicado.

Tabla 4. Número de verbos que aparecen en forma pronominal divididos en los rangos porcentuales que se indican en la primera columna. Se han dividido los porcentajes de la tabla 3 en dos bloques desiguales (80-100 % y 0,1-79 %) para mostrar que los verbos del primer bloque cumplen aproximadamente el rasgo de presentarse solo como pronominales (dejando de lado errores, variaciones expresivas y otras circunstancias del uso), mientras que los del segundo no.

	1500-1870		1871-1970		1971-2008	
Intervalos porcentuales	**n**	**%**	**n**	**%**	**n**	**%**
80-100 %	95	9,48	129	12,87	136	13,57
0,1-79 %	127	12,67	141	14,07	136	13,57
0,00 (sí OSLIN / No GBooks)	611	60,98	563	56,19	561	55,99
No OSLIN	169	16,87	169	16,87	169	16,87

Puede observarse que, en los tres períodos históricos, el porcentajes de verbos mayoritariamente pronominales (80-100 %) es menor que aquellos que pueden encontrarse también en estructuras transitivas o intransitivas no pronominales (0,1-79 %). Verbos que se encuentran en la primera franja mencionada son *abstenerse, cerciorarse, entrometerse, esmerarse, obstinarse, prosternarse, quejarse, retreparse* y otros. En cambio, *aburguesarse, desdibujarse, fosilizarse, gramaticalizarse,* etc., se encuentran entre los casos con numerosas concordancias no pronominales, como en estas frases de Google Books:

> La Revolución *aburguesó*, hasta cierto punto, la sociedad francesa.
>
> Estos dos universos *desdibujan* los límites de las distinciones culturales.
>
> ...aislamiento cultural de los españoles, que *fosilizó* la vida intelectual del país.
>
> Las lenguas codifican o *gramaticalizan* rasgos del contexto o evento del habla.

6.2. Porcentaje de verbos con usos en participio

En la tabla 3 se observa que existen verbos que los diccionarios categorizan como pronominales y que solo se han encontrado en su forma de participio, o que tienen numerosos casos de esta forma no personal. Son muy frecuentes los participios en casos como estos (tomados de Google Books):

En los delirios de mi mente *afiebrada*...

Una sonrisa maliciosa despuntó de sus labios *amoratados*.

Las capas *desclasadas* de nuestra sociedad...

Un gentío *endomingado* y devoto...

Resulta difícil o imposible averiguar la pronominalidad del verbo si este solamente se expresa como participio. Además, teniendo en cuenta especialmente los datos del período 1500-1870, si en esta franja temporal ya se encuentran casos de verbos que solamente se expresan en forma de participio, es razonable preguntarse si se trata de verbos defectivos, o simplemente de adjetivos con forma de participio, como es frecuente que ocurra[62].

6.3. Verbos no encontrados en el corpus

Es llamativa la gran cantidad de verbos de los que no se ha hallado ninguna concurrencia en el corpus[63]. Si nos limitamos a la tabla 4, se observa que más de la mitad de los verbos, en los tres períodos, no se han encontrado en ninguna de sus formas (se dejan de lado los 169 casos que el conjugador empleado no tenía). Se trata de casos como *achaplinarse, amezquindarse, chimpilinearse, desamotinarse, encaratularse, salmuerarse, trasconejarse* y muchos otros. Consideramos este el resultado más relevante en cuanto al análisis metalexicográfico.

7. Conclusiones y trabajo futuro

En este estudio se ha realizado un análisis de corpus de los 1.002 verbos que, en DRAE, DUE o DUEAE, aparecen lematizados en su forma pronominal y son considerados, por tanto, como estrictamente pronominales. Se había partido de un acercamiento previo a las tres obras y a un pequeño análisis manual de corpus, y ya se habían detectado divergencias de criterio, así como ausencia de datos empíricos. Los datos del análisis extraídos del corpus de enegramas de Google Books corroboran las intuiciones iniciales, formuladas concretamente en las dos hipótesis del apartado 3. Consideramos que:

- Se confirma la hipótesis 1, según la cual la gran variación en la representación lexicográfica es reflejo de variación en el uso. Muchos verbos no son realmente pronominales estrictos, sino que pueden entrar en construcciones no pronominales (transitivas o intransitivas), además del uso pronominal (generalmente intransitivo).
- Se confirma la hipótesis 2, según la cual verbos considerados estrictamente pronominales son más frecuentes en forma de participio o tienen concomitancias

62 El DUEAE, como cualquier otro diccionario, recoge muchos lemas que son formas en participio, no asociadas por tanto a ningún lema verbal: abuhardillado, alocado, disparatado, endemoniado, etc.

63 Se recuerda que Google Books no ha incluido en su corpus ninguna forma que haya aparecido menos de 40 veces.

con la estructura *estar* + *participio*. Muchos verbos solo se emplean en participio en esta forma, de modo que incluso queda en interrogante su estatus de verbos. Otros aparecen tan frecuentemente en esta forma que resulta difícil confirmar su naturaleza estrictamente pronominal.

Los motivos por los cuales existen estas divergencias en la representación lexicográfica sobrepasan los límites de este estudio, pero pueden estar relacionados con la escasa documentación de que se ha dispuesto hasta hace poco y con la metodología básicamente introspectiva con que se han redactado los diccionarios también hasta hace poco.

Teniendo en cuenta los datos recogidos, se extraen las siguientes conclusiones:

- La especialización morfológica que implica un verbo pronominal estricto es menos frecuente que lo esperado. Como final de un proceso de gramaticalización que a veces puede durar siglos (véase el apartado 2 para ejemplos de variación diacrónica), es corriente que no fragüe en una forma exclusivamente con pronombre, sino que pueda desglosarse en otras variantes diatéticas, que el propio idioma ya contempla como un modo de explotación de los verbos. Si se tiene *aburguesarse*, será probable que pueda existir *aburguesar*, con más o menos frecuencia de uso.
- Es necesario que el lemario de los diccionarios generales se ajuste a los datos empíricos, y no se alimente de datos que, a través de los siglos, han ido pasando de un diccionario a otro sin contrastarse con el uso real.
- El uso de corpus de grandes dimensiones arroja una nueva luz sobre el estudio de los fenómenos léxicos. En palabras de baja frecuencia, caso de muchos de los verbos estudiados, poder contar con un corpus que sobrepasa en gran medida las herramientas de que se disponía hasta ahora implica un cambio cualitativo de la investigación (aun basado en un aumento simplemente cuantitativo). La lexicología y uno de sus brazos aplicados, la lexicografía, deberían tener en cuenta la posibilidad de emplear estas herramientas de libre acceso.

Este trabajo no agota las posibilidades de análisis de los datos que se han presentado, y se considera que deben dejarse para estudios posteriores, entre otras, las siguientes cuestiones:

1. Averiguar, mediante labor documental, el origen de los lemas verbales que no se han hallado en el corpus.
2. Contrastar la semántica de estos verbos; por ejemplo, si los significados que aparecen en el diccionario son los que el corpus ofrece como más frecuentes.
3. En el mismo sentido, indagar si las formas no pronominales de los verbos tienen un significado distinto de las pronominales, o son variantes diatéticas.

Bibliografía

Battaner, P. (Dir.ª). (2003). *Diccionario de uso del español de América y España.* Barcelona: Spes Editorial.

Hanks, P. (2004). The Syntagmatics of Metaphor and Idioms. *International Journal of Lexicography, 17*(3), 245-274.

Hanks, P. (En prensa). *Lexical Analysis: Norms and Exploitations.* Massachusetts: MIT Press.

Janssen, M. Open Source Lexical Information Network. *Third International Workshop on Generative Approaches to the Lexicon*, Geneva, Switzerland.

Pomikálek, J., Rychlý, P., Kilgarriff, A. Scaling to Billion-plus Word Corpora. *Advances in Computational Linguistics*, 41, 3-13.

Moliner, M. (2001-2002). *Diccionario de uso del español*, 2.ª ed. Madrid: Gredos. Versión en CD-ROM.

Lorente, M. (2010). Verbs pronominals de font lèxica: reflexivitat i reciprocitat inherents. En I. Creus; M. Puig; J. R. Veny (Eds.), *Actes del Quinzè Col.loqui Internacional de Llengua i Literatura Catalanes* (pp. 435-445). Barcelona: Publicacions de l'Abadia de Montserrat.

Lorente, M. (en prensa). Verbs pronominals inherents: descripció i representació lexicogràfica. M. Á. Pradilla (Ed.), *Actes del III Col·loqui Internacional «La lingüística de Pompeu Fabra». Tarragona, desembre del 2008.* Barcelona: Institut d'Estudis Catalans.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., Aiden, E. L. Quantitative Analysis of Culture Using Millions of Digitized Books (2011). *Science 331*(6014), 176-182.

Otero, C. P. (1999). Pronombres reflexivos y recíprocos. En I. Bosque, V. Demonte, V. (Dirs.), *Gramática descriptiva de la lengua española* (pp. 1427-1517). Madrid: Espasa.

Real Academia Española (1726-1739). *Diccionario de autoridades*. Edición en línea en *Nuevo tesoro lexicográfico de la lengua española*: http://buscon.rae.es/ntlle/SrvltGUILoginNtlle (última consulta: 7.05.11).

Real Academia Española (2003). *Diccionario de la lengua española*, 22.ª ed. Madrid: Espasa. Versión en CD-ROM.

Real Academia Española (2009). Oraciones activas, pasivas, impersonales y medias. En *Nueva gramática de la lengua española* (pp. 3037-3112). Madrid: Espasa.

Renau, I. (En preparación). *Representación de los usos pronominales en los verbos de un diccionario de aprendizaje de español como lengua extranjera.* Tesis doctoral. Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona.

SÁNCHEZ LÓPEZ, C. (2002). Las construcciones con *se*. Estado de la cuestión. En C. Sánchez López (Ed.), *Las construcciones con «se»* (pp. 13-163). Madrid: Visor Libros.

Análisis del concepto 'habitación' en un corpus bilingüe español-inglés de páginas electrónicas de promoción hotelera

Julia Sanmartín Sáez

Universitat de València-Instituto Universitario de Lenguas Modernas Aplicadas

Nuria Edo Marzá

Universitat de València- Instituto Universitario de Lenguas Modernas Aplicadas

Resumen

El presente artículo posee como objetivo principal el análisis contextualizado de las unidades léxicas que correspondan al concepto "habitación" *a través de los datos obtenidos mediante la creación y explotación de un corpus* ad hoc *en dos lenguas de trabajo. Dicho corpus de estudio está formado por un conjunto de páginas electrónicas de hoteles de cinco y cuatro estrellas de España, y Chile (corpus de versión original en español y su traducción al inglés) e Inglaterra y EE.UU. (corpus de versión original en inglés y su traducción al español).*

A partir de dicho corpus se ha procedido al establecimiento de las posibles unidades léxicas especializadas que cubren el concepto "habitación" *(y sus tipos) en versiones traducidas y originales de las dos lenguas de trabajo.*

Palabras clave: *corpus, concepto, unidad léxica especializada, traducción, colocaciones*

Abstract

This paper is mainly aimed at analysing, from a contextual perspective, those lexical units corresponding to the concept "habitación"(room) *through the data obtained by the creation and exploitation of an* ad hoc *corpus in two work languages. Such corpus of study is made up of a series of four and five-star hotel web pages from Spain and Chile (original version corpus in Spanish together with its English translation) and from England and USA (original version corpus in English together with its Spanish translation).*

Departing from such corpus, the specialised lexical units covering the concept "room" *and its types, both in the original and translated versions and in both work languages, have tried to be retrieved and established.*

Keywords: *corpus, concept, specialised lexical unit, translation and collocations.*

1. Introducción: en el marco del Corpus Multilingüe de Turismo-Universitat de València (COMETVAL)

La presente investigación se sitúa en el marco del proyecto precompetitivo *Implementación y explotación léxica del corpus turístico multilingüe-Universitat de València (COMETVAL)*.[64] Dicho proyecto tenía como etapa inicial implementar con documentos procedentes del ámbito turístico en tres lenguas de referencia, español, francés e inglés una base de datos (véase figura 1) de discurso de este sector profesional[65], y poder, en una segunda fase, llevar a cabo un análisis lexicológico y elaborar glosarios multilingües.

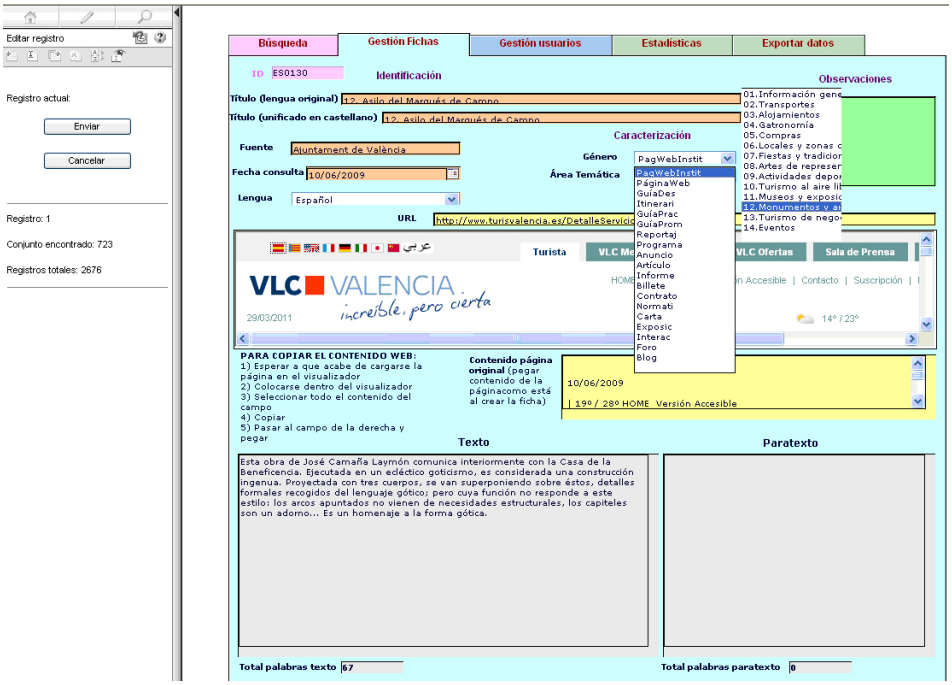


Figura 1. Base de datos inicial de COMETVAL.

Sin embargo, la base de datos inicial dio origen a una segunda base de datos, diseñada expresamente para confeccionar glosarios de la gestión hotelera (véase figura 2).

2. Metodología de la investigación: Hipótesis, corpus e instrumentos

En este marco de trabajo y a partir de esta segunda base de datos, se ubica la actual investigación, concebida como una especie de ensayo piloto que presenta como **objetivo**

64 Dicho proyecto fue concedido por la Universitat de València en abril de 2010, UV-AE-10-24407 Clave específica: 20100353.

65 Para elaborar dicha base de datos se tomó como punto de referencia la base de datos. *Enciclopedia de géneros turísticos,* elaborada por el Grupo Linguaturismo de la Universidad de Milán, dirigido por la profesora M. Vittoria Calvi. Dicha enciclopedia es de acceso restringido y no se puede consultar.

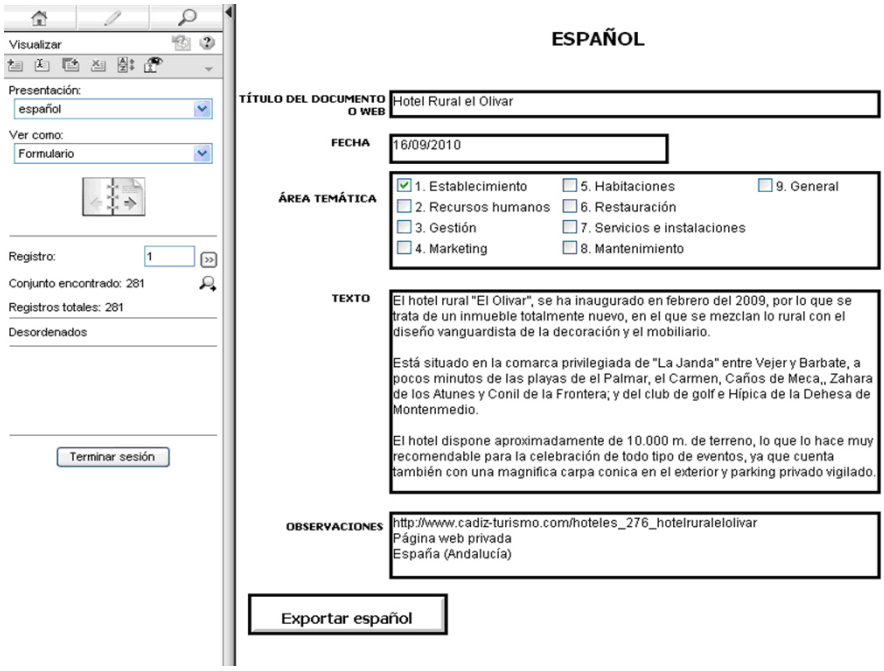


Figura 2. Base de datos de gestión hotelera.

la detección de las unidades léxicas específicas relativas al concepto de 'habitación' en un corpus[66] de páginas web de promoción y gestión de alojamientos hoteleros en dos lenguas, español e inglés. Además, se tienen en cuenta dos variables o parámetros que condicionan el objetivo inicial:

A) Se desea averiguar si estas unidades léxicas son distintas en algunas de las modalidades geográficas del español y del inglés. Por ello, los documentos vaciados procederán de dos variedades geográficas: España y Chile para el español, y Reino Unido y Estados Unidos para el inglés (véase Apéndice 1). Se parte de la **hipótesis inicial** de que las unidades de conocimiento de este ámbito son permeables a la variación geográfica y, por consiguiente, de que se encontrarán diferentes unidades y estructuras conceptuales en las distintas modalidades geográficas.[67]

B) Se pretende determinar si las versiones originales y traducidas de estos documentos coinciden en las unidades léxicas seleccionadas. En este sentido, los documentos seleccionados proceden tanto de textos redactados en su versión original como en su traducción.

66 De este modo, la investigación desarrollada se inserta en el marco epistemológico de la lingüística del corpus, entendida según Caravedo (1999: 19) como: "(…) toda orientación que, en la formulación y en el desarrollo de su programa de investigación (comprendidos la teoría y el sistema de corroboraciones o refutaciones desprendido de la actividad analítica), *depende* de la observación de un conjunto de datos extraído de la *producción real* de los individuos, y ordenados según criterios metodológicos diferentes pero *explícitos* de investigación."

67 En esta línea se sitúan los trabajos de Moreno Fernández (1999) o el de Freixà, Kostina y Cabré (2002), entre otros.

En relación con el objeto de estudio, se ha optado por acotar el corpus al tipo de discurso menos especializado del turismo, ya que estas páginas de promoción/gestión se consideran como textos de divulgación, esto es, de especialista a profano en la materia[68] (Calvi 2006). Para comprobar qué unidades léxicas presenta el concepto (*nodo, campo nocional*) 'habitación' en español e inglés se ha acotado un corpus no solo de "divulgación", sino que también se ha apostado por un corpus en línea, ya que como destacan diversos autores, "Internet se ha convertido en la mayor agencia de viajes del mundo" (Calvi 2006: 52). En concreto, se han compilado las páginas electrónicas privadas[69] de los establecimientos hoteleros de caso de cuatro y cinco estrellas por ser los que mayor diversidad de habitaciones establecen.

Una vez implementada la base de datos con el corpus seleccionado, se han utilizado los programas informáticos WordSmith y Anconc3.2. para obtener el listado de voces y las colocaciones de los términos que se consideran como pertenecientes al concepto 'habitación' de cada uno de los ocho microcorpus extraídos.[70]

La metodología empleada combina procedimientos onomasiológicos y semasiológicos: se establece un concepto 'habitación' y se determinan sus unidades léxicas a partir de los términos concretos (como *habitación*, *suite*, *individual*, *doble, fumador o discapacitado*, entre otros). Estas unidades, a su vez, permiten establecer tipos de habitaciones desde la ordenación conceptual.

3. El análisis del corpus: los resultados

La singularidad (*especificidad*) de las voces encontradas en el corpus de estudio reside en que se consideran como unidades denominativas porque así aparecen definidas en las normativas de turismo.[71] Veamos a continuación cuáles son esas voces.

3.1. El análisis de Wordlist

El análisis de frecuencia ofrecido por Wordlist (véanse tablas 3 y 4), parece confirmar que el corpus responde a los objetivos, ya que las palabras *habitación* (*habitaciones*) / *room* (*rooms*) y *suite (suites)*, aparecen en todos los microcorpus en las primeras posiciones.

68 La selección obedece a que en posteriores trabajos se desea comprobar la idoneidad o adecuación de las unidades léxicas de este discurso turístico desde la perspectiva del consumidor.

69 En principio, también se deseaba comparar la información de las páginas institucionales y de las privadas; sin embargo, la información que aparece en las institucionales de España se reduce a la ficha técnica o a la descripción mediante símbolos, como sucede en la Comunidad de Galicia o en la Comunidad Valenciana. Apenas hay información sobre las habitaciones y, por ello, quedó descartada como variable.

A partir de los datos del corpus cabría tener presente un nuevo factor como elemento importante en análisis posteriores: la cadena hotelera a la que pertenece el establecimiento turístico.

70 Los microcorpus recogen las páginas electrónicas del español de España (1), del español de Chile (2), del español traducido de Estados Unidos (3), del español traducido de Reino Unidos (4), del inglés de Estados Unidos (5), del inglés de Reino Unidos (6), del inglés traducido de España (7) y del inglés traducido de Chile (8).

71 Por supuesto, estas voces también se registran en textos especializados de turismo, como pueden ser los manuales (de carácter didáctico: de especialista a especialista en formación), como el de Blasco (coord.) (2006).

Tabla 3. Listados de frecuencia en español con las 20 unidades más frecuentes de cada microcorpus.

Español España

N	Word	Freq.
1.	HABITACIONES	292
2.	HABITACIÓN	202
3.	HOTEL	198
4.	BAÑO	149
5.	SERVICIO	97
6.	CIUDAD	91
7.	SUITE	90
8.	GRATUITO	89
9.	SERVICIO	89
10.	AC	88
11.	MURCIA	86
12.	CAMA	85
13.	SERVICIOS	81
14.	TV	76
15.	VALENCIA	74
16.	CENTRO	72
17.	COMPLETO	72
18.	ACONDICIONADO	69
19.	VISTAS	69
20.	INTERNET	66

Español Chile

N	Word	Freq.
1.	HOTEL	125
2.	HABITACIONES	104
3.	SANTIAGO	80
4.	UBICADO	64
5.	METRO	54
6.	CENTRO	40
7.	ESTACIÓN	36
8.	SERVICIO	30
9.	SUITES	30
10.	BARRIO	29
11.	PASOS	29
12.	AIRE	28
13.	CIUDAD	28
14.	BAR	27
15.	INTERNET	25
16.	SINGLE	25
17.	ACCESO	24
18.	DOBLE	24
19.	PISCINA	24
20.	PLAZA	24

Español traducido Reino Unido

N	Word	Freq.
1.	HABITACIONES	125
2.	SERVICIO	122
3.	INTERNET	115
4.	HABITACIÓN	108
5.	CAMA	73
6.	PETICIÓN	62
7.	DISPONIBLES	61
8.	CAMA	59
9.	BAÑO	55
10.	ACCESO	54
11.	DETALLES	48
12.	TÉRMINOS	45
13.	BED	41
14.	CLUB	40
15.	LOUNGE	37
16.	LEVEL	35
17.	FUMADORES	34
18.	ALTA	33
19.	SLEEPER	33
20.	SWEET	33

Español traducido EEUU

N	Word	Freq.
1.	USD	313
2.	HABITACIÓN	211
3.	INTERNET	206
4.	DETALLES	172
5.	TÉRMINOS	161
6.	RESERVAR	157
7.	FUMADORES	146
8.	HABITACIONES	144
9.	GASTOS	141
10.	IMPUESTOS	141
11.	KING	125
12.	CAMA	124
13.	SUITE	89
14.	TELEVISOR	84
15.	BAÑO	83
16.	CORTESÍA	80
17.	CAMA	78
18.	VISTA	76
19.	TARIFA	75
20.	ALTA	73

En el caso del español (en versión original) *hotel, habitación, suite, ciudad, centro, y servicio* aparecen como términos recurrentes en España y Chile; en cambio, la situación difiere en voces como *vistas*, que aparece en España; o *ubicado, metro, estación, acceso, plaza, bar, pasos* o *piscina*, que se registra solo en Chile, donde la ubicación aparece como un elemento esencial.

Diferencias similares se aprecian en las traducciones del español, en las que se documentan, como singulares, términos relativos a la gestión de reservas o a los detalles de cortesía que se ofrecen en los hoteles: *gastos, impuestos, tarifa alta* o *detalles y cortesía,* en el español de Nueva York; y *alta, términos y detalles,* en el español de Londres. En este último corpus, además, se detectan bastantes anglicismos.

Tabla 4. Listados de frecuencia en inglés con las 20 unidades más frecuentes de cada microcorpus.

Inglés original Reino Unido

N	Word	Freq.
1.	ROOM	187
2.	SERVICE	127
3.	ROOMS	113
4.	INTERNET	109
5.	BED	107
6.	AVAILABLE	99
7.	ACCESS	93
8.	REQUEST	78
9.	SMOKING	55
10.	LOUNGE	52
11.	HOTEL	48
12.	AMENITIES	47
13.	SUITE	46
14.	DETAILS	43
15.	SERVICES	43
16.	TELEVISION	43
17.	TERMS	43
18.	SPEED	42
19.	CLUB	41
20.	FREE	40

Inglés original EEUU

N	Word	Freq.
1.	USD	567
2.	ROOM	401
3.	INTERNET	241
4.	BED	233
5.	DETAILS	219
6.	ACCESS	201
7.	KING	183
8.	ROOMS	167
9.	SMOKING	162
10.	CHARGES	161
11.	TERMS	161
12.	RATE	151
13.	BOOK	150
14.	SUITE	143
15.	TAXES	141
16.	AMENITIES	137
17.	CITY	129
18.	VIEW	122
19.	SPEED	116
20.	NIGHT	115

Inglés traducido España

N	Word	Freq.
1.	ROOM	355
2.	ROOMS	258
3.	HOTEL	254
4.	SERVICE	173
5.	SUITE	120
6.	BATHROOM	95
7.	BED	95
8.	AC	92
9.	GUESTS	91
10.	SERVICES	88
11.	INTERNET	87
12.	TV	84
13.	DOUBLE	82
14.	FACILITIES	81
15.	ACCESS	78
16.	VALENCIA	78
17.	EQUIPPED	77
18.	CITY	73
19.	BAR	69
20.	AIR	67

Inglés traducido Chile

N	Word	Freq.
1.	HOTEL	93
2.	ROOMS	62
3.	SANTIAGO	52
4.	BUSINESS	29
5.	CENTER	25
6.	BAR	22
7.	SERVICE	22
8.	EQUIPPED	19
9.	POOL	19
10.	TV	18
11.	ROOM	17
12.	CABLE	16
13.	AIR	15
14.	PLAZA	15
15.	CONDITIONING	14
16.	DOUBLE	14
17.	INTERNET	14
18.	TELEPHONE	14
19.	SINGLE	13
20.	SUITES	13

En el caso del inglés, los microcorpus parecen también coherentes y representativos, es decir, bien compilados, a juzgar por el tipo de unidades altamente frecuentes que se encuentran y entre las cuales figuran en todos los casos *room* y *suite* en singular y/o plural. Se observa también la presencia de una serie de términos recurrentes en todos o casi todos los microcorpus como pueden ser las unidades *service/services*, *bed* (unidad léxica que adquiere un papel fundamental en contextos de habla inglesa a la hora de indicar la capacidad de las habitaciones) e *internet*, *TV/televisión* y *telephone*, poniendo de manifiesto la importancia concedida a la tecnología, y la coincidencia con los términos recurrentes del español.

Entre las singularidades de cada microcorpus, cabe señalar de nuevo la importancia de los términos relacionados con la gestión del alojamiento y los contratos en general muy especialmente en las versiones originales en inglés (como también se ha comentado respecto a sus versiones traducidas al español), siendo esta tendencia especialmente destacada o notoria en el caso del inglés estadounidense con términos del tipo: *USD, charges, terms, taxes* y *rate.*

Además de estas dos unidades, *habitación* / *room* y *suite,* hemos buscado en los listados de palabras todos aquellos términos susceptibles de integrar unidades denominativas simples o compuestas (habitación + modificante), ya que en los cotextos concretos puede aparecer, por ejemplo, el término *habitación individual* o simplemente *individual*. Se pueden considerar como variantes denominativas (composición sintagmática, en el primer caso; conversión categorial en el segundo):

-en el caso del español los términos encontrados han sido: *habitación, suite, individual, single uso, doble, triple, cuádruple, familiar, matrimonial, twin, room fumador, discapacitado, estándar, ejecutiva, superior, premium, contigua, comunicada*, etc.

-en el caso del inglés: *room, suite, individual, single, double, triple, X-bedded, familiar, matrimonial, twin, smoking, non-smoking, disabled, disability, standard, executive, superior, premium, adjoining, connecting, etc.*

3.2. El análisis de las agrupaciones recurrentes: unidades polilexemáticas y colocaciones

De todos los términos anteriores, hemos detectado los patrones recurrentes de construcciones sintácticas, los *cluster* (colocaciones repetidas). El análisis de estos patrones y de los *wordlist* de cada microcorpus nos permite realizar una posible clasificación conceptual de los tipos de habitaciones, una clasificación que debe caracterizarse como *relación abstracta, lógica o genérica vertical* (de género o 'tipos de'), con una polijerarquía, ya que en cada tipo de habitación se pueden establecer a su vez nuevas distinciones con criterios heterogéneos (Arntz y Picht 1995: 108-113). El posible *sistema*, considerado a modo de representación de la estructura conceptual es el siguiente (figura 5):

Estructuración conceptual propuesta

(con un término en español y en inglés como muestra)

1. Habitaciones según el número (y tipo) de personas que se alojan:

-para una persona

- con capacidad para uno: *habitación individual / single*
- con capacidad para dos: *doble de uso individual / double room for single use*

-para dos personas:

- con una cama: *doble, matrimonial / double-bedded*
- con dos camas: *habitación twin / twin room*

-para tres personas:

- tres adultos (o dos adultos y un niño): *habitación triple / triple-bedded room*
- dos adultos y un niño: *familiar/ familiar room*

-para cuatro personas:

- cuatro adultos (o dos adultos con dos niños): *habitación cuádruple*
- dos adultos y dos niños: *habitación familiar / familiar room*

2. Habitaciones según el tipo de espacios habitables:

-habitación sin salón: *habitació /, room*

-habitación con salón:

- salón independiente del dormitorio: *suite / suite*
- salón integrado en el dormitorio: *junior suite / junior suite*

3. Habitaciones especiales, según los usuarios:

- fumadores o no fumadores: *habitación para fumadores / Non smoking room*
- para personas con discapacidad: *habitación adaptada para discapacitados / Disabled accesible room*

4. Habitaciones según la ubicación en el hotel y su conexión:

- comunicada: *habitaciones comunicadas / Connecting rooms*

5. Habitaciones según una escala de comodidades-servicios y un posible usuario:

-habitación normal: *estándar/ standard*

-habitación de más lujo: *superior, premium, deLuxe, upper / superior, premium, deLuxe, upper*

-habitaciones según estilo: *classic, tradicional / classic, traditional*

-habitaciones con servicios para personas de negocios: *habitación ejecutiva / executive*

Figura 5. Propuesta de estructuración conceptual del concepto habitación.

Esta tipología se ha ampliado en forma de cuatro tablas detalladas (véanse tablas 6, 7, 8 y 9) con los resultados obtenidos del corpus, en los que figura el número de recurrencias entre paréntesis.

Tabla 6. Unidades léxicas detectadas en el corpus relativas al concepto capacidad.

CAPACIDAD								
	Orig. Español españa	**Orig. Español Chile**	**Trad. Español Usa**	**Trad. Español Uk**	**Orig. Inglés Usa**	**Orig. Inglés Uk**	**Trad. Inglés España**	**Trad. Inglés Chile**
Dos personas -con 1 cama -con 2 camas	Habitación doble (44) / habitaciones dobles (10) Dobles (2) Habitación twin (7)	Doble (24) Habitación matrimonial (1)	(camas) Dobles (22) Habitación doble (5)	Habitación doble (7) (camas) dobles (16)	Double-bedded rooms (16)	Double room (10)	Double room (33)/rooms (9) Twin room (9)/ rooms (1) Room for 1 or 2 guests (5) Matrimonial room (2)	Double (10) Superior double (3)
Una persona	Habitación individual (7) Para 1 persona (7) Single (1)	Habitación individual (1) Single (16) Habitación single (1)					Single room (4)/rooms(1)	Single (11) Superior single (2)
Dos personas (uso individual)	Doble uso single (2) Doble uso individual (8)						Double room individual use (1) Double room for single use (3) Double room for 1 person (2)	
Tres personas	Habitación triple (4) / triples (1)	Triple (19)					Triple bedded room (1) Triple room (4)/rooms (2) Double room for 3 adults (1)	Triple (9)
4 personas		Cuádruple (11)			Four confort bed...room (24)	Four confort bed...room (7)		
Dos personas + dos niños	Family room (2)	Habitación familiar (1)		Habitaciones familiares (13)		Family rooms (13)		

Tabla 7. Unidades léxicas detectadas en el corpus relativas al concepto habilitada especial.

HABILITADA ESPECIAL								
	Orig. Español España	**Orig. Español Chile**	**Trad. Español USA**	**Trad. Español UK**	**Orig. Inglés USA**	**Orig. Inglés UK**	**Trad. Inglés España**	**Trad. Inglés Chile**
FUMADOR	Habitación (para) fumadores (7)	Fumadores (3) (Habitaciones) opción fumador / no fumador (5)			Smoking room (12)		Smoking room (3)	
NO FUMADOR	Habitación para no fumadores (1)	(Habitaciones) opción fumador / no fumador (5)	Habitaciones para no fumadores (11)	Habitaciones Para no fumadores (51)	Non smoking room (52) 100% Non-Smoking Guestrooms (3) Non-smoking room (6) Smoke-free guest rooms (4)	Non-Smoking Room (4)	Non smoking room (3)	Non smoking rooms (1)
DISCAPACIT.	Habitación (acceso) minusválidos (1) Adpatadas para minusválidos Adaptadas para discpacitados (2) Habilitadas para discapacitados (2) Adaptadas para huéspedes con movilidad reducida (2) Adpatadas para personas com movilidad reducida (2) Hab.Accesibles (1)/ habitación doble accesible (1)	Habitaciones (para) (especiales) discapacitado (6) / personas con discapacidad física (3)	Habitación para personas con discapacidades (6)	Habitaciones para discapacitados (3)	Accesible room (5)/rooms (1) Disabled accessible room (1)/rooms (4) Disability accessible room (4)	Disability Accessible Room (3) Disabled Accessible Room (1)	Disability Accessible Room (1)	
COMUNICADAS	Hab.Comunicadas (1) Habitaciones conectadas (4)	Habitaciones intercomunicadas (4)	Habitaciones comunicadas (3)	Habitaciones comunicadas (17)	Connecting rooms (4)/room (1)	Connecting rooms (16)	Connecting Rooms (1) Communicating rooms (4)	Connecting rooms (1)

Tabla 8. Unidades léxicas detectadas en el corpus relativas al concepto referente individualizado.

REFERENTE INDIVIDUALIZADO								
	Orig. Español España	**Orig. Español Chile**	**Trad. Español USA**	**Trad. Español UK**	**Orig. Inglés USA**	**Orig. Inglés UK**	**Trad. Inglés España**	**Trad. Inglés Chile**
ESTÁNDAR	Habitaciones Estándar (8) / 5 En Plural Habitaciones Standard (2)	Habitación Estándar (1) /Standar (2)	Habitación Standard (2) / Estandar (2)	Habitación Standard (1)	Standard room (3)/rooms (2)	Standard room (3)/ rooms(2)	Standard room (6)/rooms (12) Standard double room (5)	
SUPERIOR	Habitación Superior (8) / 5 En Plural	Habitación Superior (Doble/ Single) (11)	Habitación Superior (11)	Habitación Superior (1)	Superior room (9) / rooms (7)	Superior room (4)/rooms (1)		
PREMIUM /PREMIER/GRAN(D) PREMIUM	Habitación Premium (8) Habitación Gran Premiun (2)				Premier room (7) Tower Premier Rooms (3)		Premium Rooms(5)/room (1)	
EJECUTIVA	Habitación Ejecutiva (6) / habitaciones Ejecutivas (4)	Habitaciones Ejecutivas (5)	Habitación Ejecutiva (3)	Habitación Ejecutiva (2) / Executiva(2)	Executive rooms (14)	Executive room	Executive room (4)/rooms(5)	Executive rooms (2)
DELUXE/LUXE/ LUXURY/GRAND LUXE	Habitación Deluxe (10)	Hab.De Luxe (1)/ Deluxe (2)	Habitación Luxe (19) Habitación Grandluxe (11)		Deluxe room (10) / rooms (4)		Deluxe room (5) / rooms (7)	Luxury rooms (2)
UPPER	Habitación Upper (3) Upper Room (15)							
CLASSIC / TRADICIONAL	Habitación Classic (2)		Habitación Tradicional (3)	Habitaciones Classic (4) Habitaciones Clásicas (2)	Traditional room (4)/rooms (1)		Classic room (1)/rooms (1)	

Tabla 9. Unidades léxicas detectadas en el corpus relativas al concepto suite.

SUITE								
	Orig. Español España	Orig. Español Chile	Trad. Español USA	Trad. Español UK	Orig. Inglés USA	Orig. Inglés UK	Trad. Inglés España	Trad. Inglés Chile
CAPACIDAD MÁXIMA					2 double-beds suite (2)			
NÚMERO DE DEPENDENCIAS	Junior suite (6)	Junior suite (2)	Junior suite (8)	Junior suite (7)	Junior suite (9)/suites (8) Two-bedroom suite (3) / two-bedroom corner suites (2) One-bedroom suite (12) Triplex suite (11) Three bedroom specialty suites (4)	Junior suite (11)	Junior suite room (1) Junior suite (24) /suites (17)	Junior suite (2) /suites (2)
CATEGORÍA Y PERFIL DE USUARIO	Upper suite (5) Suite presidencial (9) Suite real (5)			Suite presidencial (2) Máster suite (3)	Premier junior suite (2)/suites (7) Presidential suite (3) Royal suite (6) Master suite (1) Imperial suite (2) Grand suite (2)	Master suite (5)	Executive suite (1)	

A partir de la tipología presentada en la figura 5 y de los datos obtenidos y analizados en los listados de palabras y en las tablas anteriores (6-9), se extraen los siguientes resultados:

1. No se pueden reconocer como unidades léxicas especializadas aquellas denominaciones casi más cercanas al nombre propio que al común, esto es, muchas de las unidades clasificadas como si fueran tipo de unidades responden a una especie de etiqueta creada por el establecimiento hotelero para singularizar su oferta y convertirla en más atractiva para el usuario. En este sentido se encuentran desde las etiquetas que incluyen el nombre del hotel (*habitaciones AC* o *habitaciones AH*) hasta compuestos que incorporan nombres propios *Suite Malvarrosa*, *Suite Las Arenas* y *Starwood Suite*[72].

72 Además, encontramos muchos referentes individualizados totalmente personalizados y rozando la categoría de nombre propio:

-Español de España: *Red Level* (22), *Habitación AH* (6), *Habitación AC* (18)

-Traducción español de Estados Unidos: *Habitación Spectacular* (10), *Habitación cool* (7), *Habitación Mega* (7), *Habitación Wonderful* (6), *Habitación Fabulous* (5), *Habitación Sheraton Club* (4), *Habitación Atrium Club* (3), *Habitación Starwood* (2), *Habitación Tower* (5)

-Traducción español de Reino Unido: *Habitación Starwood* (1)

-Inglés de Estados Unidos: *Mega room* (6), *Spectacular room* (7), *Cool room* (6), *Wonderful room* (6), *Fabulous room* (4), *Sheraton Club room* (4), *Atrium Club room* (3), *Starwood room* (2), *Tower room* (5)

-Inglés de Reino Unidos: *Tower room* (6)/*rooms* (12), *Atrium Club Rooms* (3), *Spa-Inspired Room* (1)/*rooms* (1), *Skyline View Room* (2), *River/City View Room* (2)

2. Otras estructuras fijas recurrentes responden al esquema de sustantivo + adjetivo / adjetivo + sustantivo, a modo de sintagma nominal libre (*habitaciones confortables, stylish rooms*) sin que sea conveniente juzgarlas propiamente como unidades denominativas.[73] Se caracterizan como meras *colocaciones*, cuyo estudio también es pertinente para describir en su totalidad este tipo de léxico y sus variables en cada uno de los microcorpus, ya que muestra rasgos descriptivos distintos en cada variante geográfica e indica que en la promoción son rasgos importantes 'equipamiento', 'comodidad' y la 'amplitud' en España y en Chile, añadiéndose, además, en España el rasgo 'diseño'. En inglés ocurre algo similar ya que se priman la apariencia y características estéticas así como el confort de la habitación y las peculiaridades que la hacen especial. Curiosamente, en el caso del inglés de Estados Unidos se aprecia una marcada tendencia a rasgos más valorativos que estrictamente descriptivos: *fabulous room* o *wonderful room*. Con todo, es obvio que el uso más o menos profuso de adjetivos, la descripción más o menos exagerada de los aspectos positivos de un destino y la calidad de la información turística en general tendrán siempre un eminente sentido subjetivo, pues son muchas las variables que los condicionan y variadas son también las posibles preferencias del turista potencial. Así pues, se encuentran colocaciones y adjetivos que apelan a aspectos estético-sensoriales y hacen referencia al carácter extraordinario del elemento descrito o a su originalidad y/o exclusividad entre muchos otros, tal y como se constata en los patrones sintagmáticos señalados (Edo, e.p.).

-Español de Chile: *habitaciones equipadas (15), habitaciones cómodas (6), habitaciones confortables (5), habitaciones amplias (5)*

-Español de España: *habitaciones diseñadas (7), habitaciones equipadas (6), amplia/s habitaciones (6), habitaciones confortables (3), habitaciones decoradas (3), exclusivas habitaciones (3), espaciosa/s habitaciones (3)*

-Español de Estados Unidos: *lujosas (para habitaciones, amenidades, almohadas)*

-Inglés de Estados Unidos: *Fabulous room (6), Wonderful room (4), Spectacular room (3), fantastic suite* (25), *decorated suite* (2)

-Inglés de Londrés: *Stylish room (4), Available room (4), Wonderful suite (3), Stylish suite (2),* panoramic rooms (2), delicious room (2)

-Inglés de España: *Spacious room (8), Elegant room (2), Exclusive room* (4), *Bright room* (2), *Designed rooms* (3), *Comfortable rooms* (3), *Equipped rooms* (2), *Exclusive rooms* (2), *Selected rooms* (2), *Distinctive suites* (2), *Gorgeous suite* (2)

-Inglés de Chile: *spacious room* (3), *comfortable room* (4), *special rooms* (2)

3. Entre las denominaciones consideradas como cercanas al nombre propio y las meras colocaciones, se aprecian las denominaciones que jerarquizan las habitaciones en función

73 La frontera entre léxico y sintaxis es difusa (Val de Álvaro 1999: 4763-4765). Determinar la existencia de un concepto unitario o no supone un arduo problema: en este discurso de especialidad, los conceptos quedan establecidos de un modo explícito por la descripción de rasgos que aparecen en las definiciones de las normativas. Veánse sobre esta cuestión Piera y Varela (1999) o Ruiz Gurillo (2001).

del lujo (*estándar* / *superior, premium, deLuxe, upper),* el estilo *(classic, tradicional)* o las condiciones para el usuario *(executive*). Estas denominaciones responden a un deseo de connotar positivamente las habitaciones; sin embargo, no se registran en las normativas, de turismo correspondientes como un tipo de habitación, por lo que su concepto y definición queda pendiente de las inferencias del cotexto más inmediato, el cual cambia de un hotel a otro.

4. También cabe explicar la complejidad de establecer o delimitar estas unidades léxicas, cuyos rasgos se van combinando unos con otro, así por ejemplo, las habitaciones según los servicios complementan en muchas ocasiones las habitaciones según la capacidad o los espacios: *habitación doble superior* en el caso del español. Esto también se aprecia en *double room*, ya que en inglés se suelen anteponer los adjetivos, *superior* o *executive.* Es más, cabe reconocer que esta tendencia es especialmente importante en inglés. Esta lengua, además, presenta una diferencia de estructura informativa respecto al español, ya que el inglés suelen presentar la información de un modo más telegráfico y descriptivo: *2 Oversized Single Bed Non-smoking Superior Room* / *1 King + 1 Queen Sofa Bed Junior Suite*

5. No cabe duda de que es evidente la variedad de denominaciones en el ámbito de esta especialidad, una tendencia muy contraria a la pretendida univocidad buscada en los discursos especializados con su consiguiente monosemia y ausencia de sinonimia o connotación. En este caso, conviven diversos tipos de sinónimos con una frecuencia de aparición dispar:

-meras variantes gráficas (*habitación estándar* / *habitación standar*)

-unidades simples o compuestas (*habitación doble* / *doble*) (*room/ bedroom/guestroom*)

-término patrimonial frente a anglicismo (*individual* / *single*)

-distintas unidades compuestas (*habitación individual* / *habitación para una persona smoking room / double-bedded room*)

Además, algunos de estos sinónimos presentan una clara tendencia a asociarse a una variedad geográfica.[74] Así, el término *single* se utiliza en el español de Chile, e *individual* en el español de España. Del mismo modo, encontramos conceptos que presentan más variedad que otros. Esto sucede respecto al concepto de 'habitación para personas discapacitadas' en el español de España, donde aparecen 7 términos distintos*, habitación (acceso) minusválidos, habitación adaptada para minusválidos, habitación adaptada para discapacitados, habitación habilitadas para discapacitados, habitaciones adaptadas para huéspedes con movilidad reducidas, habitación adaptada para personas con movilidad reducida, habitaciones accesible*. Lo mismo ocurre con ese mismo concepto para el inglés de Estados Unidos, con 3 términos distintos (*accesible room, disabled accessible room* y *disability accessible room*) o con el concepto 'habitación para

74 No obstante, se ha detectado mayor variación entre las denominaciones de los establecimientos hoteleros del español de España y de Chile (Sanmartín e.p.a) que entre las denominaciones de las habitaciones, lo cual, en cierto modo, era esperable.

no fumador' también con 3 términos distintos: *non smoking room, 100% non-smoking guestrooms* y *smoke-free guest rooms*.

6. En el contraste entre lenguas y modalidades, se observa, por ejemplo, que determinados conceptos solo aparecen en alguna variedad, como sucede con el español de España, en el que se reconoce el uso individual de una habitación doble: *doble uso single, doble uso individual,* o en sus traducciones: *double room for single use, double room individual use.* Por su parte, el término *family room* aparece de forma predominante en el inglés de Reino Unido.

En esta línea, sorprende que en inglés, en lo referente a capacidad, los términos *single room* o *individual room* no aparezcan como tal en el corpus sino que *single* e *individual* suelen complementar en primera instancia a las unidades *bed* o *sofá (2 Oversized Single Bed Non-smoking Superior Room),* por lo que parece deducirse que la tendencia del inglés es a indicar la capacidad de las habitaciones en función del tipo de cama que encontramos en ellas, tal y como parece indicar también la profusión de la estructura *X-bedded rooms* (*double-bedded rooms*). Eso no sucede. sin embargo. en español, donde los términos *habitación doble* o *habitación individual* son frecuentes. De hecho, las páginas en inglés se suelen organizar en torno a tres aspectos: a) tipo de cama, b) tarifas, c) características de la habitación.

7. Por último, cabe destacar la importancia del cotexto, que no hemos podido reproducir en estas páginas, como un factor determinante para la delimitación de las estructuras conceptuales, ya que la descripción que se ofrece permite inferir una serie de rasgos o propiedades que configuran el concepto, como por ejemplo la diferencia entre *junior suite* y *executive suite/ suite ejecutiva*.

4. Conclusiones

El corpus, definido como un conjunto representativo de datos lingüísticos, y los programas informáticos de extracción de términos y colocaciones permiten comprobar los usos reales, las denominaciones, y las posibles estructuras conceptuales de los dominios especializados. No obstante, la selección correcta del corpus, en función de los objetivos planteados, resulta fundamental para valorar los resultados. En este sentido, consideramos que la estructura conceptual analizada en la investigación llevada a cabo a modo de ensayo piloto podría ser perfilada de otro modo si en el corpus se incorporan más tipos de establecimientos hoteleros y un corpus más amplio.

Sin embargo, los datos extraídos confirman la hipótesis de partida, la existencia de cierta variedad denominativa en el español de las páginas electrónicas de promoción; una variedad que no se ajusta del todo a las normativas.

Por ello, planteamos, desde la lingüística aplicada, el uso de los datos obtenidos en la investigación no solo para conocer desde el punto de vista teórico la configuración de estos

campos nocionales, sino para aprovechar los conocimientos en la mejora de los propios servicios turísticos. Este último puente entre la lingüística y la sociedad es el eslabón que nos falta transitar. Se trataría de redactar de un modo sintético las divergencias entre norma y uso con el objetivo de que las autoridades competentes adopten las medidas, si tal y como indican en las normativas desean realmente "garantizar" los derechos de los usuarios de los servicios turísticos al reservar o contratar las habitaciones.

5. Apéndice

Apéndice 1

Relación de documentos del corpus

Versión original en español y traducción al inglés

ESPAÑA

Se han tomado como corpus las páginas electrónicas privadas de al menos 10 hoteles de tres Comunidades Autónomas (Galicia, Andalucía y Comunidad Valenciana). La traducción del español al inglés es sistemática en casi todas las páginas electrónicas.

Andalucía

(5 y 4 estrellas páginas privadas y página institucional)

-privadas: Barceló Renacimiento, Hotel Cortijo Soto Real, Hotel Gran Meliá Colón, Hotel Hacienda la Boticaria, AC Palacio de Santa Paula, Hoteles Abades Guadix, AC Ciudad de Sevilla, Abba Triana Hotel, AH Granada Palace, Hotel Alay

-Institucional http://www.andalucia.org/alojamientos/ (ficha técnica)

Barceló Renacimiento, Hotel Cortijo Soto Real, Hotel Gran Meliá Colón, Hotel Hacienda la Boticaria, AC Palacio de Santa Paula, Hoteles Abades Guadix, Hotel Abades Nevada Palace, Hotel Abba Granada, Hotel Abba Triana Hotel, Hotel AC Huelva, Hotel, AC Ciudad de Sevilla, Hotel Acinipo Ronda, Hotel AGH Estepona, Hotel Alcázar, Hotel AH Granada Palace, Hotel NH Avenida de Jerez de la Frontera, Hotel NH Central Convenciones Sevilla, Hotel Alay.

Galicia

(5 y 4 estrellas, páginas privadas)

Gran hotel la toja , Hesperia Finisterre, Meliá Araguaney, AC Palacio del Carmen, Gran Hotel Nagari, Parador Hostal dos Reis Católicos, Hotel Bahía de Vigo, Augusta Spa Resort, Hotel Compostela, Hotel Congreso, Hotel Abeiras.

Comunidad Valenciana

(5 y 4 estrellas, páginas privadas)

Hospes Palau de la Mar-Valencia, Hotel Las Arenas Balneario Resort, Hotel The Westin Valencia, SH Valencia Palace, Hotel La Calderona

Hotel AC Valencia, Ayre Hotel Astoria Palace, Hotel Barceló Valencia, Beatriz Rey Don Jaime

CHILE

Se han vaciado tres páginas privadas de Chile, consideradas por ellas mismas como una especie de "guía de viajes on line", en la que se incluye una pestaña relacionada con el alojamiento. En esta pestaña se describe el hotel. En cierto modo se podrían considerar como los sustitutos de las actuales agencias de viajes, ya que estas páginas describen los lugares que se van a visitar (a modo de guía o folleto), se proponen rutas y actividades, se indican los detalles prácticos y se puede gestionar, además, la reserva del alojamiento. También se ha vaciado un folleto de alojamientos de Sernatur.

-RutaChile.com http://www.rutaschile.com/

Hotel Crowne Plaza, Hotel Grand Hyatt Santiago, Hotel Plaza San Francisco, Hotel Kennedy, Hotel Marriott, Hotel Caesar Business Santiago, Hotel Gen Suite Santiago, Hotel Fundador, Hotel Le Reve, Hotel Atton Las Condes

-**Santiago de Chile.com** http://www.santiagodechile.com/

Hotel Sheraton Santiago, Hotel NH Ciudad de Santiago, Regal Pacific Hotel, Hotel Crowne Plaza Santiago, Hotel Kennedy, Hotel Atton Las Condes, Hotel Four Points, Hotel Rugendas, Hotel Neruda, Hotel Director

-**Welcomechile.com** http://www.welcomechile.com/

Boulevard Suites Hotel Boutique, Hotel InterContinental Santiago, Plaza el Bosque Park & Suites, Sheraton San Cristobal Tower, W Santiago, Atton El bosque, Neruda Hotel, Hotel Caesar Business, Hotel Eurotel, Vespucci Suites Hotel

-**Sernatur** (Servicio Nacional de Turismo de Chile)

PDF Alojamientos en Santiago de Chile

Figura A1. Captura de pantalla de un hotel chileno incluido en el corpus.

Versión original en inglés y traducción al español

Hoteles de 5 y 4 estrellas de páginas privadas de establecimientos hoteleros En el caso del inglés, se han seleccionado dos ciudades de dos países para comprobar la variedad geolectal, Londres (Reino Unido) y Nueva York (Estados Unidos): se han analizado diez páginas electrónicas de hoteles de las categorías mencionadas de cada una de estas ciudades. El escollo fundamental de la extracción del corpus ha sido la obtención de páginas electrónicas traducidas al español.

Reino Unido (UK) - Londres

Hesperia London Victoria, NH Harrington Hall , NH Kensington, Melià White House Hotel, Sheraton Park Tower, The park lane hotel, Sheraton Skyline hotel London Heathrow, Four points by Sheraton London

ESTADOS UNIDOS (UK) –NUEVA YORK

Jolly Madison Towers (NY), New York Palace, W New York, The Westin New York at Times Square, The Manhattan at Times Square Hotel, Four Points by Sheraton, Midtown - Times Square, Four Points by Sheraton Manhattan Chelsea, Le Parker Méridien New York, The St. Regis New York, W New York – Union Square, Sheraton Lincoln Harbor Hotel, Sheraton Tribeca New York Hotel

Figura A2. Captura de pantalla de un hotel estadounidense incluido en el corpus.

6. REFERENCIAS BIBLIOGRÁFICAS

ARNTZ, REINER Y PICHT, HERIBERT (1995). *Introducción a la terminología*. Madrid: Fundación Germán Sánchez Ruipérez.

BLASCO, ALBERT (coord.) (2006). *Manual de gestión de producción de alojamientos y restauració*. Madrid: Síntesis.

CALVI, MARIA VITTORIA (2006). *Lengua y comunicación en el español del turismo*. Madrid: Arco Libro S.L.

CARAVEDO, ROCÍO (1999). *Lingüística del corpus*. *Cuestiones teórico-metodológicas aplicadas al español.* Colección: Gramática española. Enseñanza e investigación. Apuntes metodológicos. Josse de Kock. I. 6. Salamanca, Universidad.

EDO MARZÁ, NURIA. (e.p). Páginas web privadas e institucionales: El uso de la adjetivación en un corpus inglés-español de promoción de destinos turísticos. En J. Sanmartín (Coord.), *Discurso turístico e Internet: normas y usos.*

FREIXÀ, JUDIT,/ KOSTINA, I/ CABRÉ, TERESA (2002). "La variación terminológica en las aplicaciones lexicográficas". En *Actas del VIII Simposio Iberoamericano de terminología.* Cartagena de Indias

GONZÁLEZ, VIRGINIA (e.p). El discurso del turismo en Internet: hacia una caracterización general. En J. Sanmartín (coord.), *Discurso turístico e Internet: normas y usos.*

MORENO FERNÁNDEZ, FRANCISCO (1999) Lenguas de especialidad y variación lingüística. En S. Barrueco, E. Hernández, y L. Sierra (Eds.) *Lenguas para fines específicos (VI). Investigación y enseñanza, Alcalá de Henares,* 3-14.

PIERA, CARLOS/ VARELA, SOLEDAD (1999). Relaciones entre Morfología y Sintaxis. En I.Bosque y V. Demonte (Eds.) *Gramática descriptiva de la lengua española*, vol, 3. Madrid: Espasa, 4367-4422.

RUIZ GURILLO, LEONOR (2001). *Las locuciones en español actual*. Madrid: Arco.

SANMARTÍN, JULIA (e.p.a). Unidad y variación en el español del turismo: las páginas electrónicas de promoción de hoteles de España y de Chile (e.p.). En *Actas del Simposio Confini Mobili. Lingua e cultura nel discorso turístico*. Milán, 10-12 noviembre 2010.

VAL ÁLVARO, JOSÉ FRANCISCO (1999). La composición. en I. Bosque y V. Demonte (eds.): *Gramática descriptiva de la lengua española*, vol, 3. Madrid: Espasa, 4757-4843.

Visual analytics: A novel approach in corpus linguistics and the Nuevo Diccionario Histórico del Español

Roberto Therón, Laura Fontanillo Fontanillo, Andrés Esteban Marcos, Carlos Seguín Herrero

Departamento de informática y automática

Universidad de Salamanca

Abstract

The aim of this article is to introduce visual analysis in corpus linguistics. This is a novel approach that is based on the integration of automated processes and humans' unique abilities, within a common effort to gain insight into complex problems that have to deal with vast amounts of data. In particular, our intention was the advancement of the application of information visualisation techniques to the diachronic linguistics field. The proposal of novel, highly interactive, visual solutions is approached by means of the Computational Information Design methodology, an integral process that brings together fields such as information visualisation, computational linguistics, data mining and graphic design, for the creation of tools that truly support knowledge discovery and other general tasks carried out by linguists. The article discusses the choices made for the design and development of interactive visual tools triggered by the creation of the New Spanish Historical Dictionary (Nuevo Diccionario Histórico del Español, NDHE).

Keywords: *visual analytics, diachronic linguistics, NDHE.*

Resumen

El objetivo de este artículo es introducir el análisis visual en la lingüística de corpus. Se trata de un enfoque novedoso que se basa en la integración de procesos automatizados y de las habilidades únicas de los seres humanos, en un esfuerzo común para profundizar en problemas complejos que tienen que tratar con grandes cantidades de datos. En concreto, se pretende avanzar en la aplicación de técnicas de visualisación de información en el campo de la lingüística diacrónica. La propuesta de novedosas y enormemente interactivas soluciones visuales se aborda por medio de la metodología Diseño Computacional de Información, un proceso integral que reúne a campos como la visualisación de la información, la lingüística computacional, la minería de datos y el diseño gráfico, para la consecución de herramientas que realmente ayuden en el descubrimiento de conocimiento y en otras tareas generales de los lingüistas. Aquí analizamos las opciones elegidas durante el diseño y desarrollo de herramientas visuales interactivas, propiciadas por la creación del Nuevo Diccionario Histórico del Español (NDHE).

Palabras clave: *analítica visual, lingüística diacrónica, NDHE.*

1. Introduction

Our contribution begins with a failure story as Manuel Seco (Seco, 1995) narrates it: "In 1914 the director of the Real Academia Española (RAE) proposed to produce a full-scale historical dictionary that would go beyond the prescriptive dictionaries. Work began in 1927, with the first volume (*A*) being published in 1933, and the second (*B-Cev*) in 1936".

A second attempt, the Historical Dictionary of the Spanish language, was interrupted in 1996 due to the deficiency of the documentary basis, organizational difficulties and lack of resources, in order to develop a computer database and other lexical resources that would enable and expedite the drafting process. Finally, in 2006 the NDHE project was initiated.

This story serves as an example of the urgent need of both automated processes and ancillary tools that support the documentation, analysis and writing tasks.

To the best of our knowledge, our work is the first one that initiates an innovative research line in the diachronic linguistics field, by means of the development of interactive visual tools that allow carrying out tasks involving analysis of temporal phenomena related to the evolution of a lexicon. Our contribution is centred in the Spanish language, but the proposed visual solutions are general and can be used with any language, as long as the necessary historical corpus is available.

This paper focuses on the visual analysis of meaning evolution, and it is part of a more ambitious project that includes several coordinated visualisation techniques for a complete analysis in diachronic linguistics (see Figure 1).

Figure 1: The analytical tasks of the linguists are supported by (linked) Infovis techniques.

2. The Interactive Visual Analysis Approach in Diachronic Linguistics

The motivation of this article is to make the corpus linguistics community aware of how interactive visualisation can support the analytical tasks of linguists. The main idea behind our approach may be put as *letting the linguist interact with a representation until it answers a particular research question (that might be unknown to the linguist before starting the analysis).*

An example of such tool is our CorpusExplorer prototype (Figure 1.A), built to investigate interactive solutions for diachronic corpus analysis (using the corpus recently released by Google accompanying its Ngram Viewer online application[75]). The interactive two-sided word trees enhanced with small timelines (sparklines) enable linguists to identify not only which are the most common groupings of words, but also to detect which words have substituted others in use, pinpoint which specific moment in time the trends started reversing and find other words with similar patterns.

Another example is the tool (Figure 1.B) that we developed for the visual analysis of data stored in the CORDE[76]. In this case, through the coordinated interaction with a selection of well-established Infovis techniques (force directed or radial graphs, piecharts, timelines and choropleth maps), the expert is able to assess different dimensions of language variation (the example shows the use of the word *tigre (*tiger) through time (timelines: diachronic analysis), register (piecharts: diaphasic analysis) and regions (choropleth: diatopic analysis).

Thanks to our collaboration with the RAE, and most relevant for the present work, we were exposed to the map of dictionaries in which, by comparing several editions of the dictionary of the Academia (DRAE), the evolution of words and meanings in these dictionaries over a period of more than two centuries can be traced (Pascual, 2009). Figure 1.C shows the interactive visual tool we have developed for such a map of dictionaries; a highly interactive diagram highlights the temporal patterns of meaning evolution through the different editions of the normative dictionary, which also has the ability to make historical mistakes or anomalies evident (e.g., one meaning that disappears in an edition and is recovered in the next one).

Because of space limitations, we have decided to omit both detailed discussions of all mentioned tools and case studies, and we focus on how the visual analytics approach has been applied to the map of dictionaries in the rest of the paper.

3. Related work

A review of computational models for the integration of visual and linguistic information can be found in (Srihari, 1995). It is the first article to categorize the research works that have dealt with the correspondence problem, namely, how to associate visual events with words and vice versa. Nevertheless, at present there are few works dedicated to the

75 http://ngrams.googlelabs.com/

76 REAL ACADEMIA ESPAÑOLA: Banco de datos (CORDE) [on line]. Corpus diacrónico del español. http://www.rae.es (may, 2011)

interactive visualisation of linguistic data; specialized tools for language analysis still make little use of visualisations, and visualisation tools for language related information (Culy and Lyding, 2010). For a comprehensive review we refer the reader to (Collins, Penn and Carpendale, 2008).

In spite of the great achievements in computational linguistics, the introduction of interactive visual tools in linguistics is a very recent tendency and is not yet popular, particularly among linguistics professionals. The advances are mainly related to document and text mining (van Ham, Wattenberg and Viegas, 2009; Don et al., 2007), with excellent and stunning results such as the ones available at Many eyes, a website devoted to the popularization of data visualisation (Viegas et al., 2007; Wattenberg and Viegas, 2008; Collins, Carpendale and Penn, 2009). More directly related to the our work, several research works can be mentioned, like (Manning, Jansz and Indurkhya, 2001), in which software for the visual exploration of a dictionary of Warlpiri is presented, or (Derrick and Archambault, 2009), in which a tool for the presentation and exploration of grammatical trees is proposed, or even a successful commercial case, VisualThesarus[77], where interactive visual maps are used to visualise word relationships for various languages.

4. Design Decisions

Lexicographers require several types of information in order to make assessments on the historical evolution of the meanings of a word. The main task can be summarized as the process of organizing the different synchronic sequences of the dictionary to explain how the current situation was reached, and comparing the different representations of the meanings throughout time. RAE's map of dictionaries was built to provide a means of analysis of such evolution and requires the gathering, processing and representation of several types of information, including: (1) order of meanings in the different editions of the normative dictionary, (2) actual linguistic content of each meaning, and (3) relationships among meanings that can vary over the different editions of the dictionary (subject to split, merge, expand, etc.).

To be able to gain a deep understanding of the enormous quantities of data stored in the different RAE's corpora, methodologies originating in fields such as information visualisation, statistics, artificial intelligence, data mining, computational linguistics, graphic design, psychology of perception and human-computer interaction, are employed to solve the different parts of the problem.

The following sections expose the design decisions we took, based in the seven stages of the holistic approach to computational information design (CID) (Fry, 2004). Please note that these stages are not sequential but iterative.

4.1. Acquire (I) and Parsing (II)

Pre-processing and annotation had been already undertaken and the data was stored in the RAE's databases. Our interactive tools perform queries to retrieve all the information necessary to create the representations.

77 http://www.visualthesaurus.com/

4.2. Mining (III)

For the analysis of meaning evolution, having a set of ordered lists of meanings pertaining to each edition of the dictionary is not sufficient. Meanings between two different editions may disappear, appear, split, merge, etc. RAE's multidisciplinary team has developed a semi-automated algorithm for the computation of these relationships and we relied on it during a first stage. In order to provide the experts with additional information on the evolution of meanings, an implementation of the NIST algorithm (Doddington, 2002), originally intended to compare machine translation output with expert reference translations in terms of the statistics of short sequences of words (word N-grams), was included. The visual encoding (see next section) has proven to be independent of which of those two algorithms is used. The NIST algorithm provides results that are slightly different to those obtained with RAE's algorithm, which has triggered long lexicographic discussions.

4.3. Represent (IV)

A time diagram is a very natural encoding and it has the advantage of being a classical method for visualising temporal change. Furthermore, for simple data and basic analysis tasks, these approaches outperform specialized techniques, because they are easy to learn and understand (Aigner et al., 2007). However, the classical time diagram does not completely address our problem, since, starting from the most recent edition of a dictionary, a meaning may have branching relationships to none or several meanings of any of the previous editions. We had to come up with a novel time diagram, specially tailored for the evolution of meanings in historical dictionaries, that we have called *diachronlex* diagram. We built an interactive interface that, by means of automated processes, analyzes the data and provides an initial overview of the data set (see Figure 1.C, top), while enabling subsequent focus on regions of interest and access to details on demand (see Figure 1.C, bottom), which enables further analysis in turn (Shneiderman, 1996; Keim et al., 2008). Thus, diachronlex diagrams are intended for overview and analysis, while two ancillary visualisations provide further details: 1) a *meaning list view*, intended to highlight the emergence of new meanings, which displays the meanings just as ordered numbered circles in the dictionary columns, and uses a star shape instead to highlight meanings which appeared in one dictionary. 2) an *ordered references view*, intended to convey all the relationships for a certain meaning at first sight, in which each row shows a numbered circle for each of the related meanings in other dictionary editions (columns) for the leading meaning (pertaining to the most recent edition). Both ancillary views can show text boxes containing the actual meanings, preserving the grid structure and the meaning placement.

4.4. Filter (V)

The complexity of the evolution of meanings varies with words: from very simple cases with few meanings per dictionary that mostly have either a stable (meanings maintain very similar order in all editions) or a down-like pattern (meanings appearing in old editions

tend to go down and down in the meaning lists as new editions appear), to entangled cases, with up to sixty meanings in one edition that may emerge and disappear, split, merge, etc. Besides the aid of other means of interaction, we have implemented four filters to support the lexicographers' analytical tasks. The overview can filter out meanings that: disappear in any edition, are no longer present in the current edition, emerged in old editions or have branches among editions.

4.5. Refine (VI)

Improvements to the basic representation can make it clearer and more visually engaging (Fry, 2004). In this stage we decided upon the colour scheme, the shape of the connecting lines, the size of the numbers and the shape of the symbols. Perhaps the best references at this stage were the works of Tufte and Ware (Tufte, 1986; Tufte, 1990; Tufte, 1997; Ware, 2004; Ware, 2008). Every decision was taken bearing in mind that the users would be lexicographers at the beginning, and generic users at some stage in the future, once the tool is deployed as an open service.

4.6. Interact (VII)

Here we present some interaction techniques widely used in Infovis (Yi et al., 2007), explaining how we used them: *Select–mark something as interesting*. Upon clicking on the meanings or number of meaning lists, the corresponding evolution line is highlighted; *Explore–show me something else*. The possibility of hovering over the evolution lines of a complex diachronlex overview, or in the ancillary visualisations, is combined with the representations of different details; *Reconfigure–show me a different arrangement*. Zooming has been enabled in the overview visualisation, and is used once the focus of the analysis is on a particular meaning evolution; *Abstract/Elaborate–show me more or less detail*. Tooltips are active when the user hovers over number labels or time points in the diachronlex diagrams. Additionally, the user can double click on any of the representations of a meaning in the ancillary views and a contextual box appears. Finally, we have implemented animated transitions (slow-in, slow-out) when the analyst changes between any of the four variations of the ancillary views, in order to minimize the cognitive load on the users when performing an analytical task; *Filter–show me something conditionally*, (explained in section 4.4); *Connect–show me related items*. All diagrams follow a multiple linked-views approach together with a brushing technique (the representations of the selected meaning are highlighted in the other views).

Besides the interactions meant for the exploratory analysis, we have also considered users with the ability to edit, who would be able to easily change the relationships of meanings by dragging and dropping the numbered circles in any of the ancillary views. These changes are tracked back in user sessions and can be annotated.

5. Conclusions

In this paper we have applied the computational information design methodology to the domain of diachronic linguistics. To the best of our knowledge this is the first instance

of the application of information visualisation and visual analytics techniques to this particular problem. We have been lucky to work together with the multidisciplinary team of computer scientists and lexicographers of the RAE, who provided insightful recommendations while developing our tools. Although all the work has been done with the NDHE in mind, the generality of the visual encoding can be applied to any other historical dictionary. We have shown how our proposal can be integrated in a bigger system that would ease the challenge of building a historical dictionary. It would enable the application of visual analysis to many different tasks of the linguists. Finally, we hope to see the presented methodology applied to other corpus linguistics problems in the future.

Acknowledgments

This work was supported by the Ministerio de Ciencia e Innovación of Spain under project FI2010-16234

References

Aigner, W., Miksch, S., Müller, W., Schumann, H. and Tominski, C. (2007). Visualising time-oriented data-a systematic view. *Comput. Graph.*, 31(3), 401–409.

Collins, C., Carpendale, S. and Penn, G. (2009). Docuburst: visualising document content using language structure. In *Proceedings of Eurographics/IEEE-VGTC Symposium on Visualisation (EuroVis '09)*, pp. 1039–1046. Eurographics Association.

Collins, C., Penn, G. and Carpendale, S. (2008). Interactive visualisation for computational linguistics. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, pp 6–6, Morristown, NJ, USA. Association for Computational Linguistics.

Culy C. and Lyding V. (2010). Double tree: an advanced kwic visualisation for expert users. *Proceedings of the 14th International Conference Information Visualisation*, (pp. 98–103)

Derrick, D. and Archambault, D. (2009). TreeForm: Explaining and exploring grammar through syntax trees. *Lit Linguist Computing*, 53-66.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pp. 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Don, A., Zheleva, E., Machon, G., Tarkan, S., Auvil, L., Clement, T., Shneiderman, B. and Plaisant, C. (2007). Discovering interesting usage patterns in text collections: integrating text mining with visualisation. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 213–222, New York, NY, USA. ACM.

FRY, B. J. (2004). *Computational information design*. Ph.D. thesis. MIT.

KEIM, D A., MANSMANN, F., SCHNEIDEWIND, J., THOMAS, J. AND ZIEGLER, H. (2008). Visual analytics: Scope and challenges (pp. 76–90). *Visual Data Mining* . Springer-Verlag.

MANNING, C. D., JANSZ, K., AND INDURKHYA, N. (2001). Kirrkirr: Software for Browsing and Visual Exploration of a Structured Warlpiri Dictionary. *Lit Linguist Computing*, 16(2), 135–151.

PASCUAL, J. A. (2009). The preparatory stage of the NDHE: "Divide and rule" (pp. 3-28), *Perspectives on Lexicography in Italy and Europe*, Cambridge Scholars Publishing.

SECO, M. (1995). El diccionario histórico de la lengua espanola. *Int J Lexicography*, 8(3), 203–219.

SHNEIDERMAN, B. (1996). The eyes have it: A task by data type taxonomy for information visualisations. In *VL '96: Proceedings of the 1996 IEEE Symposium on Visual Languages*, p. 336, Washington, DC, USA. IEEE Computer Society.

SRIHARI, R.K. (1995). Computational models for integrating linguistic and visual information: A surrey. *Artif. Intell. Rev.*, 8(5-6), 349–369.

TUFTE, E. R. (1990). *Envisioning information*. Graphics Press, Cheshire, CT, USA.

TUFTE, E. R. (1986). *The visual display of quantitative information*. Graphics Press, Cheshire, CT, USA.

TUFTE, E. R. (1997). *Visual explanations: images and quantities, evidence and narrative*. Graphics Press, Cheshire, CT, USA.

VAN HAM, F., WATTENBERG, M. AND VIEGAS, F. B. (2009). Mapping text with phrase nets. *IEEE Transactions on Visualisation and Computer Graphics*, 15, 1169–1176.

VIEGAS, F. B., WATTENBERG, M., VAN HAM, F., KRISS, J. AND MCKEON, M. (2007). Manyeyes: a site for visualisation at internet scale. *IEEE Transactions on Visualisation and Computer Graphics*, 13(6), 1121–1128.

WARE, C. (2004). *Information Visualisation: Perception for Design, 2nd ed.,* Morgan Kaufmann.

WARE, C. (2008). *Visual Thinking for Design*. Morgan Kaufman.

WATTENBERG, M. AND VIEGAS, F. B. (2008). The word tree, an interactive visual concordance. *IEEE Transactions on Visualisation and Computer Graphics*, 14(6), 1221–1228.

YI, J. S., KANG, Y., STASKO, J. AND JACKO, J. (2007). Toward a deeper understanding of the role of interaction in information visualisation. *IEEE Transactions on Visualisation and Computer Graphics*, 13(6), 1224–1231.

Corpus, estudios contrastivos y traducción

El guión cinematográfico como corpus: un estudio contrastivo entre el español de Almodóvar y su traducción al inglés

Ángela Almela Sánchez-Lafuente

Samuel Gracia Mayor

Universidad de Murcia

Resumen

El presente estudio ofrece un análisis de la traducción al inglés de la película La Mala Educación*, realizada por uno de los cineastas españoles más internacionales: Pedro Almodóvar. A partir de nuestro corpus hemos realizado el análisis contrastivo de ambas versiones, original y traducida, a través de un método de estudio híbrido. En un primer estadio, se ha llevado a cabo un análisis cuantitativo por medio de ciertas herramientas de la lingüística de corpus que ofrecen información relevante sobre el texto analizado, tales como el estudio de las palabras clave. Este estudio de tipo cuantitativo ha sido interpretado y matizado por medio de un análisis posterior de naturaleza cualitativa del corpus meta, en el que se ha observado el trasvase a la lengua meta del lenguaje jergal utilizado, especialmente en lo referente a las expresiones expletivas dentro de un registro coloquial.*

Palabras clave: guión cinematográfico, Almodóvar, lingüística de corpus, traducción, lenguaje jergal

Abstract

The present study provides an analysis of the English version of Bad Education*, a film by one of the main Spanish directors: Pedro Almodóvar. From our corpus, we have performed a contrastive analysis of the original and the translated versions by means of a hybrid methodology. Firstly, a quantitative study has been conducted through some corpus linguistics tools, such as keyword analysis. This preliminary quantitative study has been interpreted and qualified by means of a subsequent qualitative analysis of the target corpus, in which the transference of the slang into the target language has been observed, especially expletive expressions.*

Keywords: screenplay, Almodóvar, corpus linguistics, translation, slang

1. Introducción

La investigación sobre Almodóvar se enmarca principalmente en el ámbito de los estudios culturales (Ben-Habib, 1998; Escena, 2001; Lombardo, 2004; Sorgato, 2004; Vukovic, 2004). Más recientemente, diversos elementos lingüísticos y traductológicos de su obra han sido también objeto de estudio, como en los trabajos de investigación realizados por Baldi (2004). Esta autora trata numerosos aspectos culturales a través de su materialización lingüística. Como ella misma apunta, Almodóvar crea su propio universo dentro de la *Movida,* que se ve reflejado en sus obras cinematográficas a través del *pop art*, la política, la música o el mundo de la droga. De mayor relevancia para nuestro estudio es el trabajo de investigación llevado a cabo por Moreno (2006), en el que se emplea un corpus de cinco películas de Almodóvar, siendo la más reciente la que constituye el objeto de estudio de la presente investigación: *La Mala Educación*. A partir de dicho corpus se analizan los elementos culturales dentro del contexto histórico-cultural de la época franquista y la Transición y la manera en que éstos se han trasvasado a la lengua meta. De acuerdo con la autora, los elementos más difíciles de traducir de la obra de Almodóvar pertenecen al "ámbito de la gastronomía, la tauromaquia, las manifestaciones artísticas autóctonas, el ámbito jurídico y los referentes con valor simbólico; así como refranes, dichos y modismos" (Moreno, 2006: 131). Un ejemplo dentro del ámbito jurídico lo encontramos en la traducción de *una pareja de guardias civiles* por *two patrolmen*, donde el traductor opta por utilizar un equivalente hiperonímico. En este sentido, en su traducción al inglés se observa una ligera tendencia a huir de tecnicismos mediante diferentes estrategias que van desde la descripción a la modulación o a la omisión, lo cual facilita la labor del traductor audiovisual, pero no siempre asegura la fidelidad al original.

Para conseguir un trasvase apropiado del entorno cultural del texto origen (TO), los aspectos comentados son sin duda de vital importancia. Sin embargo, creemos que, en el caso de un director tan castizo y a la vez trasgresor como Almodóvar, la adaptación a la cultura meta de la caracterización de los personajes por medio de su habla es igualmente importante. Con el término *habla* nos referimos aquí al concepto de *parole* de De Saussure (1964), *id est*, el uso particular que cada hablante hace de la lengua. Para ello han de tenerse en cuenta, en primer lugar, las limitaciones que ofrece el medio escrito con respecto al oral en términos de expresividad del hablante. Éstas vienen determinadas principalmente por las limitaciones espaciales que no permiten que los subtítulos contengan más de dos líneas de entre 28 y 40 caracteres cada una (Chaume-Varela, 2000: 78). Este esfuerzo por sintetizar impuesto conlleva a menudo una irremediable pérdida de información. Sin embargo, esta modalidad de traducción posee la ventaja de que elimina el problema del trasvase de aspectos sociolingüísticos tales como el acento de una lengua a otra en la que quizá éstos no encuentren un equivalente adecuado. De este modo, el canal escrito permite enmascarar o simplemente obviar el uso de aspectos fonéticos segmentales y/o suprasegmentales típicos de una comunidad lingüística por parte del hablante, lo cual facilita en gran medida la labor traductora. Ello no significa, no obstante, que el traductor audiovisual no deba esforzarse por respetar las características sociolectales e incluso idiolectales de otra índole. A nuestro parecer, éste es en efecto el mayor reto traductológico en la obra de Almodóvar, y muy especialmente en la obra que nos ocupa.

En este sentido, la mayoría de estudios que se ocupan de este tipo de trasvase conllevan un análisis contrastivo entre la parte hablada del guión original en la lengua origen (LO) y la lengua escrita de los subtítulos en la lengua meta (LM). Sin embargo, el presente estudio no pretende ocuparse de las restricciones propias del paso de un canal a otro, sino de ofrecer un estudio contrastivo entre el corpus de subtítulos en ambas lenguas, lo que, a nuestro parecer, constituye un objeto de estudio comparable de manera directa. Así pues, el presente estudio se centra en el trasvase del habla propia de los personajes a la lengua inglesa, especialmente en lo que respecta al uso particular que hacen del lenguaje jergal en el trabajo cinematográfico de Almodóvar *La Mala Educación.*

2. Metodología

De este modo, el objeto de estudio es el corpus de subtítulos de la película *La Mala Educación* en versión original (español) y en su versión traducida al inglés. En un primer estadio, se ha obtenido el perfil lingüístico del TO y el TM por medio del programa *WordSmith Tools* 5.0[78], que ofrece información relevante sobre los textos comparados, tales como una lista de frecuencias y el estudio de las palabras clave (*keywords*). Para esta aplicación, se ha tomado como corpus de referencia del español el corpus Cumbre[79], y el BNC (*British National Corpus*) para el inglés. También se han tomado algunos datos relevantes a nivel estadístico, como la ratio tipo/ítem estandarizada, para contrastar la densidad léxica de la versión original y la versión traducida.

Para obtener un corpus paralelo que facilitara la labor analítica de la traducción en sí, se han alineado el TO y el TM por medio del programa *TRADOS WinAlign*, incluido en la suite *SDL TRADOS 7 Freelance*[80]. Este *software* permite la alineación de archivos con contenido similar en idiomas diferentes. En una primera fase, se han alineado automáticamente ambos archivos por pares de segmentos, para lo cual el programa toma como unidad por defecto la oración delimitada por puntos. Sin embargo, en este caso los pares se han delimitado de acuerdo con las tabulaciones, ya que la unidad es el subtítulo, que, como sabemos, no tiene por qué constituir oración. Esta unidad de alineación facilita en gran medida la labor de revisión de la traducción, ya que durante la alineación automática las correspondencias 1:1 tienen un índice de aciertos mayor que cuando se trata de oraciones completas de mayor longitud. Además, los segmentos más cortos facilitan la labor revisora también a nivel visual. Una vez que la alineación está realizada, simplemente se debe localizar en el corpus paralelo la unidad de traducción en cuestión, esto es, la reunión del segmento origen con el segmento traducido.

78 Publicado por *Lexical Analysis Software Ltd.* y *Oxford University Press*, disponible en http://www.lexically.net/wordsmith

79 "CUMBRE: corpus lingüístico del español actual". Proyecto financiado por Sociedad General Española de Librería, S.A. (SGEL) (1994-1998). Más información disponible en http://www.um.es/grupolacell/proyectos/proyecto/1

80 Disponible en http://www.trados.com/

3. Resultados y discusión

3.1. Perfil lingüístico del TO y el TM

En primer lugar, resulta interesante destacar que el TO alcanza una ratio tipo/ítem estandarizada del 43%, mientras que el TM roza el 39%. Ello implica una menor riqueza léxica en la versión traducida. A ello puede haber contribuido el hecho de que en la versión en inglés hay canciones como *Quizás, Quizás, Quizás* o *Maniquí Parisien,* interpretadas ambas por Sara Montiel, que no se subtitulan. Ello, lógicamente, afecta también a la longitud del TM, en el que hay una diferencia de 400 palabras (ítems) y de 300 tipos con respecto al TO. Estos datos llevan asociados, de acuerdo con las estadísticas de *WinAlign*, un total de segmentos origen de 1.962 y un total de segmentos meta de 1.819. Más concretamente, 115 segmentos origen no encuentran equivalencia en el TM, y 28 tienen una correspondencia 2:1. La longitud media del subtítulo es similar en ambos subcorpora, de modo que podemos afirmar que en la versión en inglés se observa una mayor tendencia a la concreción que en la versión original, como muestra la Figura 1:

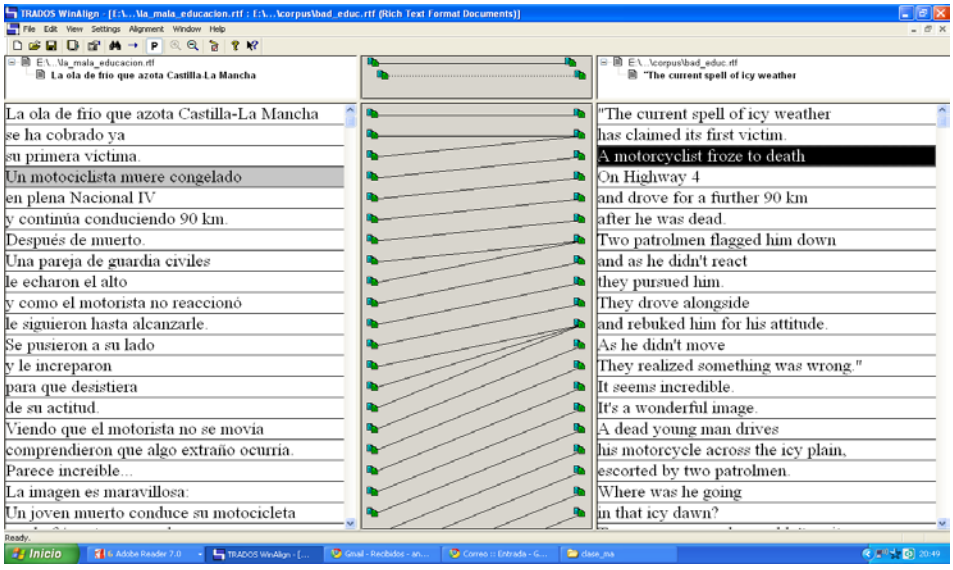


Figura 1. Alineación de Segmentos del T.O. y T.M con WinAlign

Asimismo, algunos ejemplos de subtítulos del TO que no encuentran equivalencia en el TM aparecen en la Figura 2:

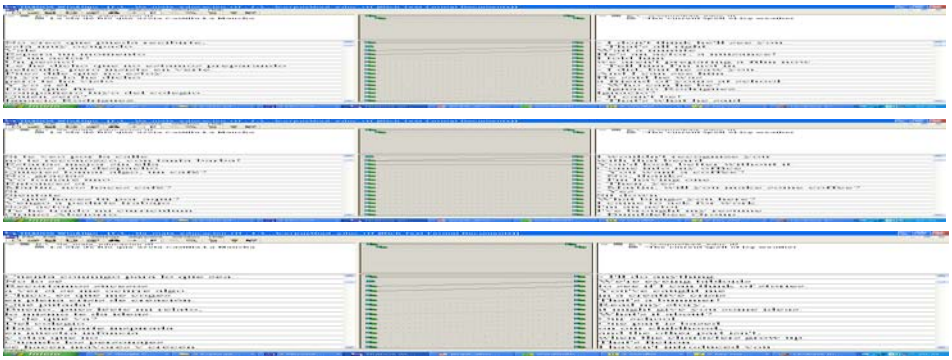


Figura 2. Subtítulos sin equivalencia en la alineación con WinAlign

El estudio contrastivo de las palabras más frecuentes del corpus, así como de las palabras clave, resulta revelador con respecto al producto de traducción. Como se ha comentado anteriormente, los dos aspectos fundamentales hallados en la lengua empleada en el guión del film son el uso constante de elementos jergales y la presencia de la feminidad en el habla de los protagonistas del cine dentro del cine para reafirmar su condición sexual. Ciertamente, la lista de palabras clave es muestra de ello (Fig. 3 y 4):

Figura 3. Palabras clave en español

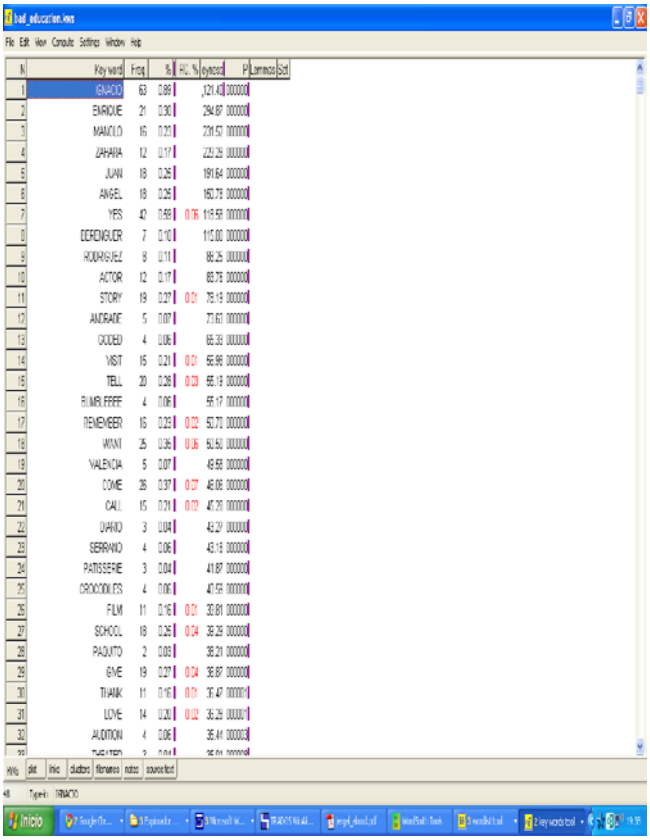


Figura 4. Palabras clave en inglés

3.2. La traducción de la jerga y las palabras tabú

Llama la atención el hecho que de las 30 primeras palabras clave en español (Fig. 3), dos de ellas sean jergales –*tío* y *maricón*– y que no lo sea ninguna en el TM. El examen de las unidades de traducción en las que aparecen estas dos palabras clave muestra que la mayor parte de veces, *tío* se traduce por *man*, vocablo propio de un registro ligeramente más elevado que en la versión original. En otros casos, directamente se omite este apelativo. En el caso de la otra palabra clave, la explicación de que no se encuentre su equivalente entre las primeras 30 palabras clave de la versión en inglés (Fig. 4) se debe a que el traductor emplea distintos equivalentes, tales como *fag*, *faggot* o *bitch*. Con ello, se pierde un rasgo idiolectal, ya que en el original este apelativo lo emplean con frecuencia determinados personajes. Dicha elección no se justifica tampoco por el número de caracteres empleados en el subtítulo, ya que cualquiera de los términos equivalentes es incluso más corto que el vocablo original. Por otro lado, existen en el TO ciertos rasgos coloquiales que se conservan tal cual en la forma escrita de los subtítulos, como las formas apocopadas

pá, tó y *ná,* difícilmente transferibles a la LM, para los que tampoco se aplica ningún mecanismo de compensación.

En la última década, el tratamiento de las expresiones jergales y las palabras tabú en traducción ha suscitado un interés creciente, ya que de su adecuado trasvase depende en gran medida la fidelidad a la caracterización de los personajes del producto original. Ejemplo de ello es el estudio llevado a cabo por Rojo López y Valenzuela Manzanares (2000), quienes hacen alusión a este fenómeno en los siguientes términos:

> [l]a proliferación en el uso de las expresiones malsonantes en el habla cotidiana se refleja en los medios de comunicación, el cine e incluso la literatura. Muchas de las obras literarias actuales, aspirando a alcanzar el mayor grado posible de autenticidad, están plagadas de expresiones malsonantes, sin que por ello dejen de ser reconocidas como obras de arte (*ibid*., 208).

En esta línea, Díaz-Pérez (1994) atribuye el constante uso de la jerga en los guiones de Almodóvar al hecho de que, desde finales de los años setenta y sobre todo a principios de los ochenta, entran en el círculo *underground* no solo los llamados "grupos marginales", sino, de igual modo, gente muy variopinta que cambia de registro dos o tres veces al día, para estar en armonía lingüística tanto en el trabajo como con la familia o en su ámbito de amigos.

La lengua del cine de Almodóvar en general, y de *La Mala Educación* en particular, consiste ante todo en la expresión de la informalidad. Entre los personajes tanto de la película como del cine dentro del cine se encuentran, por un lado, unos pertenecientes a ambientes marginales, travestidos y drogadictos, como Zahara y Paca, o transexuales, como Ignacio, y, por otro, a niveles socioculturales bajos, amas de casa de extracción rural, como la madre y la tía de Ignacio. Este contraste es típico del cine del director manchego, quien define a los grupos sociales por unos rasgos determinados que tienden a diferenciarse de los demás, y a afirmarse frente a ellos, con un estilo lingüístico especial. Hay que tener aquí presente que la distinción entre los diversos registros o niveles lingüísticos, cuando todos constituyen la lengua coloquial o informal, es muy relativa, ya que se establecen frecuentes trasvases de vocablos de uno a otro nivel (de argot, vulgar, familiar, culto o incluso poético).

3.2.1. La recurrencia de *fucking*

Especial atención merece la traducción de la voz *fucking,* debido a su uso tan extendido en inglés. A este respecto, en el estudio realizado por Rojo López y Valenzuela Manzanares (2000) mencionado anteriormente, los autores comentan que existen dos posibilidades sintácticas para su correspondencia en español. La primera consiste en un método de extracción del término expletivo de su sintagma, y la segunda en traducirlo como un modificador, ya sea anterior o posterior al núcleo (posición pre- o post-nuclear). Ello se cumple en el corpus objeto de nuestro estudio a la inversa. Un ejemplo del primer mecanismo sería la traducción de "*No seas susceptible, coño"* por "*Don't be so fucking touchy!"*. En este ejemplo, comprobamos cómo el término expletivo efectivamente se

traduce por la expresión *fucking*, que tiene a su vez función de modificador adjetival. Un ejemplo de la segunda categoría propuesta sería la traducción de "*He estado tres putos años haciendo mierdas…"* por *"I spent three fucking years doing shit…"*, donde el equivalente español para *fucking* es un modificador pre-nuclear, siendo su función en inglés la de modificador nominal. Por último, un ejemplo de modificador post-nuclear lo encontramos en la traducción del sintagma "*maricón de mierda"* por *fucking faggot.*

La frecuencia relativa de este modificador es de un 5,82% en el corpus de referencia BNC, mientras que en el TM alcanza el 8%. La adaptación de cualquiera de los ejemplos citados al inglés mediante el uso de esta voz nos parece sumamente apropiada, ya que su alta frecuencia de uso en lengua inglesa ayuda a plasmar perfectamente la naturalidad del habla empleada en el original.

3.2.2. El género marcado y la dificultad de su traducción

En el producto original, aparecen numerosas expresiones que se derivan de la identidad sexual y del tratamiento interpersonal que existe entre los personajes. En la película, los dos travestis mantienen una apelación interpersonal que se da continuamente en femenino; además, suelen utilizar entre ellos apelativos normalmente peyorativos en contextos heterosexuales y que pierden dicha connotación en contextos homosexuales. Debido a las características de la LM, la traducción del primer aspecto mencionado resulta problemático, ya que no se puede marcar el género mediante sufijación. En nuestro análisis, hemos comprobado cómo, en ciertos casos, el traductor opta por mantener el término original, como en el caso de la voz *amigas*. Es posible que la elección no esté motivada solo por la ausencia de heteronimia en el equivalente inglés *friend*. También puede deberse a que el término *amigo* se emplea en inglés americano con connotaciones despectivas en el suroeste de EE.UU., lo que puede ayudar mantener la connotación marginal del original.

En lo que respecta a la palabra *mujer,* ésta aparece en el TO 14 veces, mientras que en el TM su equivalente *woman* registra solo la mitad de ocurrencias. La explicación reside en que en las ocasiones en las que se emplea como apelativo, simplemente se suprime. Otro ejemplo de apelativo femenino común en el corpus es *cerda*, para el que el traductor emplea el equivalente *pig,* a pesar de existir un equivalente con el mismo género en la lengua meta: *slut*. En otras ocasiones, el género femenino del original se sustituye por un término con una connotación homosexual, como es el caso de *niña* traducido como *chicken,* término que en la LM se suele usar para designar a un joven homosexual.

4. Conclusión

Tras el estudio, podemos afirmar que en el subcorpus correspondiente al producto audiovisual en su versión meta se observa una mayor tendencia a la concreción que en la versión original. En algunos casos, esta tendencia no hace sino dificultar la comprensión del texto, ya que, precisamente por retratar una realidad ambientada en una cultura diferente a la del espectador, muchos de los elementos trasvasados requieren una mayor

explicitación. Tal es el caso de algunas de las canciones interpretadas, o de las unidades de traducción analizadas con correspondencia 2:1 con menor detalle que en el TO.

Además, el menor número de segmentos en la LM también se debe al hecho de que muchos apelativos empleados en la LO no encuentran correspondencia, lo cual no contribuye a configurar adecuadamente el uso tan característico que de su identidad sexual hacen los personajes. A ello se le une la carencia que supone la ausencia de flexión propia del inglés. No obstante, como hemos visto en el análisis, el traductor ha tratado de subsanarlo con algún mecanismo de compensación adecuado.

A modo de conclusión, creemos que debe prestarse mayor atención desde un punto de vista académico al estudio contrastivo de productos audiovisuales como el que nos ocupa, a lo cual pueden contribuir en gran medida las aplicaciones de software de análisis de corpus como *WordSmith Tools* y los programas de alineación de corpora tales como *TRADOS WinAlign*. Ellos pueden ayudar al traductor a evaluar las unidades de traducción resultantes de manera directa, lo que puede ayudarle en su labor futura. La importancia de ello radica en que si el traductor es capaz de plasmar en la LM el mensaje original con la misma frescura, ello tendrá una repercusión directa en el efecto conseguido en el espectador meta, ofreciéndole lo más fielmente posible una muestra de la cultura española.

Referencias Bibliográficas

Almodóvar, P. (Director) y Almodóvar, A., Almodóvar, P., y García, E. (Productores) (2004). *La Mala Educación.* España: El Deseo.

Baldi, A. (2004). *Lenguajes y lengua en la obra de Pedro Almodóvar*. Bari: Servicio de Publicaciones de la Universidad de Bari.

Ben-Habib, L. (1998). *Laberinto español: repetición y elementos culturales en la Parodia de Almodóvar.* Tel-Aviv: Universidad de Tel-Aviv.

Chaume-Varela, F. (2000). *La traducción y la interpretación en España hoy: Perspectivas Profesionales*. Granada: Comares.

Díaz-Pérez, J. C. (1994). Presencia de la comunicación jergal en la enseñanza de español para extranjeros: los guiones cinematográficos de Pedro Almodóvar. *Actas del IV Congreso ASELE*.

Escacena, M. (2001). El cine de Pedro Almodóvar. *Centro de Documentación Pedro Almodóvar.* Disponible en http://sdogma.uclm.es/uclm/html/tesis/tesis.html

Lombardo, S. (2004). Visioni di amore e di morte: Sull'ultimo cinema di Pedro Almodóvar. *Centro de Documentación Pedro Almodóvar.* Disponible en http://sdogma.uclm.es/uclm/html/tesis/tesis.html

Moreno, A. (2006). *La traducción de elementos culturales en el texto audiovisual. La obra de Pedro Almodóvar en alemán, francés e inglés.* Vizcaya: Universidad del País Vasco.

ROJO, A. Y VALENZUELA, J. (2000). Sobre la traducción de las palabras tabú. *Revista de Investigación Lingüística, 3,* 207–220.

SAUSSURE, F. (1964). *Curso de lingüística general.* Buenos Aires: Losada.

SORGATO, E. (2004). *Entre ingenuidad y consciencia: estudio de los personajes de* Hable con ella *de Pedro Almodóvar.* Padua: Universidad de Padua.

VUKOVIC, M. (2004). *Nothing is Simple. The Unique World Created by Pedro Almodóvar*. New York: The City University of New York.

Lexico-grammatical divergence in Malay translated text: A corpus-based analysis of the relative clause marker *yang*

Norsimah Mat Awal, Imran Ho-Abdullah & Intan Safinaz Zainudin

School of Language Studies and Linguistics,

Faculty of Social Sciences and Humanities

Universiti Kebangsaan Malaysia

Abstract

Laviosa (1998, 2002) suggests that corpus-based approach is the 'new paradigm in translation studies'. Since then, various translation studies utilizing corpus-based approach have been conducted. This study uses a comparable corpus to investigate the lexico-grammatical differences of the Malay relative clause marker *yang* as it is one of the salient lexical items found in the corpus. The comparable corpus is made up of texts translated into Malay and texts originally written in Malay. Comparable corpus presents an opportunity to discover features that occur more frequently in translated texts or 'translation universals'. Findings on these translation universals would be a valuable tool in the teaching and training of translators.

1. Introduction

The use of corpus in translation studies has gained a substantial enthusiasts who argue that the corpus approach in translation can offer a multitude of information that is helpful in translation process. Wilkinson (2005) states that a corpus "can be of great help in confirming intuitive decisions, in verifying or rejecting decisions based on other tools such as dictionaries, in obtaining information about collocates (words that typically co-occur), in reinforcing knowledge of normal target language patterns, and in learning how to use new expressions." With the various information that a corpus can provide, the important role of corpus in translation studies could not be denied. Laviosa (1998) suggests that the use of corpus in translation studies or corpus-based translation studies as the "new paradigm" in translation studies. A corpus-based translation analysis could either use a parallel corpus or comparable corpus depending on the focus of the analysis. This paper focuses on a specific grammatical item, namely the Malay relative clause marker *yang* by analyzing the data in a comparable corpus that consists of translated Malay text and text originally written in Malay.

2. Comparable corpus and translation studies

Baker (1993:243) explores the concept of typicality or universal features of translation. She defines universal features of translation as "features which typically occur in translated text rather than original utterances and which are not the result of interference from specific language systems." The concept of typicality or universal to the concept of norms in translation advocated by Baker (1993, 1995) is akin to that proposed by Toury (2000). Some of the typical features identified by Baker are (i) explicitation (ii) strong preference for conventional 'grammaticality' (iii) tendency to avoid repetitions and (iv) tendency to exaggerate features of target text.

Some studies using comparable corpora in translation studies that purport to examine the universal features of translations include Laviosa (1998) and Olohan and Baker (2000). Laviosa (1998) explores the core patterns lexical use in translation texts. She uses two collections of narrative prose in original English texts and a comparable corpus made up of translations from a variety of source languages. Among her significant findings which she terms as 'core patterns of lexical use' are (i) translated texts have relatively lower percentage of content words versus grammatical words and (ii) the proportion of high frequency words versus low frequency words is relatively higher in translated text.

Olohan and Baker (2000) examine the use of relative pronoun *that* in the Translational English Corpus (TEC) at the University of Manchester and compared its frequency in the reference corpus, the British National Corpus (BNC). They found that in the BNC the relative pronoun tend to be omitted more often when used in conjunction with contractions and *that* occurs more frequently with contractions in TEC. Corpus-based translation studies also gain prominence in language other than English. Tirkkonen-Condit (2005) examines the usage of particle *kin* in texts translated into Finnish and texts written in Finnish. Her study of a corpus of Finnish tanslation texts across 5 genres, found that the frequency and usage of the clitic particle *kin* in Finnish translation texts was significantly

lower (4.6 per 1000 words) compared to the use of the particle in original Finnish (6.1 per 1000 words). She hypothesized that unique elements in language tends to be under represented in translation language. These studies among others have illustrated the usefulness and effectiveness of corpus based methods in examining translation universals.

3. Methodology

The methodology of the present study is based on corpus linguistics methodology. The following section will discuss the corpus and data generated.

3.1 Research Design

The design for the study utilized data from the corpus that is generated by WordSmith program. The initial wordlist of the corpus is generated and subsequently all content words are omitted. The resulting wordlist is a list of most frequent lexico-grammatical items. A comparative list of such items was obtained from the UKM-DBP corpus of Malay text.

3.2 The Corpus

The Translation Corpus (TC) used in the present study is a specialized corpus based on the translation works by students in the translation courses taught. The courses are level 1 translation courses at Universiti Kebangsaan Malaysia. All the assignments are English to Malay translations. The students who are enrolled in the translation courses have the necessary proficiency in both source and target language. The TC currently contains 23,516 words and represents translation language. The basic statistics of the TC is as follows:

Table 1. Basic Statistics of Translation Corpus

Total tokens	23,516
Total types	4,384
Type-token Ratio	18.6

For the purpose of comparison, the UKM-DBP 5 million word corpus of Malay was used to provide statistical information on natural occurring Malay. The Keyword computer program now incorporated in the Wordsmith program, available at web site http://www.oup.co.uk/elt/software/wsmith) was used to compare frequency lists from the TC corpus and the UKM-DBP corpus. The keyword program allowed us to generate a list of frequent words (salient items) that were more significantly frequent in the TC compared to the UKM-DBP corpus. The salient grammatical items for the TC corpus are listed in the table below. It should be noted that twelve grammatical items are statistically significant in the TC as compared with the UKM-DBP corpus (Table 2).

4. Findings and Discussions

The data from Table 2 below based on keywords or salient items provide us with a principled approach to deciding which grammatical words to analyze. The pronouns (*anda* 'you'; *saya* 'I') are salient and appear more frequently in the translation corpus while and *kita* 'us' are significantly under represented in the translation corpus. The relative clause marker *yang* also appears more frequently in the translation corpus. Other significant grammatical items that are salient in the translation texts include the conjunction/ coordinator *untuk* 'for' and modal auxiliaries such as *telah* and *akan*. For the purpose of this paper only one item will be analysed further namely the relative clause marker *yang* to illustrate the principle of using corpus data in translation studies. A concordance for *yang* is generated using WordSmith Concord which provided the data for a contextual analysis of each grammatical use of *yang*.

Table 2. Salient Grammatical words in Translation Corpus

word	Frequency (TC)	% (TC)	Frequency (UKM-DBP)	% (UKM-DBP)	χ^2 score	P*
anda	380	1.60	5088	0.14	1150.49	0.00
untuk	361	1.52	30946	0.84	105.96	0.00
yang	1026	4.31	121033	3.28	73.25	0.00
telah	133	0.56	8992	0.24	70.68	0.00
tersebut	81	0.34	4477	0.12	62.57	0.00
akan	217	0.91	19512	0.53	54.44	0.00
adalah	134	0.56	10879	0.29	45.78	0.00
saya	151	0.63	13054	0.35	42.92	0.00
jikalau	8	0.03	96	-	25.66	0.00
ada	42	0.18	12958	0.35	-25.13	0.00
kita	48	0.20	16339	0.44	-38.98	0.00
itu	109	0.46	42306	1.14	-127.52	0.00

4.1 Malay relative clause marker yang

In the Malay language, *yang* is identified as a relative clause marker and functions as a modifier of relative clause. A clause that is marked by *yang* serves as a modifier to one of the aspects in the main sentence. In other words, *yang* takes the role as a modifier for the subject, object or predicate or as a modifier to a noun or noun phrase as in the examples below:

i. *Orang yang datang* itu sahabat karib saya.

The person who came is my close friend.

ii. *Universiti ini mengeluarkan siswazah yang cukup terlatih.*

This university produces graduates *who* are well-trained.

Yang also provides additional information for adverbial clause and the adverbial clause is usually marked by link words such as the prepositions *kepada* 'to', *pada* 'to' and *dalam* 'in' as in the following examples:

iii. Dia berangkat *pada hari yang bersejarah itu*.

He left on that historical day.

iv. Dia memberi topinya *kepada orang yang menghulur tangan itu*.

He gave his hat to the man who extended his hand.

Quah (2001) states that the *yang* relative clause marker is similar to the English relative pronouns 'that', 'which' and 'who'. She further states that the *yang* structure is the only relative clause form in Malay. Therefore, the *yang* structure is the most is the most productive form, especially in written Malay. The 'yang' relative clause is used to substitute noun that functions as the subject of a sentence. The relative clause structure of *yang* is *yang* + X, where X represents either an N(P), a V(P), and Adj(P) or a quantifier (Quah 2001:116). It is also claimed that of all these phrases, the *yang* structure that is most productive in written Malay is *yang* + V(P).

4.2 yang in the translation corpus

Based on Table 2 above, *yang* is the most salient in the translation corpus with a frequency count of 1026. This section will describe the most productive structure of the *yang* relative clause in the translation corpus which is *yang* + Adj(P). This is contrary to the findings on the most productive relative clause structure in written Malay which is *yang* + V(P). The basic statistics of the yang relative clause structure is as follows:

Relative clause pattern	**Frequency**
yang + Adj(P)	547
yang + V(P)	392
yang + N(P)	87
Total	1026

The high number of the *yang* + Adj.(P) structure can be attributed to the nature of adjectives that are highly descriptive. This is also in line with the nature of a translated text. Blum-Kulka (1986) states that explicitation as a process of interpretation performed by the translator on the source text that might lead to a TL text which is more redundant than the SL text. Seguinot (1988) on the other hand did not agree with the idea of explicitation as more redundant but 'reserved for additions which cannot be explained by structural, stylistic or rhetorical differences between the two languages, and addition is not the only device of explicitation'. According to Klaudy (2001) states that explicitation

takes place not only when 'something is expressed in the translation, which was not in the original', but also in cases where 'something which was implied or understood through presupposition in the source text is overtly expressed in the translation, or an element in the source text is given a greater importance in the translation through focus, emphasis or lexical choice. There are 4 types of explicitation as shown below:

1. Obligatory – dictated by the differences in the syntactic and semantic of the structure of the languages

2. Optional – dictated by the differences in text-building strategies (clauses, connective elements, relative clauses)

3. Pragmatic – dictated by different cultures

4. Translation-inherent – attributed to the nature of the translation process itself

The focus of this paper is the fourth type of explicitation that is the concept of translation-inherent explicitation. The examples below from the translation corpus show that 'yang' relative clauses does not belong to the other types of explicitation but the fourth type- the translation inherent type, the information (as underlined) is needed to complete the nominal phrase.

v. *Katakan sebuah paya yang cantik dan bersih yang merupakan habitat hidupan.*

Let's say there a clean and beautiful swamp that serves as the breeding place for habitat.

vi. *Hal ini boleh menjadi perniagaan yang sukar, kerana kebimbangan yang menghalang anda…*

This can be a difficult business because of the worries that you have may precent you…

In written Malay, it is quite common to have multiple *yang* structures within a sentence. Mohd Zain Mohd Ali (1987:109) refers to this practice as "stacking" of relative clauses where, "the first clause modifies the head noun, the second modifies the head noun already modified by the first clause, and the third modifies the head noun as in turn modified by the second clause, and so on." He also confirms that this practice is seen in Malay translation and supports the claim by Baker (1993) that explicitation as one the features of translation universals.

5. Conclusion

This study has attempted to use corpus in translation studies and apply the findings to inform the teaching of translation. In order to do this, we have first compiled a translation corpus. Data of salient differences between translation language and original language is then generated to inform our teachings. In the present case, we have investigated the over presentation of *yang* in Malay translation – where the item is seen as a 'convenient' equivalence in the translation of English *which, that* and *who*. Our investigation made

possible via corpus-based method based on a significant amount of translation sentences involving *yang* seems to indicate that *yang* which is a typicality in translation (language) is a schematic extension of its function in Malay.

References

AINON MOHD & ABDULLAH HASSAN. 2000. *Teori dan Teknik Terjema*han. Kuala Lumpur: Persatuan Penterjemah Malaysia.

ASMAH HAJI OMAR 1986 6. *Nahu Melayu Mutakhir: Edisi Baru*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

BAKER, M. 1992. *In Other Words: A Course Book on Translation.* London: Routledge.

BAKER, M. 1993. Corpus linguistics and Translation Studies. Mona Baker, Gill Franciss, Elena Tognini.Bonelli. (eds.) *Text and Technology: In Honour of John Sinclair*. Philadelphia: John Benjamins Publishing Company.

HUNSTON, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

JAKOBSON, R. (2000). On Linguistic Aspects of translation. In Venuti. L. (ed.) *The Translation Studies Reader*. London: Routledge.

LAVIOSA, S. (1998a). The corpus-based approached: A New Paradigm in Translation Studies. *META*. 13(4): 474-479.

LAVIOSA, S (1998b). Core Patterns of Lexical Use in a Comparable Corpus of English. *META*. XLII(4): 1-15.

LAVIOSA, S. (2002). *Corpus-based Translation Studies: Theory, Findings, Applications*. Amsterdam:

RODOPI B.V.

MASLIDA YUSOF. 2008. Struktur semantik preposisi 'Bertujuan': Satu analisis berdasarkan korpus. Dlm NorHashimah Jalaluddin, Imran Ho Abdullah, Idris Aman.(eds.) . *Linguistik: Teori dan Aplikasi*. Bangi: Penerbit UKM.

MUNDAY, J. 2008. *Introducing Translation Studies: Theories and Applications* (2nd edition). London: Routledge.

NEWMARK, P. (1981). *Approaches to Translation.* Oxford: Pergamon.

NIDA, E. (2000). Principles of Correspondence. In Venuti. L. (ed.) *The Translation Studies Reader*. London: Routledge.

NIK SAFIAH KARIM, FARID M. ONN, HASHIM HJ MUSA AND ABDUL HAMID MAHMOOD. 1996. *Tatabahasa Dewan Edisi Baharu*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

OLOHAN, M & M. BAKER. 2000. Reporting *that* in Translated English: Evidence for Sub-conscious Processes of Explicitation. *Across* 1: 142-172.

SEGUINOT, C. 1985. translating Implicitation. *Meta* 30: 295-8.

SINCLAIR, J. (ed.). 1991. *Prepositions*. London: Harper Collins Publisher.

TIRKKONEN-CONDIT, S. 2005. Do Unique Items Make Themselves Scarce in Translated Finnish? In: Károly, K. & Fóris, Á. (eds.) *New Trends in Translation Studies. In Honour of Kinga Klaudy* . Budapest: Akadémiai Kiadó, 177–189.

TOURY, G. (2000). The Nature and Role of Norms in Translation. In Venuti. L. (ed.) *The Translation Studies Reader*. London: Routledge.

Detección y clasificación de errores de traducción de las unidades terminológicas contenidas en un corpus paralelo multilingüe de turismo de salud y belleza

Cristina Castillo Rodríguez
Universidad de Málaga

Resumen

El objetivo de esta investigación es analizar la calidad de las traducciones publicadas en la red del material promocional del segmento del turismo de salud y belleza. Para ello se ha compilado un corpus paralelo multilingüe integrado por textos originales (TO) escritos en lengua española y sus textos traducidos o textos meta (TM) al inglés, francés e italiano. Asimismo, para procesar los bitextos contenidos en este corpus paralelo se ha llevado a cabo el proceso de alineación con ayuda de la herramienta informática WinAlign para poder analizarlos, posteriormente, con el programa de gestión de corpus paralelos ParaConc, y ofrecer una clasificación de los errores principales de traducción de la terminología empleada en este tipo de textos turísticos.

Palabras clave: corpus paralelo, traducción turística, errores de traducción, turismo de salud y belleza

Abstract

The purpose of this paper is to analyse the quality of the translations published on the net of the promotional material belonging to the segment of tourism called wellness and beauty. In order to carry out this analysis we have compiled a multilingual parallel corpus composed by texts originally written in Spanish and their translations into English, French and Italian. Besides, with the aim of managing the bitexts contained in this parallel corpus we have carried out the process of alignment through the use of the tool WinAlign *in order to proceed with the contrastive analysis using the parallel corpus management software* ParaConc*, and thus, to offer a classification of the main translation mistakes of the terminology used in this kind of tourist texts.*

Key words: parallel corpus, tourist translation, translation mistakes, wellness and beauty tourism

1. Introducción

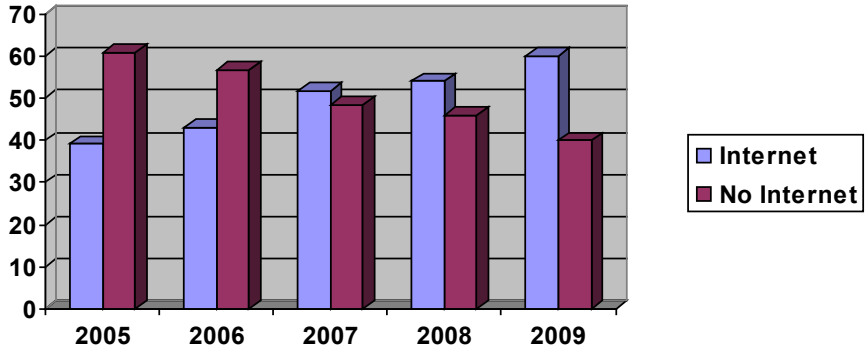
Desde el conocido "boom turístico", que se produjo durante el periodo comprendido entre los años 50 y 70 (Vogeler y Hernández, 2000), el turismo se ha convertido en una de las industrias generadoras de divisas en la mayoría de los países y, muy especialmente, en España, puesto que abarca un conjunto de actividades que tienen por objeto la explotación de las riquezas turísticas, así como la transformación de los recursos humanos, de capital y de materias primas, tanto en servicios como en productos.

Además, hoy en día, en el seno del mercado turístico se ha presenciado, sobre todo, en la última década, una auténtica revolución debido al creciente uso de las tecnologías de la información y de las comunicaciones (TIC). Por consiguiente, el empleo de la red Internet para consulta, reserva y pago de los servicios turísticos ha aumentado de forma notable en los turistas que eligen España como destino turístico vacacional.

Como se indican en los informes de Encuesta de Gastor Turístico (Egartur)[81], elaborados por el Instituto de Estudios Turísticos (IET), del Ministerio de Industria, Turismo y Comercio, cada año se aprecia una tendencia al alza en cuanto al uso de la red por parte de los turistas internacionales a la hora de consultar, reservar y contratar los servicios turísticos ofertados en España. Desde que en el año 2005 arrancara con fuerza el empleo de la red como medio para contratar estos servicios, hemos sido testigos de una creciente preferencia de este medio por parte de los turistas que no ha dejado de despuntar hasta la actualidad. De hecho, si observamos el porcentaje de turistas que usan Internet desde el año 2005 hasta el año 2009, vemos que este porcentaje ha ido incrementándose hasta desbancar el porcentaje de turistas que aún siguen reacios a utilizarlo.

Así, y según los informes de Egartur, en el año 2005 el 39,1% de los turistas utilizó Internet para planificar su viaje a España, mientras que el 60,9% no lo empleó. En el año 2006 se registró un 43,2% de los turistas que sí hicieron uso de la red Internet para contratar los servicios turísticos en nuestro país, aunque el 56,8% aún no confiaban en las facilidades disponibles en la red para contratar estos servicios. Al año siguiente, esto es, en 2007 el porcentaje de turistas que empleó la red Internet para organizar su viaje a España ascendió a un 51,6%, mientras que el 48,4% constituye el porcentaje de turistas que no utilizó Internet. En el año 2008 el 54% de los turistas que planificaron su viaje a España hizo uso de la red para consultas, reservas o servicios contratados en su totalidad, frente al 46% que prefirió no utilizarlo. El último informe de Egartur publicado se corresponde al año 2009, en el que el 60% de los turistas internacionales consultó, reservó o, incluso, contrató los servicios turísticos vacacionales en España a través de Internet, mientras que el porcentaje de turistas que no hizo uso de Internet descendió a un 40% del total de turistas. A continuación, mostramos una figura que ilustra de forma gráfica cómo ha ido aumentando esta preferencia de la red Internet por parte de los turistas internacionales como medio para planificar su viaje a España y cómo ha ido descendiendo, por el contrario, el porcentaje de turistas que no hacen uso de este medio, como acabamos de describir:

79 Para consultar cada uno de los informes de Egartur, diríjase a la siguiente URL donde se encuentran albergados: http://www.iet.tourspain.es/paginas/PubEgatur.aspx?option=egat&idioma=es-ES



Gráfica 1. Evolución porcentual de turistas que utilizan la red Internet (2005-2009)

Asimismo, en dichos informes de Egartur, así como en los informes de Movimientos Turísticos en Fronteras (Frontur), elaborados también por el IET, observamos que el alto índice de llegadas internacionales lo constituyen turistas que provienen, principalmente, del Reino Unido, Alemania, Francia, Países nórdicos e Italia. Siendo conscientes de esta situación, es decir, en vista de esta creciente tendencia en cuanto al uso de Internet por parte de los turistas internacionales y de que muchos de estos turistas son ingleses, franceses e italianos, en este trabajo nos proponemos evaluar la calidad de las traducciones inversas de los pares de lenguas español-inglés, español-francés y español-italiano, sobre todo, en lo que respecta a las traducciones de textos pertenecientes al material promocional del segmento turismo de salud y belleza.

2. Compilación de un corpus ad hoc paralelo multilingüe

Para llevar a cabo el análisis contrastivo de las traducciones y evaluar la calidad de las mismas, es necesario que, previamente, se compile un corpus *ad hoc* paralelo multilingüe del segmento del turismo de salud y belleza. De todos es conocido que la finalidad de compilar un corpus de textos reside, principalmente, en recopilar un conjunto de textos en aras de analizar ejemplos de uso real de una lengua dada. Como bien expone Bowker (2002: 43) en su propuesta de definición, el corpus es: «a collection of texts or utterances that is used as a basis for conducting some type of linguistic investigation».

Asimismo, el corpus *ad hoc* se trata de un tipo de corpus que, según describe Aston (1999) se compila: «'on the fly' by the translator in order to investigate a specific problem encountered during a particular translation». Es decir, el objeto principal de compilar un corpus *ad hoc* es de «reunir toda la documentación posible sobre un tema en muy poco tiempo» (Corpas Pastor, 2002: 195). Por último, el corpus paralelo es aquel tipo de corpus que incluye textos en lengua original (LO) y sus traducciones en una o varias lenguas meta (LM). A este respecto, Harris (1988: 8) acuña el término *bitexto* para referirse a los

pares de textos, tanto original como meta, contenidos en un corpus paralelo. Además, este autor considera que un texto original (TO) y su traducción, o texto meta (TM) no son, en realidad, dos textos sino «a single text in two dimensions, each of which is a language».

En nuestro caso, hemos compilado un corpus paralelo multilingüe de textos pertenecientes al material promocional y publicados en la red Internet que versaran sobre el turismo de salud y belleza[82]. Se han extraído textos procedentes de hoteles de 4 y 5 estrellas escritos originalmente en lengua española y sus traducciones en las lenguas inglesa, francesa e italiana; en concreto, se han compilado 111 pares de textos que conforman los bitextos español-inglés, 51 pares de textos de los bitextos español-francés y 22 pares de textos que forman los bitextos español-italiano.

3. Alineación de los bitextos del corpus turismo de salud y belleza

Dentro de la gestión de los bitextos contenidos en un corpus paralelo, la tarea de alineación de los mismos se presenta como esencial para poder así proceder al análisis lingüístico-contrastivo de cada uno de ellos. De hecho, el proceso de alineación es, en palabras de Abaitua (2002: 6) «el proceso que mayor valor añadido aporta a un corpus multilingüe». Alinear un corpus paralelo consiste en reestructurar los textos contenidos en él de forma tal que pueda establecerse una correspondencia entre párrafos, oraciones o palabras.

Rabadán y Fernández Nistal (2002: 76-77) afirman que esta tarea puede incluso llevarse a cabo de forma totalmente manual, en el caso, sobre todo, de corpus de textos de pequeña extensión, aunque advierten que para corpus más extensos es importante y necesario contar con herramientas informáticas que nos ayuden a realizar esta tarea de forma automática.

Por otro lado, en otro estudio (Castillo Rodríguez, 2010) ya advertimos que para alinear corpus de textos de estructura similar, como pueden ser las fichas técnicas de productos sanitarios para diagnóstico *in vitro*, bastaría con emplear aplicaciones informáticas que albergaran un módulo de alineación sencillo, como, por ejemplo, el programa *ParaConc*. No obstante, para textos que ofrezcan una estructura diferente, como pueden ser los textos pertenecientes al material promocional del segmento del turismo de salud y belleza, es necesario que se cuente con un alineador que permita una alineación flexible y semiautomática de las unidades de alineación, como, por ejemplo, *WinAlign*, contenida en el paquete de herramientas de *TRADOS*.

Para poder llevar a cabo con éxito el análisis contrastivo de las traducciones contenidas en nuestro corpus paralelo multilingüe, dado que está integrado por bitextos cuyo contenido del TO difiere estructuralmente del contenido del TM, hemos empleado, en primer lugar, la herramienta *WinAlign*, ya que, gracias a las opciones que ofrece esta herramienta (como editar segmentos, dividirlos, unirlos, entre otros), permite al usuario llevar a cabo una alineación de forma semiautomática.

En segundo lugar, una vez que se han guardado todos los bitextos alineados, éstos se

82 Para la compilación de nuestro corpus paralelo multilingüe hemos seguido la metodología protocolizada propuesta por Seghiri Domínguez (2006) para la compilación de corpus comparables multilingües, aunque aplicable a la compilación de corpus paralelos.

integran en el programa de gestión de corpus paralelos *ParaConc*, el cual, además de ofrecer una alineación sencilla de los bitextos (este proceso puede resultar necesario, ya que, a veces, algunos segmentos han podido *desalinearse*), permite al usuario llevar a cabo un análisis contrastivo de bitextos integrados en un corpus paralelo.

4. Clasificación de errores principales

Una vez que se han alineado los bitextos y analizado el contenido de los mismos con el programa de gestión de corpus paralelo *ParaConc*, como mencionábamos en el apartado anterior, hemos extraído una serie de errores en los bitextos del corpus paralelo que nos han servido para ofrecer una catalogación de los mismos: 1) gramaticales; 2) ortográficos; 3) concordancia; 4) sentido; 5) no traducción; 6) traducción a otra LM; 7) precisión; y, por último, 8) omisión, los cuales pasamos a ejemplificar en los siguientes subapartados.

4.1. Errores gramaticales

Los errores gramaticales detectados en el corpus paralelo se han clasificado, a su vez, en: a) formación del plural del adjetivo en inglés: *essentials oils*, *facials treatments* o *aromatics oils*; b) colocación del adjetivo en inglés: *oil aromatic* o *treatment body*; c) formación del plural irregular en italiano: *sopraccigli*, *tintura di supracciglì*[83] o *cigli*; d) formación del plural en francés: *traitements facials* o *soins facials*.

4.2. Errores ortográficos

Por su parte, los errores ortográficos, que, por otro lado, han sido los numerosos de todos los tipos de errores, se han clasificado de la forma siguiente: a) por omisión de letras: *sopra**[c]**ciglia* (IT), *ma**[s]**sage* (FR), *Fi**[n]**nish* (EN); b) por inclusión de letras: *essent**e**ielles* (FR), *drai**n**nage* (FR), *hydrom**m**assage* (EN); c) por cambio de letras: *draina**j**e* (EN); ***v**ody* (EN), *tir**a**d legs* (EN), *h**i**dromassage* (FR), ***hy**dromassaggio* (IT) *s**u**pracciglì* (IT); d) por omisión o inclusión inapropiadas de tildes: *entra**i**neur* (FR), *compl**e**te* (FR), *th**é**rmale* (FR), *f**á**ccia* (IT); y e) por uso inapropiado de mayúsculas o minúsculas: ***s**hiatsu* (EN), ***uva** rays* (EN), *raggi **Uva*** (IT).

4.3. Errores de concordancia

Los errores de concordancia también han sido abundantes y se han dividido en dos: a) errores de concordancia de género: *eau thermal*, *sauna finlandaise* (FR), *jambe complet* (FR), *pulizia della viso* (IT); y b) errores de concordancia de número: *jambes complète* (FR); *oli aromatico* (IT), *olio essenziali* (IT).

4.4. Errores de sentido

Entre los errores principales de sentido encontrados en los bitextos analizados destacan: *hydrotherapy circuit* (de *masaje subacuático* del TO), *warm water wrap* (de *bañera de agua caliente* del TO), *anti-age treatment* (de *limpieza de cutis* del TO) o *tiring legs* (de

83 Además, este término comporta otro tipo de error, esto es, error ortográfico.

piernas cansadas del TO), *enveloppes culturelles* (de *envoltura corporal* del TO), *top waxing* (de depilación de piernas enteras) o *soin personalisé* (de entrenador personal).

4.5. Errores de no traducción

También se han encontrado errores de no traducción de determinadas secuencias terminológicos, es decir, se ha optado por dejarlas en LO (española): *tinte de pestañas*, *Rayos UVA* o *limpieza de cutis*, (en TM ingleses); *sauna finlandesa*, *baño turco* o *piernas cansadas* (en TM franceses); *tratamientos corporales* o *tratamientos faciales* (en TM italianos).

4.6. Errores de traducción a otra LM diferente

Otro tipo de errores, aunque menores, lo constituyen los errores de traducción a otra lengua diferente a la del TM en cuestión: *oil massage*, *massage with oils* o *chocolate massage* (traducción realizada al inglés en TM franceses); *Lymphdrainage* (traducción realizada al alemán en TM italianos) o *bany turc* (traducción realizada al catalán en TM franceses).

4.7. Errores de precisión

Los errores de precisión más importantes encontrados en los bitextos analizados son: *wine therapy treatment* (de *tratamiento corporal con cava* del TO), *bains turcs* y *bagni turchi* (de *baño turco*, en singular, del TO).

4.8. Errores de omisión

Los errores de omisión se han dividido en dos subtipos, a saber: a) errores de omisión total: *bain turc* (UT omitida en TM franceses) y *huiles essentielles* (UT omitida TM franceses); b) errores de omisión parcial: *oil* (de *aceites esenciales* del TO), *enveloppement* (de *envoltura corporal* del TO), *circulation* (de *circulación sanguínea* del TO) o *wrap* (de *envoltura corporal* del TO).

5. Conclusiones

En esta investigación se ha ofrecido una catalogación de los principales errores de traducción de la terminología contenida en un corpus paralelo multilingüe, integrado por TO en lengua española y TM en inglés, francés e italiano. Como hemos ofrecido en la parte introductoria, hoy en día, España es uno de los destinos vacacionales preferidos por la mayoría de los turistas internacionales, especialmente, los procedentes del Reino Unido, Francia e Italia. Además, en la actualidad, existe una tendencia al alza por parte de estos turistas en cuanto al uso de la red Internet para consultar, reservar e, incluso, contratar los servicios turísticos en España.

No obstante, y como se ha reflejado a lo largo de estas páginas, muchos de estos textos muestran errores de traducción, muchos de los cuales constituyen errores muy graves que, incluso, llegan a confundir al turista internacional que contrata sus servicios turísticos confiando en la veracidad de los mismos. Los errores que inducen a confusión y que, por tanto, dejan en mal lugar a España son los errores de sentido y de precisión, ya que, por ejemplo, el turista puede pensar que está contratando servicios de circuitos de hidroterapia y de tratamientos antiedad, cuando la realidad muestra que lo que oferta el establecimiento de salud y belleza (en nuestro caso, hoteles de 4 y 5 estrellas) son masajes subacuáticos y tratamientos de limpieza de cutis, respectivamente, en el caso de los errores de sentido, así como puede hacerse a la idea de que el establecimiento en sí dispone de varias salas de baño turco, cuando, en realidad, sólo hay uno, en el caso de los errores de precisión.

Estos tipos de errores, así como el resto de errores que se han mostrado en la clasificación de los mismos podrían haberse evitado con una profunda revisión del texto traducido, aunque también con la compilación de corpus comparables de textos especializados en el segmento del turismo de salud y belleza en las distintas lenguas incluidas en el estudio. De esta forma, se evita también la mala imagen que puedan mostrar los establecimientos turísticos de España y que nuestro país quede, por tanto, a ojos del turista internacional, en una posición inversamente proporcional al lugar que ocupa como destino vacacional preferido mundialmente.

6. Referencias bibliográficas

Abaitua Odriozola, J. (2002). Tratamiento de corpora bilingües. En M.A. Martí Antonín y J. Llisterri Boix (Eds), *Tratamiento del lenguaje natural* (pp. 61-90). Barcelona: Universidad Autónoma de Barcelona. 61-90.

Aston, G. (1999). Corpus use and learning to translate. *Textus, 12*, 289-314.

Bowker, L. (2002). *Computer-Aided Translation Technology. A Practical Introduction*. Ottawa: University of Ottawa Press.

Castillo Rodríguez, C. (2010). La dificultad del proceso de alineación para el estudio contrastivo de traducciones: un caso práctico con corpus paralelo multilingüe. En R. López-Campos Bodineau, C. Balbuena Torezano y M. Álvarez Jurado (Eds.), *Traducción y modernidad. Textos científico-técnicos, jurídico-socioeconómico, audiovisuales y de interpretación* (pp. 239-250). Córdoba: Servicio de publicaciones de la Universidad de Córdoba.

Corpas Pastor, G. (2002). Traducir con corpus: de la teoría a la práctica. En J. García Palacios y M.T. Fuentes Morán (Eds.), *Texto, Terminología y Traducción* (pp. 189-226). Salamanca: Almar.

Harris, B. (1988). Bi-text, a New Concept in Translation Theory. *Language Monthly, 54*, 8-10.

Seghiri Domínguez, M. (2006). *Compilación de un corpus trilingüe de seguros turísticos (español-inglés-italiano): aspectos de evaluación, catalogación, diseño y representatividad*. Málaga: Servicio de publicaciones de la Universidad de Málaga.

VOGELER RUIZ, C. Y E. HERNÁNDEZ ARMAND. (2000). *El mercado turístico. Estructura, operaciones y procesos de producción*. Madrid: Centro de Estudios Ramón Areces, S.A.

Deictic neutralization and overmarking: demonstratives in the translation of fiction (English-Catalan)

Maria Josep Cuenca
Josep Ribera
Universitat de València

Abstract

Demonstratives exhibit a complex behavior that translation highlights. In this paper two translation strategies are described in a corpus of English fiction translated into Catalan, namely, neutralization (i.e., deletion or translation by a non-deictic unit) and overmarking (i.e., new demonstratives added to the target text). Corpus analysis shows differences regarding the use of demonstratives in both languages that can only be explained by considering syntactic and discourse factors that go well beyond the classical analysis of deixis. The two strategies analyzed are mainly syntactically conditione. Neutralization is associated with the anaphoric functioning of non-situational demonstratives and implies the underspecification of the subjective perspective of the source text. Conversely, overmarking is associated with a different behaviour of forms such as pronouns and possessives in the two languages and implies the introduction of the speaker's subjectivity.

Keywords: deixis, demonstratives, translation strategies, English, Catalan,

Resumen

Los demostrativos muestran un comportamiento complejo que la traducción ayuda a destacar. Partiendo de un corpus de novela en inglés traducida al catalán, en este trabajo describimos dos estrategias de traducción: la neutralización (esto es, la elisión o traducción del demostrativo por un elemento no deíctico) y el sobremarcaje (la adición de demostrativos en el texto meta). El análisis de corpus muestra diferencias en cuanto al uso de los demostrativos en las dos lenguas que solo pueden explicarse teniendo en consideración factores sintácticos y discursivos que van más allá del análisis clásico de la deixis. Las dos estrategias analizadas están condicionas sintácticamente sobre todo. La neutralización se relaciona con el funcionamiento anafórico de los demostrativos no situacionales e implica la infraespecificación de la perspectiva subjetiva del texto origen. En cambio, el sobremarcaje se relaciona con el diferente funcionamiento de formas como los pronombres y los posesivos en ambas lenguas e implica introducción de la subjetividad del hablante.

Palabras clave: deixis, demostrativos, estrategias de traducción, inglés, catalán.

1. Introduction

Demonstratives are deictic units typically analysed in situational terms, that is, as linguistic items that point to elements of the situational ground of utterance, thus marking space distance with regard to the deictic origin (cf. Bühler, 1934). However, demonstrative expressions exhibit a complex behaviour that translation highlights. First, not all demonstratives are situational (see Himmelmann, 1996; Diessel, 1999); in fact, both in oral and written texts, non-situational uses outnumber the cases in which demonstratives indicate proximity or distance with respect to the addressor[84]. Among non-situational ones, text deictics (i.e., demonstratives with an antecedent trigger in the linguistic context) are very frequent and have an important cohesive role as devices that help to maintain discourse reference together with other phoric markers as third person pronouns (Halliday & Hasan, 1976). As several authors point out (e.g. Ariel, 1990; Cornish, 1999; Strauss, 2002), demonstratives represent an intermediate point in a continuum of referential markers between zero-anaphora and unstressed pronouns, at the one end, and definite expressions, at the other end. This fact is related with two other features that corpus analysis shows (cf. Cuenca, 2010; Cuenca & Ribera, 2010): (i) non-situational demonstratives are frequently neutralised (i.e., translated by a non deictic unit or deleted) in translation, and (ii) new demonstratives, mainly non-situational ones, show up in the target text (that is what we call *deictic overmarking*).

The main aim of this paper is to describe neutralization and overmarking as prominent strategies in the translation of English fiction into Catalan. This research is based on a corpus consisting of a hundred pages of Donna Leon's *Through a Glass Darkly* and one hundred pages of Jed Rubenfeld's *The Interpretation of Murder* and the translation of the texts into Catalan.[85]

2. General description of the data

English and Catalan exhibit different deictic systems, as shown in Table 1, which includes the forms and number of tokens identified in the corpus.

84 For instance, situational demonstratives are less frequent in our English corpus (17.2%) than non-situational ones (82.8%), and among these text deictics represent 74% (362 cases) of all demonstratives (489 cases).

85 We are indebted to Edicions 62 for having offered us the electronic versions of the Catalan translations.

Table 1. Deictic systems and forms in the corpus

Deictic space		Proximal		Distal	
		D1 ('this')	D1/2 ('that-near')	D2 ('that-far')	
English	forms	this these	that those		
	#	236 (48.3%)	253 (51.7%)		489
Catalan	forms	aquest, aquesta aquests, aquestes això així		aquell, aquella, aquells, aquelles allò	
	#	277 (58.6%)		196 (41.4%)	473

English demonstratives are organized in a binary system: addressor space (D1) vs. non-addressor space (D2) (see Halliday & Hasan, 1976; Lyons, 1977; Stirling and Huddleston, 2002, among others). Catalan exhibits a different binary deictic system: proximal or addressor and addressee space (D1 and D1/2) vs. distal or 3rd person space, i.e., 'that-far' sphere (Hualde, 1992; Badia i Margarit, 1994).[86]

However, the translation of the demonstratives in our corpus does not always correspond to differences in the deictic systems. In order to describe the strategies followed by the translators, the English demonstratives *this/these* and *that/those* and their counterparts in the translated versions have been analysed. Four strategies regarding demonstratives have been identified: a) maintenance, b) shift, c) neutralization, and d) overmarking.

The general results shown in Table 1 put forward that there is no one-to-one correspondence between English demonstratives and Catalan ones. Catalan demonstratives only translate English demonstratives in 286 cases (58.5%). This is because English demonstratives are frequently neutralized, that is, do not always result in a Catalan demonstrative or deictic (D > non-D, including zero translation), as in (1), where *this* is translated by *ho* (a neutral unstressed pronoun roughly equivalent to *it*):

(1)

After a brief pause, Brunetti said: 'You're right, but we'd better do *this* in person.'. (Glass, 8)	Després d'una petita pausa, Brunetti va dir: — Tens raó, val més que *ho* fem en persona.

When English demonstratives are not neutralized, the deictic centre can be either maintained or shifted. Maintenance implies that the deictic centre remains unchanged in the target text (D1>D1, D2>D2).

86 Medieval Catalan and some of its nowadays dialects exhibit a ternary deictic system with different forms corresponding to D1, D1/2 and D2.

(2)

The coroner said he could by no means allow it: in cases of homicide, the decedent's body must by law be taken into custody for an autopsy. 'Not *this* body,' answered Banwell. (Murder, 27)	El forense li va dir que no ho podia permetre de cap manera: en casos d'homicidi, el cos del difunt quedava custodiat fins que li practicaven l'autòpsia. —Doncs *aquest* cos no —va dir en Banwell

Deictic shift occurs when the deictic centre is reversed (D1>D2; D2>D1).

(3)

'It's Wednesday,' the man said. 'So there'll be liver. It's good.' . 'With polenta?' Brunetti asked. 'Of course,' the man said, pausing to glance aside at *this* man who spoke Veneziano yet who had to ask if liver was served with polenta. (Glass, 93)	Avui és dimecres, o sigui que hi haurà fetge. El fan bo. —¿Amb polenta? —va demanar Brunetti. —És clar —va respondre l'home, aturant-se per donar un cop d'ull a *aquell* ('that') paio que parlava veneziano però havia de preguntar si el fetge se servia amb polenta.

On the other hand, demonstratives in the target language are used to translate English non-demonstrative units. It is possible to identify 184 demonstratives in Catalan showing deictic overmarking, i.e. an additional deictic marking with respect to the original, as in (4):

(4)

'Come and meet my wife' Brunetti followed him over to *the* woman.... (Glass, 45)	Vingui a conèixer la meva dona. Ell el va seguir fins a *aquella* ('that') dona...

As for frequency, neutralization is the most frequent strategy (196 cases out of 489 English demonstratives, 40.1%), especially with non-situational deictics. Maintenance, which is predominant with situational demonstratives but only the second preferred strategy with non-situational ones, applies in 155 cases (31.7%), whereas shift occurs in 138 cases (28.2%). Overmarking (184 cases, 38.9%) is almost as prominent as neutralization in our corpus.

The following sections will be devoted to neutralization and overmarking, contrary translation strategies implying the underspecification and the introduction, respectively, of subjective and intersubjective values in the narration.

3. Neutralization

Neutralization occurs when a demonstrative is deleted (5) or substituted by a non-deictic unit (6), mainly a definite article or a pronoun.

(5)

'All they make is *that* tourist crap. You know, the porpoises leaping up out of the waves. And toreadors.' (Glass, 26)	—Només fan rampoines per a turistes. Ja saps què vull dir: balenes blanques saltant per sobre les ones. I toreros.

(6)

The horse in question belonged to a carriage that had just emerged from a construction site on Fortysecond Street [...]. The man driving *this* carriage was superbly attired (Murder, 141)	El cavall en qüestió era d'un carruatge que acabava de sortir d'una obra situada al carrer Quaranta-dos [...]. L'home que conduïa *el* ('the') carruatge anava de vint-i-un botons

As we can see in the previous examples, neutralization implies a loss of deictic force and sometimes also the empathetic nuance, which affects the implication of the character or the narrator in the narration (cf. Ribera, 2007).

Neutralization is syntactically conditioned. Since Catalan is a pro-drop language, a deictic pronoun in a subject position tends to be deleted.

(7)

'I won't hear another word of this ludicrous slander. Now go home. You are not fit to be in your office in this state. Get some rest. *That*'s an order.' (Murder, 196)	I no penso sentir més calúmnies sense solta ni volta. I ara vagise'n a casa. En aquestes condicions no pot treballar. Descansi una mica. És una ordre.

Another syntactic context that favours neutralization is when a deictic-NP or PP can be pronominalised by means of a 3rd person pronoun.

(8)

'He says that Marco doesn't love me and that he married me for my money.' She did not look at him as she said *this*. (Glass, 66)	—Diu que en Marco no m'estima i que només es va casar amb mi pels diners. —*Ho* (*'it*-unstressed') va dir sense mirar Brunetti a la cara.

A change in the overall construction also increases the possibility of neutralization.

(9)

'I wondered if you had a moment,' Vianello said, using the familiar tu and not referring to Brunetti as 'sir', thus increasing the likelihood that *this* would be an informal conversation. (Glass, 2)	—Em preguntava si tindries un moment —va dir Vianello adreçant-se-li amb el tu familiar i sense afegir «senyor», cosa que augmentava les probabilitats de tenir-hi una conversa informal.

Finally the presence of two or more deictics in the same linguistic context also enhances the possibility of neutralizing one of them.

(10)

'He loves me. He'd never hit me. He'd cut off his hand first.' Strangely enough, Brunetti believed *this*, too. 'I see,' he said, and then added, *'That* must make *this** even more painful for you.' (Glass, 75)	Ell m'estima. No em pegaria mai. Abans es tallaria la mà. —Per estrany que fos, Brunetti també va creure *aquest* comentari. —Ja ho veig. *Això* encara *ho* deu fer tot plegat més dolorós, per a tu.

In conclusion, Catalan shows a tendency to avoid deictic marking in syntactic contexts where a demonstrative could be interpreted as too focal or somehow emphatic.

4. Overmarking

Overmarking is the addition of a demonstrative in a context where the source language included no deictic. This strategy, contrary to neutralization, adds deictic force to the source text. It usually introduces a character or the narrator point of view.

(11)

Believe me, he hates De Cal more than I do, so if he said *the* old bastard didn't start it, then he didn't.' (Murder, 98)	Creguin- me, ell odia De Cal més que no pas jo. Si ell diu que *aquest* malparit no va començar la baralla, és que no la va començar.

Overmarking often affects an English phoric pronoun, especially *it*.

(12)

Both Brunetti and Vianello said that the meal had' been excellent, and Navarro seemed more pleased than the waiter to hear them say *it*. (Glass, 101-2)	Brunetti i Vianello van dir que havia estat un dinar excel.lent, i Navarro semblava més complagut que el cambrer amb *aquest* comentari.

The alternation 3rd person pronoun / demonstrative can be accounted for, on the one hand, by the different behaviour of Catalan strong pronouns and possessives, which do not usually refer to inanimates and are less used with animates than in English, and, on the other hand, by the lack of a stressed pronoun equivalent to *it*.

Moreover, pronominalisation, possessive reference or ellipsis would often imply an unnatural construction in Catalan or could create ambiguity in the establishment of co-reference.

(13)

He bent and picked her up [...] Not-quite-forgotten habit slipped into operation and he put her over his shoulder, noticing the insubstantiality of *her*. (Glass, 83)	Es va ajupir i la va agafar [...]. Aleshores va posar en pràctica un hàbit no del tot oblidat i se la va repenjar a l'espatlla, notant la insubstancialitat d'*aquell* cos.

In (13) a strong pronoun "la insubstancialitat d'ella" would not be sound natural, whereas a possessive ("la seva insubstancialitat") would be ambiguous (the possessor could be either the baby or the main character). The demonstrative, often followed by a synonym or a hyperonym, avoids ambiguity.

Articles and other determiners are also frequently translated by a demonstrative. The repetition of the antecedent acting as a topic, especially when the NP includes an adjective or another complement, enhances this correspondence.

(14)

And the gentleman had a way of moving, a fluidity when he swung the young lady into the carriage [...] He did not resent *the* young gentleman, and he liked Betty, the maid, better than he liked *the* angelic young lady (Murder, 138)	I el jove tenia una manera de moure's, una fluïdesa a l'hora de balancejar la dama cap a l'interior del vehicle [...] *Aquell* jove no li va fer gens de tírria i, a més, a ell li agradava més la Betty, la cambrera, que no pas *aquella* dama angelical.

It is noteworthy that the translation of *such* systematically implies the introduction of a demonstrative in our corpus.

(15)

A less magnanimous man would have been crowing about *such* a thing, rubbing the invitation in the others' noses (Murder, 210)	Un home menys magnànim hauria presumit d'un fet com *aquell* i els hauria refregat la invitació pels nassos

A change in the overall structure can also favour overmarking. This is especially the case when an elliptic structure is translated by a fuller equivalent and when the Catalan construction is a free relative or corresponds to "*això/allò de...*" ('this/that of').

(16)

'Done,' said the mayor. 'You can have the most seasoned man on the force.'	—Fet —va dir l'alcalde—. Pot triar l'home amb més experiència del cos.
'Exactly what I don't want,' replied the coroner. (Murder 22)	—*Això* és precisament el que no vull —va dir el forense—.

(17)

it's almost a point of pride with her never to be satisfied with their performance.' (Glass, 35)	Hi ha gairebé un punt d'orgull, en *això* de no estar mai satisfeta amb el rendiment dels seus alumnes.

The addition of a demonstrative is also increased by a change in the structure resulting in a dislocation. The dislocated component includes a deictic, and a weak pronoun represents its function in the VP.

(18)

it's the same clay I saw up in Miss Riverford's room, I'm sure of *it*.' (Murder, 95)	és la mateixa argila que vaig veure a l'habitació de la senyoreta Riverford. D'*això* n'estic segur.

The construction in (18) is very frequent in Catalan speech and thus its use in translation makes dialogues more natural.

Finally, some English connectives are translated by complex expressions containing a text deictic whose deictic force is more or less weakened.

(19)

By comparison with the gleaming Vanderbilt mansion, the Astors' fine old brick townhouse had suddenly looked small and drab. *Therefore* Mrs Astor unceremoniously razed it and built herself a double-sized French chateau. (Murder, 68)	Comparada amb la resplendent mansió dels Vanderbilt, la bonica casa antiga de maons dels Astor havia quedat petita i apagada. *Per això* la senyora Astor l'havia fet ensorrar sense miraments i s'havia fet construir un château francès el doble de gran.

In conclusion, Catalan shows a tendency to add a demonstrative in several contexts. Overmarking often implies a change in the narration by foregrounding a character or introducing the narrator's point of view.

5. Conclusions

English and Catalan demonstratives exhibit differences in use and frequency that cannot be attributed to differences in their respective deictic systems. Translation often implies the underspecification of the subjective perspective of the source text (neutralization), or the introduction of the speaker's subjectivity (overmarking).

Neutralization is the most frequent strategy when translating non-situational demonstratives. This strategy is associated with their anaphoric functioning: demonstratives alternate with other phoric processes, such as ellipsis or 3rd person pronouns, and are dispreferred in several syntactic contexts.

Overmarking is also very frequent in the translation of English fiction into Catalan. It adds deictic force to the source text and usually introduces the narrator's or a character's point of view. This strategy is associated with a different behaviour of pronouns, possessives and forms as *such* in the two languages. Catalan makes a more extensive use of demonstratives in phoric contexts in order to retrieve a topic and also in a number of constructions, such as dislocations, free relatives or connectives.

Corpus analysis highlights interesting differences regarding English and Catalan demonstratives that can only be explained by considering syntactic and discourse factors that go well beyond the classical analysis of deixis.

References

Ariel, M. (1990). *Accessing Noun-phrase Antecedents*. London: Routledge.

Badia i Margarit, A. M. (1994). *Gramàtica de la llengua catalana. Descriptiva, normativa, diatòpica, diastràtica*. Barcelona: Enciclopèdia Catalana.

Bühler, K. (1934). *Sprachtheorie: die Darstellungfunktion der Sprache*. Stuttgart: Gustav Fischer [1982].

Cornish F. (1999). *Anaphora, Discourse and Understanding. Evidence from English and French*, Oxford: OUP.

Cuenca, M. J. (2010). Díctics espacials i gramàtica de les narracions orals», *Estudis Romànics, XXXII*, 101-123.

Cuenca, M. J. & Ribera, J. (2010). Disappointing deictics. A contrastive approach to non-situational uses of demonstratives in written narratives», Paper presented at the *43rd Annual Meeting of the Societas Linguistica Europaea*. Vilnius, Latvia, September, 2-5.

Diessel, H. (1999). *Demonstratives. Form, Function, and Grammaticalization*. Amsterdam/Philadelphia: John Benjamins.

Halliday M. A. K. & Hasan, J. (1976). *Cohesion in English*. London: Longman.

Himmelmann, N. P. (1996). Demonstratives in narrative discourse: a taxonomy of universal uses. In B. Fox (Ed.). *Studies in Anaphora*. Amsterdam: John Benjamins, 205-254.

Hualde, J. I. (1992). *Catalan*. London: Routledge.

Lyons, J. (1977). *Semantics, II*. Cambridge: Cambridge University Press.

Ribera, J. (2007). Text deixis in narrative sequences. In J. Valenzuela, A. Rojo & P. Cifuentes (Eds.), *Cognitive Linguistics: from words to discourse. International*

Journal of English Studies (IJES) Special Issue, *7* (1), 149-168. Murcia: Servicio de Publicaciones de la Universidad de Murcia.

STIRLING, L. & HUDDLESTON, R. (2002). Deixis and anaphora. In R. Huddleston & G. K. Pullum (Eds.) *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press, chap. 17, 1451-1554.

STRAUSS, S. (2002). *This, that*, and *it* in spoken American English: a demonstrative system of gradient focus», *Language Sciences*, *24*, 131-152.

CORPUS REFERENCES

DONNA LEON, *Through a Glass Darkly*. London: Arrow Books, 2006. Catalan translation: *Cristall enverinat*. Barcelona: Edicions 62, 2006. Translators**:** Anna Mauri i Batlle, Anna Roca Puntí and Joan Puntí i Recasens.

JED RUBENFELD, *The Interpretation of Murder*. London: Headline Review, 2006. Catalan translation: *La interpretació del crim*. Barcelona: Edicions 62, 2007. Translator: Albert Torrescasana Flotats.

Métodos de la lingüística de corpus aplicados a los estudios descriptivos de traducción

Rosa Currás Móstoles

Universidad Católica de Valencia

Miguel Ángel Candel-Mora

Universidad Politécnica de Valencia

Resumen

La comparación lingüístico-textual de un solo texto traducido con su original es una técnica reciente, que sin embargo, debe constituir la base imprescindible sobre la que realizar el comentario crítico de los textos traducidos y sacar conclusiones empíricamente fundamentadas acerca de lo que implica globalmente lo que llamamos traducción literaria.

El objetivo de este trabajo consiste en demostrar la consolidación del método de trabajo interdisciplinar, en este caso por medio de la combinación de métodos de análisis procedentes de la Lingüística de Corpus y de los Estudios Descriptivos de Traducción para el análisis de traducciones teatrales.

Entre las características mas destacables cabe mencionar la adaptación de los métodos tradicionales de alineación al método más centrado en el teatro, así en lugar de alinear por segmentos de traducción se procede a alinear por réplicas, como unidad mínima de significado.

Palabras clave: diseño de corpus; estudios descriptivos de traducción; traducción teatral

Abstract

The linguistic and textual comparison of a single translated text with its original version is a recent technique, which however, must be the essential basis on which to perform a critical analysis of translated texts and draw evidence-based conclusions about the overall meaning of what we call literary translation.

The aim of this paper is to demonstrate the consolidation of the interdisciplinary working method, in this case through a combination of analysis methods from Corpus Linguistics and Descriptive Translation Studies towards the analysis of drama translations.

Among the most remarkable features is the adaptation of the traditional alignment method focused on the theater play format: instead of the usual translation segment alignment it is necessary to align texts at the replica level, as the minimum unit of meaning.

Keywords: corpus design; Descriptive Translation Studies; drama translation

1. Introducción

Los Estudios de Traducción se han convertido en una disciplina académica con gran reconocimiento, pero la poca literatura existente sobre la traducción teatral se ha estudiado como un sub-apartado dentro de la traducción literaria, Por otra parte, este abandono que muestra el aparato teórico no se corresponde con la práctica, lo cual se ilustra con la evolución que ha experimentado el teatro clásico gracias al denominado *teatro de director*, el cual proporciona un amplio corpus de textos traducidos y retraducidos.

La razón aducida para la escasa investigación en la traducción teatral es principalmente la posición marginal del teatro en prácticamente toda la teoría crítica y social contemporánea, si bien una mayor dedicación a este campo podría aportar mucho a la teoría de la traducción, a la vez que proporcionaría medios excelentes para realizar análisis más detallados sobre la función del texto escrito dentro del proceso dinámico de dotar al teatro de significado (McAuley, 1995).

No obstante, la incorporación de técnicas de estudio y metodologías propias de la lingüística de corpus aplicadas al estudio de obras teatrales pueden suponer un giro en la tendencia observada hasta el momento

El objetivo final en los Estudios de Traducción debería ser la conjunción de fuerzas entre las distintas disciplinas para contribuir a la práctica traductora y por consiguiente a su posterior análisis donde se valore su adecuación teniendo en cuenta la confluencia de circunstancias en el hecho traductor. Amparo Hurtado visiona un futuro investigador en traducción múltiple, interdisciplinario y empírico: múltiple porque hay cabida para todos los enfoques, interdisciplinario, puesto que el estudio de la Traductología requiere el contacto con otras disciplinas, y empírico y experimental para que por medio de la recogida de datos de modo que las propuestas se transformen en principios y modelos validados (Hurtado, 2001:632).

El objetivo de este trabajo consiste en demostrar la consolidación del método de trabajo interdisciplinar, en este caso por medio de la combinación de métodos de análisis procedentes de la Lingüística de Corpus y de los Estudios Descriptivos de Traducción para el análisis de traducciones teatrales: "The two main sources of influence to Corpus-based Translation Studies are Corpus Linguistics and Descriptive Translation Studies" (Laviosa, 2000: 5).

La primera parte del trabajo hace un breve recorrido por las particularidades del texto teatral y su traducción. En segundo lugar, se describe la aportación de la lingüística de corpus y la metodología para la elaboración de un corpus específicamente orientado al estudio de una traducción teatral.

2. Características de la traducción teatral

En España, la presencia del teatro traducido sigue siendo mayor que el de producción original. Según Merino (1995), no se puede comprender la historia del teatro en España sin

la presencia constante y significativa de obras extranjeras, la mayoría de las cuales están escritas originalmente en inglés. La paradoja es que tal abundancia de teatro traducido no haya generado corpus alguno de reflexión teórico-critica, aunque sí considera importante la diferenciación entre dos tipos de traducción. La traducción teatral prescinde de textos concretos y atiende a una especulación sobre la especificidad del texto teatral per se, mientras que las traducciones teatrales atienden a lo particular, a uno o varios textos traducidos, títulos determinados y traductores con nombre y apellido (Santoyo, 1995: 16).

Por otra parte, otros autores (Bassnett, 1991; Espasa, 2001) no destacan la escasez de estudios sobre este campo, sino la falta de sistematización, por lo cual exige abandonar el estudio de la traducción teatral como un fenómeno contingente y pasar a la investigación sobre la especificidad de la traducción teatral, o que la poca literatura existente se centra exclusivamente en la problemática de la traducción del texto dramático para la representación.

En la práctica, la existencia de múltiples variedades de traductor de obras de teatro ha contribuido a una falta de consenso. Entre las distintas posibilidades están: el traductor teatral, que puede trabajar por encargo para un director particular o para una compañía; puede ser un traductor contratado para una editorial; en ocasiones el traductor es el mismo dramaturgo o uno de los actores, e incluso se da el caso de que es el dramaturgo quien -sin conocer la lengua origen- produce un texto basándose en una traducción literal hecha por otra persona (McAuley, 1995). Esto plantea, según este mismo autor, dos campos de estudio muy interesantes, por una parte el estudio del grado de adaptación y reescritura que se produce en el trasvase, y por otra la investigación sobre la tendencia tan extendida de elaboración de un texto partiendo de traducciones ya existentes.

En lo que respecta a la investigación en el campo de la traducción teatral, la tipología es la siguiente: aquellos que provenientes de otros campos de estudio, como la filología, la literatura comparada o la lingüística; y por otro lado, los profesionales del campo dramático (actores, directores, técnicos, dramaturgos, traductores dramáticos) que reflexionan sobre su quehacer dentro del teatro, si bien a estos últimos se les reprocha que ignoren las aportaciones de la semiología del teatro, y que se centren en la representación y los aspectos no-verbales de la misma en detrimento del texto literario.

Drama Translation es, por tanto, un término muy amplio dentro del cual se aglutinan múltiples variedades, entre las cuales se incluyen no sólo la traducción de una obra de teatro de una lengua a otra o de la página al escenario, sino también traducción entre culturas, entre diferentes públicos, medios, periodos, estilos, géneros o grupos de acción (Gostand, 1980: 1-9).

3. Lingüística de corpus aplicada a la traducción

Tomando como punto de partida la consideración de los estudios de corpus aplicados a la traducción como línea de trabajo consolidada y con entidad propia (Laviosa, 2002:1) y el

hecho de que en los últimos años, las mejoras en velocidad de proceso de datos, capacidad de almacenamiento y disponibilidad de textos en formato electrónico han potenciado el auge y la consolidación de los métodos de la lingüística de corpus en la investigación en traducción y su adaptación al medio de estudio de la Traducción, como demuestran por ejemplo las diferentes pautas para el diseño de corpus, y las diferentes técnicas de explotación de corpora.

Aunque al hablar de corpus se hace referencia a una recopilación de textos, o a una recopilación de fragmentos de una lengua, un corpus diseñado para el estudio de la traducción puede estar formado únicamente por dos obras, una en la lengua origen y otra en la meta (Laviosa, 2002:33).

La literatura sobre estudios de traducción basados en corpus divide tradicionalmente la orientación del corpus hacia la docencia, la investigación y el ejercicio profesional. Por su parte, Laviosa (2002:34) propone una clasificación de las aplicaciones de corpus para la traducción en la que incluye la línea de estudios contrastivos propuesta en este trabajo. Para esta autora, las tres grandes líneas de trabajo giran en torno al estudio de la traducción como proceso y como producto, la formación de traductores, y los estudios de lingüística contrastiva.

La validez de la comparación lingüístico-textual de un texto traducido con su original viene determinada porque supone interpretación, que no necesariamente evaluación, y puesto que la interpretación del traductor implica decisiones "en el mejor de los casos intersubjetivas", se puede inferir la toma de decisiones sobre la interpretación que éste hace sobre el material lingüístico (Sánchez García, 1996: 357-378). Por otra parte, el panorama sobre los estudios empíricos de traducción produce una disparidad entre la naturaleza misma de los análisis sobre traducciones específicas, ya que los estudios basados en corpus extensos colisionan con la obligada minuciosidad de la comparación entre un texto original y su traducción (Merino, 1994).

El objeto de estudio principal está compuesto por el texto origen en inglés (Bolt: 1990), y el texto meta: la traducción que realizó Luis Escobar en 1967. Con la finalidad de facilitar el análisis, se ha elaborado un corpus paralelo por medio de la alineación de todos los segmentos (tanto el diálogo como el marco teatral) de ambas obras.

Este estudio se basa en un análisis contrastivo a nivel microestructural cuyo punto de partida es el marco contextual de la cultura origen. La dimensión extratextual del teatro como espectáculo impide su tratamiento del mismo modo que el resto de géneros por la complejidad del contexto que la rodea, por lo que además del texto traducido, se tendrán en cuenta el resto de elementos que forman parte del proceso teatral, puesto que afectan al texto impreso y al modo en que éste ha sido traducido y publicado.

Züber-Skerritt planteó por primera vez la diversidad de códigos no lingüísticos que operan en el trasvase y que se suman al campo de la traducción dramática, y que suponen la apertura de nuevos campos de estudio y de dificultades añadidas para el traductor, entre los que se encuentran los aspectos culturales, los no verbales y los de escenificación (1984: 3-11).

Con respecto a la segmentación del texto para la realización del análisis, es necesario que aunque sea realizada con criterios conceptuales, debe producir "manageable chunks" (Toury, 2004) que funcionen en la comparación textual como unidades de traducción. A este respecto, y debido a la especificidad propia del género teatral se ha delimitado la unidad de análisis microtextual en la réplica. Esto permite un acercamiento al texto traducido y posibilita el análisis comparativo con el original. La réplica, en una obra teatral constituye la unidad mínima del campo dramático, es específica del género textual y su identificación resulta sencilla, ya que aparece indicada con el nombre del personaje precediendo a la acotación escénica y al discurso correspondiente, y tanto el marco o didascalia como el diálogo o texto están indicados tipográficamente en una réplica mediante cursiva, negrita, paréntesis o corchetes. Tampoco se ha prestado atención al componente estructural de la obra de teatro objeto de estudio, puesto que la acción se desarrolla en dos actos y ni el autor ni el traductor utilizan la subdivisión en escenas. No obstante, y se ha utilizado el término escena para delimitar fragmentos de los actos en función del desarrollo de la acción, y que como ocurre en este caso, no están reflejados gráficamente en el texto impreso.

El estudio del contexto de la cultura meta y de la dramaturgia del autor, como consecuencia de la profunda imbricación de la obra en dicha cultura ha resultado determinante en la identificación y categorización de los referentes culturales que asoman a lo largo de la obra. Pese a la existencia de numerosas clasificaciones de éstos en la literatura sobre traducción, se han encontrado dificultades en la aplicación de estas clasificaciones por ser demasiado generalistas y globales. Es por ello que una de las características definitorias de la propuesta de análisis señalada en este trabajo es que no se ha efectuado con carácter prescriptivo, sino que las características especiales de la obra de teatro y su conformación según ciertas coordenadas espacio-temporales conducen a un modelo de clasificación ad hoc, de esta manera se fundamenta su elección en un texto real traducido y en el contexto en el que se utiliza, logrando así la cobertura de todos los aspectos específicos del texto que se pretende analizar.

En la parte del análisis confrontado de las dos obras, los parámetros macrotextuales que rigen en la cultura receptora inciden directamente en la toma de decisiones del traductor. Así, el historial interlingüístico en la traducción de ciertos referentes culturales como nombres propios, unidades monetarias, o medidas es uno de los determinantes en la traducción.

En segundo lugar, la función inicial que el autor del texto original otorga a un referente cultural en el texto origen no constituye de por sí un motivo suficiente para su traducción, hecho que se demuestra con la inconsistencia en aplicación de técnicas de traducción de referentes pertenecientes a la misma categoría.

La variedad de los resultados obtenidos en el análisis prueba que las estrategias de traducción de los referentes culturales es múltiple y amplia, y que la intervención del traductor posibilita la realización de una gradación en cuanto a las técnicas de traducción adoptadas. No obstante, la disparidad existente entre el conocimiento del lector medio de la cultura origen y el del lector de la cultura meta no ha sido motivo suficiente para una

mayor intervención traductora que favorezca la explicitación de toda la carga informativa contenida en el referente cultural, principalmente cuando esta carga connotativa es inexistente u opaca en la lengua meta. En algunas ocasiones, sin embargo, se aprecia un esfuerzo por parte del traductor por trasvasar el mensaje atendiendo a la funcionalidad del referente en la cultura origen, mientras que en otras se proporciona al lector de la cultura receptora un equivalente generalizado del referente aparecido en la cultura origen, con la consiguiente pérdida de información.

4. Conclusiones

Las principales dificultades que se plantean en traducción están relacionadas con la identificación de los aspectos espacio-temporales específicos de la obra. El análisis confrontado a nivel de la réplica entre el texto origen y el texto meta revela datos de interés en cuanto al tratamiento de los referentes culturales propios de la obra.

Como consecuencia, el conocimiento requerido para el traductor literario debe ser de dos tipos, lingüístico y extralingüístico, y este último puede ser subdividido en conocimiento textual y extratextual. Mientras que el primero es el que se extrae del texto al que se aproxima el traductor, el conocimiento extratextual es el que engloba el conocimiento general del mundo, y el conocimiento de fondo o especializado.

El traductor ha de valorar de qué manera los contextos lingüísticos y extralingüísticos pueden ayudar a los destinatarios de la recepción a inferir el significado de manera parecida al que ofrecía el texto original. Por ello, se considera imprescindible un estudio en profundidad del contexto y de los personajes de la obra original con el fin de conocer la carga informativa y cultural encubierta tras el aspecto lingüístico y en consecuencia, poder hacer una comparativa textual de ambas traducciones con la finalidad de corroborar la existencia de lagunas o vacíos culturales que surgen durante el trasvase.

Entre las características mas destacables cabe mencionar la adaptación de los métodos tradicionales de alineación al método más centrado en el teatro, así en lugar de alinear por segmentos de traducción se procede a alinear por réplicas, como unidad mínima de significado.

5. Bibliografía

Bassnett, S. (1991). Translating for the theatre: The case against performability. *Traduction, Terminologie, Redaction*, *4, 1,* 99-111.

Baker, Mona (1995). Corpora in Translation Studies. An Overview and Suggestions for Future Research, *Target 7(2),* 223-43.

Bolt, R. (1967). *Un hombre para la Eternidad. A Man for All Seasons*. Traducción de Luis Escobar. Madrid: Ediciones Iberoamericanas.

Bolt, R. (1969). *A Man for All Seasons*. Nueva York: Vintage International.

Espasa, E. (2001). *La Traducció dalt de l'escenari*. Barcelona: Eumo.

Franco Aixellá, J. (2000). *La traducción condicionada de los nombres propios inglés-español*. Salamanca: Almar.

Gostand, R. (1980). Verbal and non verbal communication: Drama as translation. En O. Zuber (Ed.).*The languages of theatre*. (pp. 1-9). Oxford: Pergamon Press.

Laviosa, S. (2002). Co*rpus-based Translation Studies: Theories, Findings, Applications*. Amsterdam/New York: Rodopi.

McAuley, G. (1995). Translation in the performance process. *Revista About Performance. Translation and Performance*. Centre for Performance Studies, 111-125. Sydney: Universidad de Sydney.

Merino Álvarez, R. (1994). *Traducción, tradición y manipulación. El teatro inglés en España 1950-1990*. León: Universidad. Secretariado de publicaciones.

Olohan, M. (2004). *Introducing Corpora in Translation Studies*. London and New York: Routledge.

Sánchez García, J. M. (1996). La comparación intertextual en una aproximación al texto traducido dentro de la Traductología Descriptiva. *Epos. Revista de Filología*, 12, 357-378.

Santoyo, J. C. (1995). Reflexiones, teoría y crítica de la traducción dramática. Panorama desde el páramo español. En F. Lafarga, y R. Dengler (Eds.). *Teatro y Traducción* (pp. 13- 23). Barcelona: Universidad Pompeu Fabra.

Toury, G. (2004). *Los estudios descriptivos de Traducción y más allá. Metodología de la investigación en Estudios de Traducción*. Madrid: Cátedra.

Zuber-Skerritt, O. (Ed.). (1984). *Page to Stage. Theatre as Translation*. Amsterdam: Rodopi.

COMENEGO (Corpus Multilingüe de Economía y Negocios): corpus estable vs. metodologías *ad hoc (web as/for corpus)* aplicadas a la práctica de la traducción económica, comercial y financiera

Daniel Gallego Hernández *(Universidad de Alicante)*
Ramesh Krishnamurthy *(Aston University)*

La práctica de la traducción económica requiere desarrollar la competencia instrumental, que implica el uso de textos paralelos. Las actuales posibilidades tecnológicas permiten explotar dichos textos empleando metodologías de corpus. En este sentido, los corpus, concebidos como un conjunto de textos paralelos, pueden ayudar a satisfacer las necesidades informativas del traductor. Estos recursos pueden estar disponibles en Internet. En cambio, si el traductor se enfrenta a un texto cuyo campo no se encuentra entre los textos de dichos corpus ya compilados, es él quien puede buscar y compilar sus propios textos. En el caso de la traducción económica en francés y español, pocos son los corpus virtuales que pueden servir de recurso documental. COMENEGO puede ayudar a que el traductor reduzca el tiempo invertido con las metodologías ad hoc y explote directamente los textos. En este trabajo trataremos los temas relacionados con el diseño y la creación de este corpus, así como algunas de sus diferencias respecto de dichas metodologías.

traducción económica, COMENEGO, textos paralelos, corpus multilingüe

The practice of business translation requires the development of instrumental competence which involves using parallel texts. Current technological possibilities allow translators to exploit these texts by using corpus methodologies. In this sense, corpora conceived as sets of parallel texts may help the translator to satisfy his information needs. These corpora can be distributed via the Internet. However, if the translator is facing a text whose area of specialisation is not represented in the texts of the corpora currently available on the internet, he may compile his own ad hoc corpus. In the case of French-Spanish and Spanish-French business translation, there are very few corpora that can assist translation practitioners. COMENEGO may help the translator to save time when using ad hoc methodologies, by offering a ready-made variety of texts. The aim of this work is to discuss the issues relating to the design and creation of this corpus, as well as its differences from ad hoc methodologies.

business translation, COMENEGO, parallel texts, corpus, multilingual

1. Introducción

El traductor de textos económicos que trabaja en francés y español puede acceder, a través de internet, a una importante cantidad de información económico-financiera para satisfacer las necesidades informativas que, en forma de problemas y dificultades de traducción, pueden presentársele durante su trabajo.

A día de hoy, las posibilidades tecnológicas permiten diversas estrategias para aprovechar al máximo estas fuentes de información. El traductor puede navegar por la web y consultar los textos paralelos (textos comparables respecto de la función, tema o situación comunicativa de los textos originales objeto de traducción) que se encuentra, utilizar la web como si fuera un corpus, compilar en su ordenador dichos textos y consultarlos con aplicaciones de análisis de corpus y, por supuesto, consultar los corpus ya compilados disponibles en línea.

Sobre utilizar la web como si fuera un corpus *(web as corpus),* son pocos los trabajos aplicados a la práctica de la traducción que se han llevado a cabo (Gallego Hernández, 2010*a*). Esta metodología supone emplear Google u otros buscadores como si fueran extractores de concordancias. Requiere el establecimiento de una serie de parámetros referidos tanto al texto original como a los textos paralelos, así como el uso estratégico de las funcionalidades y operadores del buscador utilizado. Los textos no se descargan en el ordenador, sino que se consultan a partir de los descriptores y resultados que devuelven los buscadores al interrogarlos con ecuaciones de búsqueda.

En cuanto a utilizar la web para compilar corpus *(web for corpus),* son diversos los trabajos que, en su aplicación a la práctica de la traducción, han debatido el tema (Bernardini & Zanettin, 2000; Corpas Pastor, 2002; Zanettin *et al.,* 2003; Sánchez Gijón, 2004; Beeby *et al.,* 2009; Rodríguez Inés, 2009; Gallego Hernández, 2010*b*). El diseño y la compilación de este recurso documental suponen, en esencia, reunir en un tiempo reducido una serie de textos paralelos cuya información término-fraseológica y conceptual permita al traductor resolver aquellos problemas o dificultades de traducción que puedan surgirle durante un encargo de traducción específico. En líneas generales, su compilación implica localizar los textos, descargarlos al ordenador, convertirlos a un formato reconocible por los programas de concordancias y, si cabe, limpiarlos.

Respecto de utilizar corpus ya compilados y disponibles en línea, los corpus de corte económico que pueden ser de aplicación a la traducción francés-español y español-francés escasean. En este sentido, el corpus técnico del IULA, aunque es de libre acceso, solo contiene un subcorpus español de economía de alrededor de un millón de palabras (Cabré & Martorell, 2004: 174). Por su parte, CLUVI permite consultar un subcorpus español de economía, EGAL (0,4 millones de palabras), y otro de consumo, CONSUMER (1,8 millones de palabras en español). Sin embargo, no tiene textos en francés. El MLCC Multilingual and Parallel Corpora contiene un subcorpus genérico de artículos financieros de periódicos en francés y español, pero es de pago y los textos son de 1990-1994. Vicente (2007) posee un corpus representativo del lenguaje comercial en francés y español de la prensa generalista y especializada, pero es privado.

2. *COMENEGO: Corpus Multilingüe de Economía y Negocios*

Ante este panorama, como formadores de traductores para el ámbito de la economía y los negocios, nos vemos obligados actualmente a implementar metodologías de consulta *ad hoc* de textos paralelos que implican destinar parte del tiempo disponible en el aula a su explicación y desarrollo. Por ello, nos parece interesante compilar COMENEGO, un corpus con el que, entre otras cosas, pretendemos reducir el espacio de tiempo dedicado al desarrollo de subcompetencias tecnológicas y aumentar el tiempo dedicado a la traducción de la especialidad.

El acrónimo se corresponde con *Corpus Mulitilingüe de Economía y Negocios,* donde *corpus* es, en términos de Sinclair (1996), «a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language», *multilingüe* equivale a francés y español, al tiempo que deja espacio a otras lenguas, y *economía y negocios* puede entenderse en el sentido de Mateo Martínez (2007), referido al lenguaje de la economía teórica, es decir, un lenguaje académico propio de expertos investigadores que utilizan un discurso poco transparente de gran especialización terminológica, y al de los negocios, un lenguaje profesional propio de usuarios no solo teóricos, sino también especialistas del mundo del comercio y las finanzas, periodistas e incluso personas sin ninguna especialización.

2.1. Finalidad de COMENEGO

Como acabamos de comentar, COMENEGO surge de la necesidad que, como formadores de traductores, tenemos de reducir el tiempo dedicado a la compilación de corpus *ad hoc* tanto en el aula como en casa. Ahora bien, el corpus no solo está concebido como recurso de documentación o fuente lingüística especializada. También puede servir como herramienta docente (detección de terminología y fraseología de interés, elaboración de ejercicios de terminología, traducción o revisión, etc.), así como objeto de análisis con fines de investigación.

2.2. Fases en la compilación de COMENEGO

En líneas generales, la compilación de un corpus implica diferentes fases que, en conjunto, suponen un proceso cíclico: 1) documentación (identificación, por ejemplo, en una hoja de cálculo, de las características esenciales de los textos: idioma, variedad, campo, modo, etc.); 2) selección de textos en función de la finalidad y disponibilidad del corpus, de criterios externos relativos a características textuales como el idioma y sus variedades, el modo (discursos orales, escritos, textos electrónicos, etc.), el tipo de texto (libros, artículos, etc., en el caso de textos escritos; seminarios, presentaciones, etc., en el caso de discursos orales), la especialidad, el año y otras características; o de criterios internos fundamentados en las características propias de los textos; 3) obtención de permisos de usos (si se va a distribuir el corpus); 4) adquisición de textos (descarga de archivos, inclusión de metadatos, etc.) y 5) conversión a texto plano y limpieza de archivos (eliminación de ruido, caracteres no reconocidos, etc.).

En nuestro caso, los recursos textuales vienen acompañados de una hoja de cálculo que contiene diversos campos (URL, fecha de descarga, tipo de texto, etc.). La selección de textos, como hemos comentado, responde fundamentalmente a fines de formación y práctica de la traducción. La tabla 1 del apéndice contiene ejemplos de los tipos de textos recuperados, así como las categorías en las que nos hemos basado para agruparlos en diferentes subcorpus.

Algunas de estas categorías se corresponden con los tipos funcionales de comunicación en las organizaciones que, según unos criterios pragmático-discursivos (función, interlocutores, estructura y estilo), identifica Cassany (2004: 53-55). Se trata, en concreto, de las categorías *técnico (TEC)* y *científico (SCI),* que tienen una función principalmente referencial, intentan transmitir información de manera objetiva y se dan en áreas técnicas de la organización (proyectos, auditoría, investigación). Nuestra clasificación también tiene en cuenta su concepción del discurso *organizativo (ORG),* que puede tener una función conativa, referencial o metalingüística, trata de ordenar y regular la actividad de la organización, y suele tener cabida en áreas concretas (dirección, personal, administración, evaluación, calidad). Aunque el autor incluye dentro de esta misma categoría lo que denomina *lenguajes administrativo y jurídicos,* hemos creado una categoría aparte, *legal (LEG),* muy relacionada con la organizativa, pero referida, en esencia, a textos que tienen como objetivo no regular la actividad de una organización concreta, sino del conjunto de organizaciones de uno o varios países. De su clasificación también hemos recuperado el discurso *comercial (COM),* cuya función ronda entre la conativa y la referencial, pretende influir sobre la opinión y la conducta del destinatario, y suele ser propio de áreas específicas (*marketing,* publicidad, comunicación, ventas).

A estas categorías hemos añadido otras dos, *didáctica (DID)* y *prensa (PRS),* concebidas principalmente según criterios pragmáticos. La categoría *didáctica* se refiere a textos que surgen del círculo formativo y que, por tanto, tienen unos fines de formación no solo en el ámbito académico, como pueden ser cursos o apuntes, sino también en el ámbito profesional, como pueden ser textos que explican al inversor qué es la bola o guías para consumidores. La categoría *prensa,* por su parte, se corresponde con textos de carácter informativo de corte económico, comercial o financiero, publicados por la prensa general y especializada, así como con los comunicados o notas de prensa de las organizaciones.

La tabla 2 del apéndice muestra la distribución de estas categorías en ambos idiomas según el número de archivos *(files),* palabras *(tokens),* media de palabras por archivo *(ave)*, tipos de palabra *(types)* y ratio *token-type (ratio)*. Estas cifras se refieren a los archivos TXT convertidos y limpiados (ruido y caracteres no reconocidos). Respecto de la obtención de permisos de usos, hemos llevado a cabo algunas peticiones que han recibido una respuesta afirmativa. En cualquier caso, se trata de un proceso todavía por completar, que probablemente hará variar estas cifras.

2.3. Análisis de COMENEGO

La lingüística de corpus dispone de diversas herramientas de análisis, como WordSmith Tools, AntConc (utilizado en el presente análisis), etc., que permiten extraer diversos

datos en forma de 1) listados de palabras frecuentes: «the rank of a word-form in a corpus frequency list has some relationship to the importance of that word-form in the linguistic system» (Krishnamurthy, 2001), 2) concordancias (listados de palabras clave contextualizadas, que pueden ordenarse o expandirse según las necesidades del análisis), 3) colocaciones (otras palabras que aparecen inesperadamente con más frecuencia alrededor de palabras clave) y 4) n-gramas (listados de secuencias de palabras más frecuentes).

A continuación presentamos un breve análisis preliminar que puede servir para ilustrar el tipo de análisis que es posible llevar a cabo en investigaciones futuras, al tiempo que resaltamos en algunos puntos la necesidad de dar nuevos pasos en el estudio del corpus.

En la tabla 2 del apéndice podemos ver las similitudes de cada corpus respecto de su tamaño (9115352 *tokens* en español frente a 9086627 en francés), pero hemos de tener en cuenta que el cómputo de palabras puede verse afectado por las diferencias morfológicas de cada idioma.

El tamaño de los subcorpus o categorías es muy parecido: una media de 1,3 millones de palabras con variaciones de entre 1,19 y 1,37 millones. Estos datos pueden verse afectados por las posibles diferencias en la tipología textual de cada lengua.

Ahora bien, existen algunas diferencias significativas entre el corpus español y el francés: 1) el español tiene 10050 archivos, pero el francés solo 8880, lo que significa que los textos franceses son, de media, más largos que los españoles (1023 palabras frente a 907 palabras); 2) aun así, hay variaciones en las diferentes categorías: mientras que los subcorpus COM, DID, LEG y TEC tienen más textos en español que en francés (5255/3909, 1491/1121, 211/21 y 351/133, respectivamente), y, por tanto, los textos españoles son más cortos, los subcorpus ORG, PRS y SCI tienen, en cambio, más textos en francés que en español (634/429, 2859/2214 y 203/99, respectivamente), y, por tanto, sus textos en francés son, de media, más cortos que los españoles; por último 3) los textos más cortos en ambas lenguas pertenecen al subcorpus COM; por su parte, los más largos en español se encuentran en el subcorpus SCI, y los más largos en francés, en el subcorpus LEG.

2.3.1. Palabras frecuentes

En cualquier idioma las palabras más frecuentes suelen ser palabras gramaticales, como determinantes, preposiciones, etc. Sin embargo, precisamente por su elevada frecuencia, su polifuncionalidad y su complejidad de uso, conviene analizarlas mejor con más detenimiento. Por ello nos centramos ahora en las palabras no gramaticales más frecuentes, también denominadas *palabras de contenido (content words* o *vocabulary words)*. La tabla 3 del apéndice contiene las 15 palabras más frecuentes en cada idioma (nótese que hemos optado por no distinguir entre mayúsculas y minúsculas). Este tipo de listados puede ayudar a identificar, por una parte, las palabras que los traductores deberían conocer dentro de un amplio abanico de contextos, significados y usos, y, por otra parte, un elevado número de palabras que, por su frecuencia mínima dentro del corpus, pueden, en principio, no resultarles de tanto interés.

Observemos, en primer lugar, los ítems similares en cada lengua. Podemos ver que 7 de las 15 palabras más frecuentes de cada corpus son cognados y, por tanto, los primeros equivalentes que proponen la mayoría de los diccionarios bilingües (ahora bien, hemos de tener en cuenta que pueden esconder diferentes significados y usos en ambas lenguas). También apreciamos algunas variaciones menores: en español, la forma singular *cuenta* coincide con las formas *compte* en singular y *comptes* en plural. Por supuesto, debemos consultar el listado completo de formas en español, pues *cuentas* aparece con una frecuencia algo menor. Ocurre algo parecido con el plural *empresas,* representado en francés con su forma en singular *(entreprise).* En análisis posteriores, será importante comparar un mayor número de ítems y analizar de manera más precisa las diferencias en la posición que ocupan.

Observemos ahora los ítems que aparecen en un solo idioma: *información, mercado, valores, millones, general, valor, riesgo,* en el caso del español, y *assurance, actions, conseil, france, conditions, ans,* en el caso del francés. De nuevo, deberemos analizar los listados de palabras completos para comprobar si se trata de diferencias significativas o de simples variaciones menores de frecuencia en cada corpus.

2.3.2. Concordancias

Como dijimos anteriormente, las concordancias permiten observar cualquier ítem aparecido en los listados de palabras frecuentes, tal como aparece en los textos, con los contextos que le rodean. Se trata del nivel de análisis más detallado, y permite investigar cada palabra en cada texto. Con ellas es posible identificar clases de palabras, significados, usos, colocaciones y fraseologismos, patrones gramaticales, usos pragmáticos y características específicas de los géneros textuales.

Incluso en el pequeño listado de 10 de las 11879 concordancias de *cuenta* (tabla 4 del apéndice), podemos apreciar algunos patrones: 4 líneas contienen el verbo *cuenta* seguido de *con,* 2 tienen la secuencia *de ser cliente de cuenta Nómina.* En futuros análisis, el estudio de la totalidad de sus contextos ayudará a determinar qué patrones referidos a *cuenta* pueden ser de utilidad para los traductores, extraer la terminología especializada de interés, etc.

También apreciamos algunos patrones en el caso de las 12 concordancias en francés (de 10342) aparecidas en la tabla 5 del apéndice: 4 vienen seguidas de la preposición *de,* 3 de las cuales vienen precedidas de *tenir/tenant,* y una, de *rendant* (también encontramos un ejemplo de *compre rendu*); 3 ejemplos de *votre compre,* dos de los cuales vienen seguidos de *sera* y uno, precedido de *sur;* un ejemplo con *compte-titres;* además *titres* aparece en otra posición dentro de otra línea; un ejemplo con *ouvrir un compte* e incluso uno con el verbo *qui compte.* Nuevamente, una investigación más detallada revelará qué patrones son significativos y cuáles no. Por su parte, la elevada frecuencia de la forma en plural *comptes* (9149) también puede ayudar a identificar distintos usos de esta forma.

2.3.3. Colocaciones

Como acabamos de apreciar, las concordancias pueden ayudarnos a obtener diferentes informaciones sobre las características y comportamientos de una palabra. Ahora bien,

también existen otras herramientas que pueden acelerar los procedimientos iniciales de detección de información. Se trata de las colocaciones, una herramienta de mayor complejidad que requiere del analista conocimientos lingüísticos avanzados. Esta función ofrece un análisis cuantitativo de las palabras que aparecen en las proximidades del núcleo o palabra clave. Algunos trabajos, como el editado por Krishnamurthy (2004), muestran que los colocativos más significativos suelen aparecer dentro de una distancia de 4 palabras alrededor del núcleo o palabra clave. AntConc nos permite seleccionar esta distancia.

Si observamos el listado de la tabla 6 del apéndice, observamos que parece haber menos correspondencias formales que en el listado anterior de palabras frecuentes (tabla 3): solo *corriente-courant* y *depósito-dépôt.* Tras un examen más detenido, apreciamos dos equivalentes semánticos potenciales: *ahorro-épargne* y *vivienda-logement,* así como la presencia del nombre de entidades bancarias: *Caixa-Bred.* El hecho de que no sea posible asociar fácilmente los colocativos de una y otra lengua sugiere que las palabras *cuenta* y *compte* se emplean en diferentes contextos y fraseologismos en ambos idiomas. Ello demuestra la necesidad de ser cautos en futuros análisis: las palabras con similitudes formales/etimológicas pueden ser falsos amigos.

Existen otros aspectos de la herramienta de colocaciones que pueden utilizarse para futuros análisis, como la posibilidad de determinar exactamente en qué posición aparece cada colocativo respecto de la palabra clave o núcleo.

2.3.4. 4-gramas

Otra herramienta que puede ser útil para identificar las características lingüísticas de los textos son los n-gramas. En lugar de crear listados de palabras frecuentes, esta herramienta crea listados de secuencias de dos palabras (2-gramas), tres palabras (3-gramas), etc.

Tal como ocurre con los listados de colocativos (tabla 6), los listados de 4-gramas aparecidos en la tabla 7 del apéndice muestran más diferencias que similitudes. Ahora solo queda un ítem paralelo del listado de palabras frecuentes (tabla 3): *artículo-article.* Algunas palabras continúan apareciendo en una de las listas, pero no en la otra: por ejemplo, *euros* aparece en el listado español, pero no en el francés; *comptes* sigue estando en el listado francés, el verbo c*ompter* aparece en su forma infinitiva, pero *compte* desaparece, tal como hace *cuenta* en el listado español. Asimismo, en este nuevo listado aparecen ahora nuevos ítems, como *cadre,* que no aparecían en los listados de palabras frecuentes. Nótese que hemos tenido en cuenta la diferencia entre mayúsculas y minúsculas para mostrar que ahora aparecen algunos nombres propios de instituciones y órganos de gobierno. También es posible identificar algunos cognados: *dispuesto-dispositions.*

3. Conclusiones

En este trabajo hemos presentado COMENEGO en contraposición con metodologías *ad hoc web as/for corpus* y hemos llevado a cabo un análisis preliminar con herramientas de explotación de corpus.

Respecto de sus diferencias con las metodologías *ad hoc web for corpus,* creemos que, con COMENEGO, tenemos posibilidades de omitir, entre otras cosas, la fase de compilación tanto para traductores como para formadores. En consecuencia, es posible disponer de más tiempo para dedicar a la práctica de la traducción propiamente dicha y a la planificación docente. Asimismo, el desarrollo de la competencia instrumental se ve reducido en su dimensión tecnológica, pues, una vez compilado y disponible el corpus, los usuarios no requerirán de conocimientos informáticos avanzados. Otro punto que puede jugar a favor es la fiabilidad de los textos, que puede ser mayor respecto de las metodologías *ad hoc,* con las que es posible recuperar textos no fiables o ruidosos. Desde el punto de vista del acceso a los recursos documentales, la centralización de los textos en un servidor puede suponer una ventaja respecto de las metodologías *ad hoc,* que requieren tener en el ordenador del usuario no solo los archivos del corpus, sino también el *software* de lectura. Además, la explotación de corpus puede verse mejorada, al menos respecto de las metodologías *web as corpus* que emplean buscadores comerciales y cuyas estrategias de búsqueda (truncación, palabras en contexto, visualización de resultados, etc.) suelen verse limitadas.

Por su parte, el análisis presentado tan solo es la punta del iceberg y muestra el tipo de estudio que puede llevarse a cabo sobre COMENEGO. No hemos querido ir más allá, pues todavía estamos ajustando y equilibrando el corpus y sus componentes. Ahora que disponemos de un corpus piloto, podemos embarcarnos en un análisis global más riguroso aprovechando al máximo las herramientas de explotación de corpus. Las observaciones y datos mostrados simplemente dejan entrever la variedad de resultados que pueden obtenerse con el análisis, así como sus potenciales beneficios para los traductores.

En cualquier caso, COMENEGO es un corpus piloto todavía en construcción. El análisis de sus distintos subcorpus puede ayudar a definir con más precisión las categorías textuales establecidas. Es posible complementar la selección de textos incluyendo nuevas tipologías a partir, por ejemplo, de encuestas dirigidas tanto a los iniciadores de traducción como a los propios traductores profesionales con el propósito no solo de conocer sus necesidades de traducción, sino también de conocer qué tipos textuales, temas, campos, etc., son los que suelen traducirse en el ámbito profesional. Por otra parte, las funciones del corpus son, de momento, limitadas, pues la fase de solicitud de permisos de uso de los recursos textuales no está completada, lo que hace que se trate actualmente de un corpus privado con fines de investigación y de uso propio, y tampoco se dispone de una plataforma o sistema virtual que permita la consulta de los textos.

4. Apéndice

Tabla 1: categorías textuales en COMENEGO

TIPOS DE TEXTOS	CATEGORÍAS
promoción de productos bancarios, productos financieros y de seguros, páginas web corporativas, etc. (sitios web comerciales)	comercial
cursos en línea, guías para consumidores, inversores y clientes, etc. (sitios web comerciales e informativos: personales de profesores, universidades, instituciones)	didáctico
leyes, códigos, decretos, etc. (sitios web informativos: ministerios y agencias)	legal
estatutos, reglamentos, actas de juntas, etc. (sitios web corporativos e informativos)	organizativo
notas de prensa, noticias, boletines de noticias, etc. (sitios web corporativos y de periódicos)	prensa
artículos y trabajos, etc. (sitios web informativos: revistas especializadas)	científico
cuentas e informes anuales, resultados, informes de gestión, análisis técnicos, planes de marketing, informes sectoriales, etc. (sitios web corporativos e informativos)	técnico

Tabla 2: distribución de recursos en COMENEGO

CAT	ESPAÑOL					FRANCÉS				
	FILES	TOKENS	AVE	TYPES	RATIO	FILES	TOKENS	AVE	TYPES	RATIO
COM	5255	1329915	253	35321	37,65	3909	1325544	339	29316	45,21
DID	1491	1276089	856	40641	31,39	1121	1304585	1164	35937	36,30
LEG	211	1342698	6363	23077	58,18	21	1293704	61605	14772	87,57
ORG	429	1337822	3118	29417	45,47	634	1365468	2154	21885	62,39
PRS	2214	1329029	600	37314	35,61	2859	1308418	458	37928	34,49
SCI	99	1311731	13250	34483	38,03	203	1301102	6409	32710	39,77
TEC	351	1188068	3385	40777	29,13	133	1187806	8931	24646	48,19
TOTAL	10050	9115352	907	113100	80,59	8880	9086627	1023	89133	101,94

Tabla 3: listado de palabras frecuentes no gramaticales en COMENEGO

ESPAÑOL			FRANCÉS		
RANK	FRECUENCIA	PALABRA	RANK	FRECUENCIA	PALABRA
29	17540	información	28	26843	article
30	17531	mercado	41	15921	cas
31	17339	artículo	42	15294	société
35	15557	euros	50	11854	capital
36	15015	caso	52	11535	assurance
38	13921	sociedad	55	10898	actions
39	13582	valores	61	10379	entreprise
40	13378	millones	62	10342	compte
41	12421	capital	63	10335	conseil
42	12407	general	69	9556	france
43	11879	cuenta	70	9532	groupe
45	11777	valor	73	9149	comptes
47	11358	empresas	74	9027	conditions
49	11141	grupo	76	8821	ans
50	10969	riesgo	78	8607	euros

Tabla 4: listado de concordancias en español (cuenta)

…so de ser cliente de cuenta Nómina) o con Certifi…
…eneral de la Policía cuenta con 350 Oficinas de E…
…so de ser cliente de cuenta Nómina) o con el Cert…
…comendamos tengan en cuenta. Ante cualquier duda,…
…total o parcial, por cuenta de gobiernos o autori…
… capital liberada. A cuenta Complementario Total …
…07/07/880,1800,144 A cuenta Complementario Total …
…d, abertis logística cuenta con cerca de 933.000 …
…dad El Grupo abertis cuenta con una plantilla med…
… de los trabajadores cuenta con contrato indefini…

Tabla 5: listado de concordancias en francés (compte)

…st ajusté pour tenir compte de la modification du…
…inatif pur Ouvrir un compte au nominatif pur Pour…
…'Administration, qui compte désormais 12 membres …
… du dividende. Votre compte sera crédité dans les…
…irectement sur votre compte. Au nominatif adminis…
…e 17 mai 2010. Votre compte sera crédité dans les…
… suivants pour tenir compte des délais de traitem…
…ts et en lui rendant compte de son examen : organ…
… d'une action tenant compte des opérations ayant …
…ier pour la tenue du compte-titres. Ils représent…
…s titres inscrits en compte nominatif pur. Droit…
…umenté au moyen d'un compte rendu écrit Avec le …

Tabla 6: listado de colocativos (cuenta y compte)

COLOCATIVOS DE CUENTA	COLOCATIVOS DE COMPTE
nómina	unités
corriente	bancaire
servicio	épargne
naranja	titres
vivienda	bred
caixa	dépôt
condiciones	courant
ahorro	logement
producto	numéro
crédito	livret
vista	ouverture
tarjeta	gestion
depósito	titulaire
valores	relevés

Tabla 7: listado de 4-gramas

ESPAÑOL	FRANCÉS
de millones de euros	dans le cadre de
del Mercado de Valores	du Code de commerce
de la Ley de	dans les conditions prévues
la Ley de de	à compter de la
del Consejo de Administración	alinéa de l'article L
Nacional del Mercado de	Crédit Agricole S A
Comisión Nacional del Mercado	de commerce et d'industrie
la Comisión Nacional del	droit préférentiel de souscription
el Consejo de Administración	en application de l'article
en el caso de	valeurs mobilières donnant accès
artículo de la Ley	la mise en place
Consejo de Administración de	par lettre recommandée avec
el artículo de la	dans la limite de
los millones de euros	l'Autorité des marchés financiers
lo dispuesto en el	le commissaire aux comptes
en el artículo de	des commissaires aux comptes
En el caso de	le cadre de la
que se refiere el	dispositions de l'article L

5. Bibliografía

Beeby, A.; Rodríguez Inés, P. & Sánchez Gijón, P. (Eds.). (2009). *Corpus Use and Translating.* Amsterdam/Philadelphia: John Benjamins.

Bernardini, S. & Zanettin F. (Eds.). (2000). *I corpora nella didattica della traduzione: Corpus use and learning to translate.* Bologna: CLUEB.

Cabré Castellví, M. & Bach Martorell, C. (2004). El corpus tècnic del IULA: corpus textual especializado plurilingüe. *Panacea, 16,* 173-176.

Cassany, D. (2004). Explorando los discursos de las organizaciones. In van Hooft Comajuncosas, A. (Ed.). *Textos y discursos de especialidad. El español de los negocios* (pp. 49-70). Amsterdam/New York: Rodopi.

Corpas Pastor, G. (2002). Traducir con corpus: de la teoría a la práctica. In García Palacios, J. & M.ª T. Fuentes Morán (Eds.). *Texto, terminología y traducción* (pp. 189-226). Salamanca: Almar.

Gallego Hernández, D. (2010a). *Traducción económica y textos parlelos en internet. Aproximación teórica y metodológica.* Tesis doctoral, Universidad de Alicante.

GALLEGO HERNÁNDEZ, D. (2010b). Acquiring instrumental sub-competence by building do-it-yourself corpora for business translation. Paper presented at the *Using Corpora in Contrastive and Translation Studies*.

KRISHNAMURTHY, R. (2001). Size Matters: creating Dictionaries from the World's Largest Corpus, pp 169-180 in 8th Annual KOTESOL Conference Proceedings, Taegu: KOTESOL.

KRISHNAMURTHY, R. (ED.) (2004). English collocation studies: The OSTI Report by John Sinclair, Susan Jones and Robert Daley. London: Continuum Books.

MATEO MARTÍNEZ, J. (2007). El lenguaje de las ciencias económicas. In Alcaraz Varó, E., *et al.* (Eds.). *Las lenguas profesionales y académicas* (pp. 191-203). Barcelona: Ariel.

SÁNCHEZ GIJÓN, P. (2004). *L'ús de corpus en la traducció especialitzada: compilació de corpus ad hoc i extracció de recursos terminològics.* Barcelona: Universitat Pompeu Fabra.

SINCLAIR, J. (1996). Preliminary recommendations on Corpus Typology. Retrieved from http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html.

VICENTE, C. (2007). Lingüística de corpus y traducción especializada: aplicaciones a la traducción francés-español de la economía. Paper presented at the *XXV Congrès international de linguistique et de philologie romanes.*

Elaboración de glosarios a partir de corpus paralelos *ad hoc*. Aplicación a la interpretación de conferencias en el ámbito socioeconómico

Daniel Gallego Hernández

Universidad de Alicante

Miguel Tolosa Igualada

Universidad de Alicante

Resumen:

La interpretación de conferencias no da pie a que los profesionales que se dedican a ella puedan documentarse, al menos no de manera exhaustiva, durante el proceso de escucha activa-reformulación. El trabajo documental deberá, pues, llevarse a cabo antes de la celebración del evento. El intérprete tiene la posibilidad de compilar en su ordenador textos relacionados con el evento con el objetivo básico de extraer, en forma de glosarios, el vocabulario de sus lenguas de trabajo, y anticiparse así, en la medida de lo posible, a los eventuales problemas y dificultades que puedan presentársele durante la interpretación. En este trabajo nos proponemos reflexionar sobre los pasos que, en la etapa de documentación previa a la conferencia, el intérprete puede dar para elaborar este tipo de glosarios a partir de corpus paralelos compilados ad hoc.

Palabras clave: interpretación, documentación, terminología, corpus paralelos

Abstract:

Conference interpreting does not allow professionals to carry out a documentation work, at least not exhaustively, in the process of active listening and reformulating. This work must therefore be carried out before the event. Interpreters have the possibility to compile in their computers texts related to the event with the basic purpose of creating a glossary in their working languages and anticipating, as far as possible, any problems and difficulties that they may come across during the interpretation. The aim of this paper is to show some steps involved in the documentation work before the event that interpreters may follow to extract terminology from ad hoc parallel corpus.

Key words: interpreting, documentation work, terminology, parallel corpus.

Estado de la cuestión: interpretación y documentación

Una lectura pormenorizada de la bibliografía en torno a la investigación en interpretación permite vislumbrar la ingente cantidad de artículos, libros, conferencias, etc. dedicadas, por ejemplo, a las destrezas o habilidades cognitivas que debe reunir el intérprete para serlo. Pero ¿por qué no se ha dedicado tanta atención a la preparación documental y terminológica que debe llevar a cabo todo intérprete antes, durante y después de su prestación para que ésta concluya de manera satisfactoria? En este mismo sentido, al plantearnos en qué medida se han vinculado corpus e interpretación en el marco de la investigación, de la docencia y de la profesión de dicha actividad, nos damos cuenta de que los traductólogos han relacionado ambas nociones en el campo de la investigación mayoritariamente. Dicho de otro modo, se ha utilizado la metodología de corpus para investigar la interpretación como actividad (Shlesinger, 1998; Corpas, 2008: 95-98), pero no hemos encontrado tantos trabajos que expliquen las ventajas e inconvenientes que plantea la utilización de corpus para la confección de glosarios desde una perspectiva profesional y/o docente, si exceptuamos algunos trabajos (Szabó, 2003). Existe, sin embargo, otro grupo de trabajos (Blasco & Jiménez, 2003; Bianchessi *et al.,* 2011), ciertamente reducido, que sin entrar a valorar en profundidad la pertinencia (o falta de ella) de confeccionar glosarios a partir de corpus, sí ponen de relieve la importancia de la preparación temática y terminológica a la hora de realizar una interpretación profesional, sobre todo si ésta aborda cuestiones técnicas o científicas, y también las implicaciones que tal preparación podría tener para la docencia (Choi, 2005; Foulquié & Navarro, 2011).

Así pues, es bien sabido y reconocido en el mundo de la interpretación que una buena preparación documental y terminológica previa de las conferencias (Gile, 1985, 1995; Choi, 1998, 2005; Seleskovitch & Lederer, 2002; Nolan, 2005) es condición necesaria para aspirar a un trabajo operativo y funcional tanto para los asistentes o participantes en el evento como para los intérpretes que trabajan en *relé,*[87] aunque desgraciadamente, y con demasiada frecuencia, no resulta suficiente (Gulyás, 2003) por razones ajenas al propio intérprete.

Caracterización del trabajo del intérprete

Sin ánimo de ser exhaustivos, podemos diferenciar dos ámbitos de actuación en los que los intérpretes desarrollan su actividad:

- la interpretación de conferencia y
- la interpretación en los servicios públicos.

En la interpretación de conferencias se suele incluir la interpretación consecutiva, la interpretación simultánea y la interpretación susurrada o *chuchotage*. ESPaiic (http://espaiic.es), asociación española de profesionales de la interpretación de conferencias miembros de la Asociación Internacional de Intérpretes de Conferencia (AIIC), explica

87 Práctica consistente en trabajar a partir de la interpretación de otra cabina en vez de partir directamente del discurso original.

en su página web en qué consiste la interpretación consecutiva en los siguientes términos: "el intérprete, situado junto a los oradores, toma notas del contenido de una intervención, que puede alargarse varios minutos, y seguidamente reproduce el discurso con toda exactitud". Por lo que se refiere a la interpretación simultánea, ESPaiic afirma que se trata de la modalidad en la que "el intérprete, sentado en una cabina insonorizada frente a un micrófono, escucha mediante auriculares las intervenciones de los oradores y las traduce en tiempo real a otro idioma para los delegados, que escuchan a través de receptores". En la "interpretación susurrada", los intérpretes interpretan sin cabina turnándose para susurrar el mensaje original "al oído" de uno o dos delegados como máximo.

Respecto de la interpretación en los servicios públicos, Wadensjö (1998: 33) considera que tiene como objetivo facilitar la comunicación entre el personal oficial y los usuarios inmigrantes. Para Abril Martí (2006: 5) es "aquella que facilita la comunicación entre los servicios públicos nacionales –policiales, judiciales, médicos, administrativos, sociales, educativos y religiosos– y aquellos usuarios que no hablan la lengua oficial del país y que habitualmente pertenecen a minorías lingüísticas y culturales".

Reflexión

En ambos ámbitos de actuación, es obvio que la preparación documental y terminológica del intérprete resulta fundamental para ofrecer un producto final de calidad. Sin embargo, en no pocas ocasiones, el intérprete profesional recibe escasa documentación o información de los organizadores del evento. Y es que por mucho que ESPaiic afirme que "para la calidad de la interpretación, es sumamente importante la colaboración de los organizadores del acto, que pueden facilitar a los intérpretes documentación, material de referencia o glosarios. Esta colaboración es crucial cuando se prevé que un documento sea leído o citado. En este caso, los intérpretes deben recibir una copia con antelación", aquél que haya trabajado como intérprete *free-lance*, en este caso de conferencias, en el mercado privado español habrá vivido en sus propias carnes lo que es y supone enfrentarse a una interpretación "a tientas". Es evidente que forma parte de la profesionalidad del intérprete documentarse antes de interpretar pero, cuando la única información con la que se cuenta es el lugar, hora y programa sucinto (muchas veces modificado a última hora porque algún ponente o conferenciante ha causado baja), esta operación se convierte en una verdadera lotería. Por mucha especialización, cultura, competencia lingüística, intuición, horas de vuelo, capacidad camaleónica y de reacción que atesore el intérprete, éste puede verse sorprendido por la especificidad temática y terminológica de la conferencia en la que está interpretando. De este modo, la "interpretación a tientas" suele venir como consecuencia de tres situaciones que, lamentablemente, pueden llegar a concurrir en un mismo evento:

- No contar con documentación[88] relacionada con el evento que se va a interpretar.
- No contar con tiempo suficiente para preparar el evento en condiciones, por haber contratado al intérprete a ultimísima hora.

88 Otro problema menos frecuente, al menos desde nuestra propia experiencia, es justamente el contrario: contar una ingente cantidad de información, muchas veces irrelevante, enviada por la propia organización del evento que resulta absolutamente imposible asimilar al no contar con tiempo suficiente para hacerlo.

- Desviación o digresión del ponente respecto del tema que se suponía que iba a abordar en su intervención.

Ante la tercera de las situaciones, la solución pasa por estar con los cinco sentidos puestos ya no en lo que el orador "dice", sino en lo que "quiere decir". De todos modos, si el orador decide "darse un paseo" por vericuetos retórico-temático-terminológicos con un alto de grado de especificidad, el intérprete poco podrá hacer, más allá de esperar que el mal trago pase cuanto antes.

A las dos primeras situaciones creemos que sí se les puede encontrar una vía de solución, tal vez provisional, que consiste en explorar las grandes posibilidades que la metodología de corpus ofrece a la hora de confeccionar glosarios bilingües *ad hoc* de una manera rápida y eficaz y que trataremos de aplicar al siguiente caso real.

Caso concreto

Supongamos que nos contratan para trabajar en simultánea en una conferencia de un día de duración que lleva por título: "I Jornada de cooperación internacional al desarrollo en el ámbito de la agricultura y la ganadería". La única información con la que contamos es el programa de dicho evento, disponible en *http://asimov.sav.us.es/internacional/uploads/blog/ FICHA%20INS%20VSF.pdf,* en el que aparecen el nombre de los conferenciantes, su organismo o institución de procedencia y el título de su ponencia. Además, la jornada tendrá lugar a los dos días de la recepción del documento. En tales circunstancias, ¿qué puede hacer el intérprete más allá de rechazar la propuesta?

Compilación del corpus: pasos

Una posibilidad es compilar corpus comparables *ad hoc* en el menor tiempo posible y a bajo coste con el propósito de extraer de ellos terminología en forma de glosarios. Sobre los procesos que pueden seguirse para compilar este tipo de corpus se han escrito bastantes trabajos: a las actas de los congresos CULT (Corpus Use and Learning to Translate) pueden sumársele otros trabajos de autores como Corpas Pastor (2002), Sánchez Gijón (2004) o más recientemente, y en el terreno de la economía Gallego Hernández (2010, 2011). Ahora bien, si lo que pretende el intérprete es elaborar un glosario bilingüe sin tener que traducir él mismo los términos extraídos de un corpus monolingüe, puede recurrir a la compilación de corpus paralelos.

A la hora de compilar este tipo de corpus, podemos tener en cuenta el trabajo de Castillo Rodríguez (2009), que muestra los pasos que sigue para la compilación de un corpus paralelo *ad hoc* con fines de investigación sobre la calidad de la traducción de textos turísticos. En concreto, desarrolla un modelo cuatrifásico: recopilación de los datos, almacenamiento, conversión de formatos y alineación. La primera fase tiene dos subtareas: elaboración de los criterios de diseño (propósito, tamaño, medio, tema, tipo textual, autoría, la fecha de publicación y lenguas) y búsqueda de la información

en organizaciones e instituciones (Organización Mundial del Turismo, Consejería de Turismo, Comercio y Deporte de la Junta de Andalucía, y de una serie de cadenas hoteleras) y con palabras clave (utiliza Google y ecuaciones como spa OR balneario AND turismo AND hotel). Respecto de la fase de almacenamiento, incide en el tipo de codificación empleado para identificar los recursos bitextuales y sugiere la creación bases de datos textuales en hojas de cálculo (cada registro viene referido a cada archivo, y cada campo, al código del texto, título, dominio, tipo y URL). Respecto de la fase de conversión a texto plano, alude concretamente al programa pdf2txt; también se vale de diversas versiones demo de algunos programas comerciales. La autora alude igualmente a la limpieza de textos, que ejemplifica con la corrección de tildes no detectadas o caracteres no reconocidos en el proceso de conversión, la adición de nuevo código, así como la supresión del texto relacionado con los menús propios de las páginas web. Para llevar a cabo esta fase, la autora alude a la función «buscar y remplazar» de Word. Por último, respecto de la alineación de bitextos, la autora menciona WinAalign de Trados, así como una función integrada en ParaConc.

En nuestro caso concreto, centrado en la compilación de corpus paralelos *ad hoc* para fines profesionales (disponemos, por tanto, de menos tiempo a la hora de compilar los textos), podemos seguir los siguientes pasos y emplear otras herramientas:

En primer lugar, identificar los posibles sitios web multilingüe. En este sentido, a sabiendas de que las instituciones internacionales suelen disponer gran parte de sus textos y publicaciones en más de un idioma, podemos conocer diferentes instituciones a partir, por ejemplo, de la ecuación *related:www.fao.org,* con la que Google recupera sitios web similares a la de la Organización de las Naciones Unidas para la Agricultura y la Alimentación, como, entre otros, *www.codexalimentarius.net, www.who.int/es/index.html, es.wfp.org, www.unicef.org, www.ifad.org, www.wto.org, www.oecd.org, www.oie.int/es, www.europa.eu.int, www.cgiar.org/languages/lang-spanish.html.*

Dado que los sitios recuperados pueden no corresponderse con el tema de las conferencias, conviene hacer una prospección de estos sitios no solo con el propósito de conocer cuáles contienen temas y, por tanto, vocabulario de interés, sino también cuáles están estructurados de modo que, con tan solo cambiar aquellos elementos de sus URL relativos al idioma, sea posible acceder a las versiones en francés y en español. Desde este último punto de vista interesan estructuras del siguiente tipo:

- http://www.wto.org/spanish/news_s/news11_s/tnc_dg_infstat_29mar11_s.htm
- http://www.wto.org/french/news_f/news11_f/tnc_dg_infstat_29mar11_f.htm
- http://www.unicef.org/spanish/media/media_4537.html
- http://www.unicef.org/french/media/media_4537.html

Una vez seleccionados los sitios de los que vamos a recuperar la información, el siguiente paso tiene que ver con la selección de un sistema de recuperación de información y con la formulación de una ecuación de búsqueda que permita recuperar textos paralelos

relacionados con el encargo. En este sentido, una posible ecuación con la que interrogar a Google puede ser la siguiente:

- "modelo agroalimentario"|"escuela agropecuaria"|"distribución de agua"|"cooperación al desarrollo" site:codexalimentarius.net OR site:who.int/ OR site:unicef.org/ OR site:wto.org OR site:ec.europa.eu/agriculture/ -ext:pdf

Con ella, conseguimos recuperar textos paralelos que contienen las palabras clave de los títulos de las ponencias del congreso dentro de los sitios de CODEX Alimentarius (Normas alimentarias FAO/OMS), la Organización Mundial de la Salud, Unicef, la Organización Mundial del Comercio, y un directorio de la Comisión Europea dedicado a la agricultura. Asimismo, descartamos de la recuperación de textos los archivos PDF, pues prevemos que su conversión y posterior alineación puede presentar problemas (de todos modos, puede incluirse este formato de archivos y estudiar si, por el contrario, permite una correcta alineación).

Recuperados los archivos es posible editar la página de resultados con vistas a obtener un listado de URL referidos a textos paralelos en español y generar, mediante procesos de buscar y reemplazar, un nuevo listado de URL, esta vez, referido a sus correspondientes versiones en francés.

Generados dichos listados, pasamos a la descarga de archivos al disco duro. Cualquier gestor de descargas o volcador que permita, como WinHTtrack, descargar archivos específicos de una lista de URL puede ser válido.

Realizada la descarga, conviene, por supuesto, comprobar que el número de archivos en español se corresponde con el número de archivos en francés, y que, además, se trata de los mismos archivos. Ello es posible, por ejemplo, comprobando los archivos de auditoría que suelen generar los programas anteriores o confrontando los archivos de cada lengua en dos ventanas distintas. En caso de que no se corresponda dicho número, es posible eliminar, con el propósito de no perder más tiempo, aquellos archivos que solo contengan una versión en una lengua.

El siguiente paso consiste en convertir en dos archivos TXT los ficheros de cada lengua, y alinearlos. Para ello, es posible emplear cualquier programa que permita convertir por lotes archivos HTM, PDF, DOC, etc. a txt, así como emplear el comando *copy* de MS-DOS para juntar los archivos en uno solo. La versión en línea de You Align permite ahora alinear de manera automática los archivos, siempre que no superen 1 Mb.

Análisis del corpus

El corpus compilado tiene un total de 43 archivos en español y otros tantos en francés. Según Antconc, el subcorpus español tiene un tamaño de 432.724 *tokens,* y el francés, de 435.709. Los textos proceden fundamentalmente de www.unicef.org, www.who.int, www.wto.org. Solo se recuperaron textos a partir de dos palabras de la ecuación de búsqueda: *distribución de agua* y *cooperación al desarrollo*. Con el resto de palabras no conseguimos textos en las dos versiones, debido, en esencia, a la omisión del formato

PDF y a que algunos textos solo estaban disponibles en un idioma. En cualquier caso, es posible extender las búsquedas a partir, por ejemplo, de las palabras clave más representativas de los sitios web de las organizaciones a las que pertenecen los ponentes (corpus comparable monolingüe *ad hoc*). Los listados de palabras y grupos de palabras más frecuentes (*cf.* apéndice) ayudan a dar una idea del contenido del corpus *ad hoc.*

Glosario: creación, estudio y utilidad para la interpretación

Es posible automatizar la extracción bilingüe de la terminología contenida en el corpus paralelo con la aplicación Linguoc LexTerm (Oliver *et al.,* 2007). Utilizando los listados de palabras gramaticales que proporciona este programa y extrayendo terminología de entre dos y cuatro palabras hemos obtenido un glosario de unas 170 entradas. El análisis de este glosario deja vislumbrar ya una serie de ventajas, toda vez que el intérprete:

- Enriquece su léxico *ad hoc.*
- Es consciente de los rasgos diacrónicos, diatópicos, diastráticos e idiolectales de los ponentes interpretados.
- Enriquece su competencia fraseológica.
- Está sobre aviso en relación con los posibles falsos amigos que suelen aparecer al abordar las cuestiones analizadas en este tipo de eventos.
- Enriquece su capacidad de reconocimiento y aplicación de la variación denominativa.

Por otra parte y centrándonos ya en la utilidad de este tipo de glosarios para la interpretación, creemos conveniente hacer una distinción espacio-temporal. Dicho de otro modo, tenemos por más oportuno por arrojar, creemos, mayor luz sobre la cuestión estudiada, analizar las ventajas y los inconvenientes que presenta el glosario antes, durante y después de la propia interpretación. Así las cosas, consideramos que antes de la interpretación las ventajas en relación con la compilación del corpus paralelo son la semiautomatización del proceso y la instalación y detonación de la subcompetencia instrumental; las ventajas respecto de la elaboración del glosario vendrían dadas por la extracción asistida por ordenador de terminología bilingüe y el ahorro de tiempo. Finalmente, por lo que se refiere a las ventajas de la explotación del glosario antes de la interpretación, podemos afirmar que el intérprete dispone de terminología contextualizada, el glosario obtenido se puede utilizar en sistemas de explotación de glosarios multilingües para intérpretes como Interplex y, por último, favorece el aprendizaje y enriquecimiento *ad hoc* del léxico especializado. Los inconvenientes respecto de la compilación del corpus paralelo antes de la interpretación serían la posibilidad de "toparnos" con ruido y silencio informativos, la necesidad de contar con unos conocimientos y pericia informáticos previos, las limitaciones que la técnica todavía presenta (imposibilidad de procesar ciertos archivos con esta metodología basada en el software libre), limitaciones derivadas de la ausencia de las diversas lenguas de trabajo en la red. Respecto de los inconvenientes previos a la prestación referida a la

elaboración del glosario debemos mencionar el "peligro" de utilizar la frecuencia como criterio de extracción y la obligación de tener que seleccionar manualmente el candidato a término. Finalmente, el inconveniente respecto de la explotación del glosario pasa por el hecho de no poder combatir, ni siquiera con esta mitología, los estragos que pueden provocar una desviación o digresión imprevista por parte del orador en relación con el tema de la interpretación.

Por otra parte, las ventajas de la metodología durante y después de la interpretación en cuanto a la compilación del corpus paralelo se ubican en la posibilidad de afinar las ecuaciones de búsqueda según lo escuchado en las conferencias. Respecto de la elaboración del glosario: actualización a partir de nuevos textos, aumento de la precisión y de la pertinencia de los términos, retroalimentación a partir de lo escuchado en las conferencias. Finalmente, respecto de la explotación del glosario y tal y como hemos comentado más arriba: el intérprete dispone de terminología contextualizada, puede utilizar el glosario en sistemas de explotación de glosarios multilingües para intérpretes (Interplex) y gana en precisión y pertinencia en su prestación.

Vías futuras de investigación

Tal y como hemos podido comprobar en la bibliografía especializada en interpretación consultada y a la luz de nuestra propia experiencia como profesionales de la interpretación, parece una tendencia comportamental generalizada aquella según la cual el intérprete prepara sus futuras prestaciones primero desde un punto de vista temático y, a continuación, desde un punto de vista terminológico. Dicho de otro modo, el intérprete lee la documentación con la que cuenta y, a partir de ahí, extrae los términos o palabras clave a partir de los cuales se hace sus propios glosarios monolingües que luego traduce a sus lenguas de trabajo y asimila. Nosotros, sin embargo, teniendo en cuenta los resultados obtenidos en la presente investigación, creemos posible establecer la hipótesis contraria, la cual quedaría enunciada de la siguiente manera: es posible que a través de la utilización de la metodología de explotación y aplicación de corpus presentada en estas páginas y recorriendo el camino inverso al recorrido tradicionalmente por los intérpretes, es decir, yendo del término al documento y no del documento al término ganemos en pertinencia y precisión informativa y ganemos un tiempo precioso que el intérprete podrá invertir en asimilar e interiorizar la nueva información y terminología a la que deberá enfrentarse en cabina algunas horas más tarde.

APÉNDICE

Palabras clave del corpus paralelo

servicios	services
productos	pays
países	produits
sector	développement
desarrollo	secteur
empresas	commerce
millones	production
producción	droits
OMC	OMC
actividades	activités
comercio	Services
país	tourisme
Ley	entreprises
derechos	millions
transporte	Ministère
Acuerdo	État
Estados	prix
Gobierno	Accord
comerciales	Loi

Grupos de 4-gramas del corpus paralelo

la cláusula de habilitación	la clause d habilitation
el por ciento de	dans le cadre de
de dólares ee uu	pour cent de la
los países en desarrollo	consulté sur http www
millones de dólares ee	de la clause d
en el caso de	des pays en développement
en el marco del	millions de dollars eu
el por ciento del	au titre de l
en el marco de	disponible sur http www
consultado en http www	les pays en développement
de las políticas comerciales	le cadre de l
por ciento de la	est de pour cent
puede consultarse en http	de l accord sur
con el medio ambiente	examen des politiques commerciales
de la cláusula de	dans le cadre du
examen de las políticas	l accord sur les

Ejemplo de terminología extraída en el glosario

Cláusula de Habilitación	Clause d'habilitation
países en desarrollo	pays en développement
República Centroafricana	République centrafricaine
francos CFA	francs CFA
servicios financieros	Services financiers
Examen de las Políticas	examen des politiques commerciales
Políticas Comerciales	politiques commerciales
derechos e impuestos	droits et taxes
celebración de consultas	ouverture de consultations
productos agrícolas	produits agricoles
derecho de autor	droits d'auteur
necesidades de desarrollo	besoins du développement
propiedad intelectual	propriété intellectuelle
medio ambiente	environnement
Trinidad y Tabago	Trinité-et-Tobago

Bibliografía

Abril Martín, M. (2006): *La interpretación en los servicios públicos: caracterización como género, contextualización y modelos de formación. Hacia unas bases para el diseño curricular.* Tesis doctoral, Universidad de Granada.

Beeby, A.; Rodríguez Inés, P. & Sánchez Gijón, P. (Eds.). (2009). *Corpus Use and Translating.* Amsterdam/Philadelphia: John Benjamins.

Bernardini, S. & Zanettin F. (Eds.). (2000). *I corpora nella didattica della traduzione: Corpus use and learning to translate.* Bologna: CLUEB.

Bianchessi et al. (2011): El glosario como herramienta en la interpretación consecutiva. Estudio de un caso práctico: la conciliación en Ruanda. *Entreculturas, 3.*

Blasco, M. J. & Jiménez, A. (2003): Elaboración de glosarios terminológicos para interpretar. En Collados, A. *et al.* (Eds.), *La evaluación de la calidad en interpretación: docencia y profesión* (pp. 225-234). Granada: Comares.

Castillo Rodríguez, C. (2009): La elaboración de un corpus paralelo ad hoc multilingüe. *Tradumàtica, 7.*

Choi, J. (2005): Qualité et préparation de l'interprétation. Évolution des modes de préparation et rôle de l'Internet. *Meta, 50, 4*.

Choi, J. W. (1998): *Introduction to Interpretation and Translation.* Seoul: Shinronsa.

Corpas Pastor, G. (2002): Traducir con corpus: de la teoría a la práctica. En García Palacios, J. & M.ª T. Fuentes Morán (Eds.), *Texto, terminología y traducción* (pp. 189-226). Salamanca: Almar.

Corpas Pastor, G. (2008): *Investigar con corpus en traducción: los retos de un nuevo paradigma.* Frankfurt am Main: Peter Lang.

Foulquié, A. & Navarro, N. (2011) Calidad de la búsqueda de información en la preparación de la interpretación: su importancia en la docencia de técnicas de interpretación. Póster presentado en el *II Congreso Internacional sobre Calidad en Interpretación*, 24-26 de marzo de 2011, Almuñécar (Granada).

Gallego Hernández, D. (2010): Acquiring instrumental sub-competence by building do-it-yourself corpora for business translation. *Proceedings of The International Symposium on Using Corpora in Contrastive and Translation Studies*. Ormskirk: Edge Hill University.

Gallego Hernández, D. (2011): Documentación aplicada a la traducción económica, comercial y financiera: estrategias de compilación ad hoc de textos paralelos, *V Congreso AIETI.* Castellón: Universidad.

Gile, D. (1985): Les termes techniques en interprétation simultanée. *Meta, 30, 3,* 199-210.

Gile, D. (1995): *Regards sur la recherche en interprétation de conférence*, Lille, Presses Universitaires de Lille.

Gulyás, R. (2003): Become a cybrarian – make the best of the Web. En Szabó, C. et al. (Eds.), *Interpreting: From Preparation to Performance. Recipes for practitioners and Teachers.* British Council: Budapest.

Nolan, J. (2005): *Interpretation Techniques and Exercises.* Toronto: Multilingual Matters.

Oliver, A.; Vázquez, M. & Moré, J. (2007): Linguoc Lexterm: una eina d'extracció automàtica de terminologia gratuïta. *Translation Journal, 11, 4.*

Sánchez Gijón, P. (2004): *L'ús de corpus en la traducció especialitzada: compilació de corpus ad hoc i extracció de recursos terminològics.* Barcelona: Universitat Pompeu Fabra.

Seleskovitch, D. & Lederer, M. (2002): *Pédagogie raisonnée de l'interprétation.* Paris: Didier érudition.

Shlesinger, M. (1998): Corpus-based Interpreting Studies as an Offshoot of Corpus-based Translation Studies, *Meta, 43, 4,* 486-493.

SZABÓ, C. ET AL. (Eds.). (2003): *Interpreting: From Preparation to Performance. Recipes for practitioners and Teachers.* Budapest: British Council.

WADENSJÖ, C. (1998): Community interpreting. En Baker, M. (Ed.), *Encyclopaedia of Translation Studies* (pp. 33-37). Manchester: Multilingual Matters.

ZANETTIN, F.; BERNARDINI, S. & STEWART, D. (Eds.) (2003). *Corpora in Translator Education.* Manchester/Northampton: St. Jerome.

El fenómeno *pro-drop* en portugués de Brasil y español peninsular

Iria Gayo
U. de Santiago de Compostela

Luz Rello
Universitat Pompeu Fabra

Resumen: español y portugués son lenguas pro-drop. No obstante, diversos estudios indican que el portugués muestra diferencias respecto a este fenómeno entre sus variedades europea y brasileña. Estas diferencias han llevado a considerar el portugués de Brasil una lengua parcialmente pro-drop. En este trabajo se analiza el fenómeno pro-drop en portugués de Brasil a través de una comparación con el español peninsular, utilizando para ello corpus comparables. Nuestros resultados tratan las diferencias entre las dos lenguas prestando especial atención a las distintas posibilidades de realización sintáctica del sujeto.

Palabras clave: pro-drop; corpus comparable; sujeto explícito; sujeto nulo; construcción impersonal; pasiva refleja.

Abstract: Spanish and Portuguese are pro-drop languages. However, the differences between some varieties of Portuguese have led researchers to consider Brazilian Portuguese as partial pro-drop language. This paper explores the pro-drop phenomenon in Brazilian Portuguese and compares it to Iberian Spanish using comparable corpora. Our results discuss the differences found between these two languages with a special focus on the syntactic possibilities of subject realization.

Keywords: pro-drop; comparable corpora; explicit subject; null subject; impersonal construction; reflex passive.

Introducción

Este artículo presenta un estudio de corpus de tipo cuantitativo en el que se comparan dos variedades lingüísticas: español peninsular (en adelante, EP) y portugués de Brasil (en adelante, PB). El objetivo de este estudio es verificar si el PB, como lengua *parcialmente pro-drop*, presenta diferencias respecto al EP en relación al fenómeno *pro-drop*. De los múltiples rasgos relacionados con este fenómeno, nos ocuparemos de analizar las distintas posibilidades de expresión sintáctica del sujeto. Utilizaremos para ello un corpus comparable formado por textos escritos originalmente en EP y en PB.

La motivación de nuestra investigación es doble: teórica y práctica. En primer lugar nos interesa comprobar si los datos de nuestro corpus corroboran o refutan la hipótesis del PB como lengua *parcialmente pro-drop*. En segundo lugar, este tipo de investigación comparativa puede ser útil para aplicaciones de áreas relacionadas con el procesamiento del lenguaje natural como la traducción automática.

La estructura del artículo es la que sigue: en el apartado 1 presentamos el fenómeno *pro-drop* y sus pecualiaridades respecto al PB; en la sección 2 detallamos las características del análisis comparativo llevado a cabo así como del corpus utilizado; en el apartado 3 presentamos los resultados de nuestro análisis; finalmente, en el punto 4 recogemos las principales conclusiones del estudio y algunas ideas para investigaciones futuras.

1. El fenómeno pro-drop y la realización del sujeto

En términos generales se consideran lenguas *pro-drop* aquellas que no requieren la aparición *obligatoria* de un sujeto explícito, es decir, aquellas que pueden presentar o no un sujeto realizado sintácticamente (Chomsky, 1981).

Son lenguas *pro-drop* el español (a), el portugués (b) o el italiano (c). Este tipo de lenguas se oponen a aquellas que, como el inglés (d) o el francés (e), requieren obligatoriamente la presencia de un sujeto explícito:

a. *(Ella)*[89] Vino ayer.

b. *(Ela)* Veio ontem.

c. *(Lei)* É venuta ieri.

d. *She* came yesterday.

e. *Elle* est venue hier.

La omisión del sujeto recibe en lingüística diferentes etiquetas: *sujeto nulo, sujeto vacío, sujeto elíptico, sujeto elidido, sujeto tácito*, etc. (Rello, 2010a: 6), mientras que en el campo del procesamiento del lenguaje natural suele denominarse *pronombre cero* (Rello, 2010a: 6). En este trabajo utilizaremos *sujeto explícito* para los sujetos realizados sintácticamente y *sujeto nulo* para los no realizados sintácticamente.

89 Los paréntesis en los ejemplos significan opcionalidad.

1.1. Portugués de brasil: lengua parcialmente pro-drop

Como hemos visto, el español y el portugués se consideran lenguas *pro-drop*. Sin embargo, en lo que respecta al portugués, diversos estudios (Barbosa, Duarte & Kato, 2003, 2005) han constatado que existen diferencias en relación a este aspecto sintáctico entre dos de sus variantes geográficas: el portugués europeo (en adelante, PE) y el PB. Mientras que el PE se comporta como una lengua prototípicamente *pro-drop*, el PB muestra una serie de características que han llevado a definirlo como una *lengua parcialmente pro-drop* (Duarte, 1995; Kato & Negrão, 2000; Barbosa, Duarte & Kato, 2003).

La principal característica que define el PB como lengua parcialmente *pro-drop* es la *pérdida progresiva del sujeto nulo*, substituído fundamentalmente por estructuras de tipo pronominal:

> Duarte (1993, 1995) shows that spoken Brazilian Portuguese (henceforth BP) is gradually displaying an increase in the use of overt pronominal subjects, even with non-human antecedents.
>
> **These appear in contexts where a null subject would show up in EP**[90], namely when they are anaphorically related to a matrix subject. (Barbosa, 2005: 5).

Duarte (1993) muestra además que esta pérdida progresiva de sujetos nulos afecta más a la primera y la segunda persona gramatical que a la tercera. Este hecho parece ir en contra de lo que se esperaría si se tiene en cuenta la morfología verbal del PB, donde segunda y tercera persona del singular se solapan (f) de manera que *la tercera persona del singular es la forma no marcada*.

f. Você *ama.* / Ele *ama.*

Tú amas. / Él ama.

La pérdida progresiva del sujeto nulo en PB ha sido estudiada tanto desde una perspectiva diacrónica en la variante hablada (Duarte, 1993, 1995), como sincrónica y comparada con el PE en la variante escrita (Barbosa, Duarte & Kato, 2003, 2005). Asimismo, se han estudiado algunas consecuencias sintácticas que esta tendencia genera (Kato, 2009; Cavalcante & Duarte, 2008).

2. Análisis comparativo del sujeto en EP y PB

Partiendo de la idea de que el PB es una lengua *parcialmente pro-drop* y el EP una lengua *pro-drop*, planteamos un estudio comparativo EP-PB centrado en el análisis de la realización del sujeto en ambas lenguas. Se trata por tanto de un estudio en la línea de Barbosa, Duarte y Kato (2003, 2005). Nuestra intención es comprobar si nuestro corpus muestra, como cabría esperar, diferencias entre ambas lenguas en la elección de las distintas posibilidades para la realización del sujeto. Tres son los aspectos en los que nos centramos:

1. ¿Muestra el PB diferencias respecto al EP en la utilización de sujetos explícitos *vs.* sujetos nulos?

90 La sigla EP se corresponde con "European Portuguese" (portugués europeo).

2. ¿Existen diferencias entre ambas lenguas en el uso de otras estructuras sintácticas como las impersonales o la pasiva refleja?

3. ¿Muestra el PB diferencias en su preferencia por sujetos explícitos *vs*. sujetos nulos según la persona gramatical?

El primer aspecto constituye el punto de partida de este trabajo: comparar el uso de sujetos explícitos frente a sujetos nulos en EP y PB. Nuestra intención es comprobar si, como lengua *parcialmente pro-drop*, el PB presenta una mayor inclinación por la utilización de sujetos explícitos frente sujetos nulos que una lengua *prototípicamente pro-drop* como el EP.

Con el segundo aspecto se pretende comparar la aparición en EP y PB de dos estructuras sintácticas relacionadas también con la realización del sujeto: pasiva refleja e impersonales. La pasiva refleja (g) constituye un tipo especial de construcción donde el sujeto sintáctico se comporta como objeto directo y suele ir colocado tras el verbo. Las impersonales (h), a su vez, se caracterizan por no tener sujeto.

g. *Se venden pisos.*

h. *Hubo fuegos artificiales.*

Con el tercer aspecto se pretende verificar la tendencia observada por Duarte (1993) para el PB (en este caso no hay comparación con el EP).

2.1. El corpus comparable

El corpus utilizado para nuestro estudio está compuesto por un conjunto de textos extraídos de dos corpus comparables EP/PB: ESZIC_ES (Rello, 2010b) (textos en EP) y ESZIC_PT[91] (Rello & Gayo, 2011) (textos en PB). ESZIC_ES y ESZIC_PT presentan las siguientes características:

1. 17 textos por lengua: escritos originalmente en EP y en PB; correspondientes al mismo periodo de tiempo
2. textos parseados (ESZIC_ES con el parser para el español de *Connexor*[92] y ESZIC_PT con el parser *PALAVRAS*[93])
3. 2 géneros: legal (8 textos por lengua) y de salud (9 textos por lengua)
4. anotados manualmente: 3398 instancias anotadas en ESZIC_ES; 3547 instancias anotadas en ESZIC_PT

La anotación de ambos corpus se ciñe a las formas verbales finitas. Cada verbo finito tiene asignada una categoría dependiendo de su naturaleza y de la de su sujeto correspondiente. Se manejan tres grandes categorías[94] (Rello, 2010b):

91 Ambos corpus se encuentran disponibles como recurso electrónico en la siguiente url: http://www.luzrello.com/Projects.html
92 http://www.connexor.eu/technology/machinese/index.html
93 http://beta.visl.sdu.dk/visl/pt/
94 Todos los ejemplos utilizados a partir de ahora pertenencen a ESZIC_ES y ESZIC_PT. Los ejemplos en portugués se traducen al español. Se marca en negrita el verbo anotado.

(1) Sujeto: para los casos de sujeto explícito.

i. *Las fuentes **son** la ley, la costumbre y los principios generales del derecho.*

(2) Cero: para los casos de sujeto nulo.

j. ø *Las leyes no **tendrán** efecto retroactivo si Ø no dispusieren lo contrario.*

(3) Impersonal: para las oraciones impersonales.

k. *Cuando **hay** un diagnóstico.*

Estas tres grandes categorías están divididas a su vez en 13 subcategorías que tienen en cuenta otros aspectos gramaticales tales como la voz o la naturaleza del constituyente que funciona como sujeto (Rello & Gayo, 2011):

(1) Sujeto explícito en oración activa.

l. *Os agentes consulares brasileiros **poderão** servir de oficiais públicos na celebração e aprovação dos testamentos de brasileiros.*

Los agentes consulares brasileños **podrán** servir de oficiales públicos en la celebración y aprobación de los testamentos de brasileños.

(2) Sujeto explícito en oración pasiva:

m. *Redação dada por a lei nº 10.149, de 21.12.2000) citado por 1§ 2o a empresa estrangeira **será notificada** e intimada de todos os atos processuais [...].*

Redacción dada por la ley nº 10.149, de 21.12.2000) citado por 1§ 2o la empresa extranjera será notificada e intimada de todos los actos procesuales […].

(3) Sujeto explícito en oración pasiva refleja.

n. *Ao mesmo tempo em que se **visita** outra pessoa essa experiência muda o próprio visitante.*

Al mismo tiempo que se visita otra persona esa experiencia cambia al propio visitante.

(4) Sujeto omitido en oración activa.

ñ. *Ø C**onvergem** também no conceito de que os transtornos mentais surgem a partir de interrelações dimensionais, complexas, em múltiplos níveis [...].*

Ø **Convergen** también en el concepto de que los trastornos mentales surgen a partir de las interrelciones dimensionales, complejas, en múltiples niveles [...].

(5) Núcleo omitido en el sujeto en oración activa.

o. *Em a clínica, é comum a sobreposição de sintomas, o Ø que **promove** dificuldades na distinção de categorias tão diversas.*

En clínica, es común la superposición de síntomas, lo Ø que promueve dificultades en la distinción de categorías tan diversas.

6) Sujeto no nominal en oración activa.

p. ***É*** *possível que um processo menos rebuscado não comprometa a qualidade do instrumento final.*

Es posible que un proceso menos rebuscado no comprometa la calidad del instrumento final.

(7) Sujeto omitido en oración pasiva.

q. *Ø* ***É tomado*** *por o desejo de ser amado por ela e de expressar o amor que sente por ela.*

Ø Es tomado por el deseo de ser amado por ella y de expresar el amor que siente por ella.

(8) Núcleo omitido en el sujeto en oración pasiva refleja.

r. *Os Ø que, embora naturalmente divisíveis, se* ***consideram*** *indivisíveis por lei, ou vontade das partes.*

Los Ø que, aunque naturalmente divisibles, se consideran indivisibles por la ley, o voluntad de las partes.

(9) Sujeto no nominal en oración pasiva refleja.

s. *Para a retificação de dados, quando não se* ***prefira*** *fazê-lo por processo sigiloso judicial ou administrativo.*

Para la rectificación de los datos, cuando no se prefiera hacerlo por proceso secreto judicial o administrativo.

(10) Núcleo omitido en el sujeto en oración pasiva.

t. *O Ø que* ***foi discutido*** *e solucionado na etapa 3.*

Lo Ø que fue discutido y solucionado en la etapa 3.

(11) Sujeto no nominal en oración pasiva.

u. *Sobre a auto-percepção do estado de saúde,* ***deve ser observado*** *que a maioria […].*

Sobre la auto-percepción del estado de salud, debe ser observado que la mayoría […].

(12) Construcción impersonal con se.

v. *Também* ***há*** *uma distinção entre a qualidade de vida global, as dimensões ou domínios da qualidade de vida e os componentes de cada dimensão.*

También **hay** una distinción entre la calidad de vida global, las dimesiones o dominios de la calidad de vida y los componentes de cada dimensión.

(13) Construcción impersonal sin se

(w) *Para tal finalidade,* ***assistia****-se aos quinze minutos de filmagem quantas vezes fossem necessárias na atribuição do escore do item em questão para depois passar ao item seguinte.*

Para tal finalidad, se **asistía** a los quince minutos de filmación cuantas veces fuesen necesarias en la atribución de la puntuación del ítem en cuestión para después pasar al ítem siguiente.

Como apuntábamos, el corpus comparable utilizado en este trabajo está compuesto por los textos del género de salud de ESZIC_ES y ESZIC_PT (9 textos por lengua). Se trata de textos de tipo científico del campo de la psiquiatría[95].

2.2. Categorías tenidas en cuenta para el análisis

Como se ha indicado, en ESZIC_ES y ESZIC_PT se utilizan 13 posibles subcategorías de anotación para cada verbo finito. De estas 13 subcategorías (Rello & Gayo, 2011) se han utilizado para este estudio:

1. Sujetos explícitos: de carácter nominal o pronominal. Todas las voces (activa, pasiva y pasiva refleja).
2. Sujetos nulos: todos los tipos excepto los de *núcleo omitido*. Todas las voces (activa, pasiva y pasiva refleja).
3. Impersonales: los dos tipos, *impersonales* e *impersonales con se* (Rello, 2010a).

3. Resultados

Recordemos que tres son los aspectos que nos interesa analizar en nuestro estudio comparativo:

- Uso de sujetos explícitos *vs.* sujetos nulos en EP y PB.
- Utilización de pasiva refleja e impersonales en EP y PB.
- Presencia de sujetos explícitos *vs.* sujetos nulos en primera y tercera persona gramatical (a partir de ahora 1p y 3p) en PB.

3.1. Sujeto explícito vs. Sujeto nulo

Nuestro corpus ofrece los siguientes datos:

Tabla 1. Sujeto explícito vs. Sujeto nulo en EP y PB.

	Sujeto Explícito*	Sujeto Nulo	TOTAL
EP	1827 (63%)	1063 (37%)	2890
PB	2497 (78%)	683 (21%)	3180

95 La decisión de no utilizar también los textos del género legal responde a la idea de que el lenguaje del género legal es un lenguaje más formular y menos natural que el de los textos científicos, con construcciones sintácticas propias y particulares, en muchos casos ajenas al lenguaje general. No obstante, en un futuro no se descarta el análisis de los textos legales.

Los datos recogidos en la Tabla 1 muestran que EP y PB sí presentan diferencias en su elección de sujetos explícitos *vs.* sujetos nulos: el PB se inclina más que el EP por el uso de sujetos explícitos (78% frente a 63%), mientras que el EP tiene una mayor preferencia que el PB por la utilización de sujetos nulos (37% frente a 21%).

Estos datos apoyan la hipótesis del EP como lengua *pro-drop* y el PB como lengua parcialmente *pro-drop*.

3.2. Pasiva refleja e impersonales

En primer lugar se analiza la presencia de la pasiva refleja en contraposición con los otros dos tipos de voz anotados: activa y pasiva. Este factor voz se combina a continuación con la aparición de sujeto explícito *vs.* sujeto nulo.

En segundo lugar nos ocupamos de las construcciones impersonales.

3.2.1. Sujeto explícito *vs.* Sujeto nulo + voz

Los datos globales para los tres tipos de voz anotados (A-activa, P-pasiva y PR-pasiva refleja) son los siguientes:

Tabla 2. Uso de la voz.

	A	P	PR
EP	2943 (90%)	65 (2%)	264 (8%)
PB	2933 (87%)	359 (11%)	86 (2%)

Ambas lenguas muestran valores similares en la utilización de la voz activa (90% frente a un 87%). En cuanto a voz pasiva, el PB presenta valores más altos que el EP (11% frente a un 2%). Finalmente, en lo referente a la pasiva refleja, el PB presenta una tendencia clara a usar menos esta estructura que el EP (2% frente a un 8%).

Combinando el factor sujeto explícito *vs.* sujeto nulo con el factor voz, obtenemos los siguientes resultados:

Tabla 3. Sujeto explícito (SE) vs. Sujeto nulo (SN) + voz.

	SE_A	SE_P	SE_PR	SN_A	SN_P	SN_PR	TOTAL
EP	1836 (56%)	56 (2%)	**221 (7%)**	1107 (34%)	**9 (0,3%)**	43 (1%)	3272
PB	2275 (67%)	334 (10%)	**83 (2%)**	658 (19%)	**25 (0,8%)**	3 (0,08%)	3378

El análisis en detalle muestra datos esperables para casi todas las combinaciones: el EP presenta en general un mayor uso que el PB de las estructuras con sujeto nulo, mientras que el PB presenta una mayor utilización de estructuras con sujeto explícito. Solo dos

combinaciones van en contra de esta tendencia: las construcciones con sujeto explícito en pasiva refleja (SE_PR) y las construcciones con sujeto nulo en pasiva (SN_P). Las construcciones con sujeto explícito en pasiva refleja son, contrariamente a lo esperado, más comunes en EP que en PB. Este hecho se debe probablemente a la escasa presencia de la pasiva refleja en PB (cf. Tabla 2). Las construcciones con sujeto nulo y voz pasiva son más comunes en PB que en EP. Esto a su vez puede deberse a la preferencia clara del PB por la utilización de la voz pasiva (cf. Tabla 2).

3.2.2 Impersonales

El corpus ofrece los siguientes datos:

Tabla 4. Presencia de construcciones impersonales.

	EP	PB
IMPERSONALES	64 (2%)	38 (1%)
IMPERSONAL	45 (1%)	17 (0,5%)
TOTAL	**109 (3%)**	**55 (1,5%)**

Al igual que ocurría con la pasiva refleja, en términos globales el PB presenta menos casos de impersonales que el EP, concretamente, un 50% menos. Esta diferencia se mantiene además para los dos tipos de impersonales analizados.

3.3. Sujeto explícito vs. Sujeto nulo + persona gramatical en pb

El último aspecto que nos interesa analizar concierne tan solo al PB. Recordemos que nos interesa verificar la tendencia observada por Duarte (1933) según la cual 1p y 2p presentan más casos de sujetos explícitos que 3p, que se inclina más por el uso de sujetos nulos, contrariamente a lo esperado.

En nuestro caso, comparamos 1p y 3p, ya que no existen casos de 2p en el corpus (cf. apartado 1.1.).

La siguiente tabla recoge los resultados al respecto:

Tabla 5. Sujeto explícito vs. Sujeto nulo + persona gramatical en PB.

	1P	3P
Sujeto Explícito	14 (12%)	1756 (83%)
Sujeto Nulo	103 (88%)	360 (17%)
TOTAL	117	2116

Nuestros datos no verifican la tendencia observada por Duarte (1993), sino que la contradicen: los sujetos nulos son mucho más comunes para la 1p (88%) que para la 3p (17%), mientras que los sujetos explícitos son mucho más comunes para la 3p (83%) que

para la 1p (12%). Recordemos, sin embargo, que esta sería la tendencia lógica desde el punto de vista lingüístico, ya que la 3p es la no marcada en PB (cf. apartado 1.1.).

4. Conclusiones

Hemos presentado un análisis cuantitativo utilizando un corpus compuesto por textos de dos lenguas diferentes: EP y PB. El objetivo principal de la comparación era comprobar si nuestro corpus mostraba las diferencias esperables al ser el EP una lengua *pro-drop* y el PB una lengua parcialmente *pro-drop*. Estas diferencias se manifiestan fundamentalmente en la utilización de sujetos explícitos *vs.* sujetos nulos en EP y PB. Gracias a la riqueza de anotación de nuestro corpus, hemos podido comparar también la relación entre este factor y la voz, el uso de estructuras impersonales y la presencia de la pasiva refleja en ambas lenguas. Por último, hemos analizado el uso de sujetos explícitos *vs.* sujetos nulos en 1p y 3p en PB.

En suma, nuestros datos muestran que:

- El PB presenta, frente al español:
    1. más casos de sujeto explícito que de sujeto nulo. Estos datos apoyan la hipótesis del PB como una lengua parcialmente *pro-drop*.
    2. menos casos tanto de construcciones impersonales como de pasiva refleja. Esta tendencia tal vez pueda relacionarse con la del punto anterior: el PB se inclina más a usar estructuras que impliquen la expresión de un sujeto explícito.
    3. más casos de voz pasiva que el español, siendo incluso más abundantes que en español los casos de sujeto nulo en esta voz.
- El PB muestra un uso mucho más abundante de sujetos explícitos con 3p, mientras que la 1p prefiere claramente los sujetos nulos.

Estos datos contradicen la tendencia mostrada en Duarte (1993), no obstante, parecen confirmar el comportamiento lingüístico esperable teniendo en cuenta que la 3p es la no marcada en el PB.

4.1. Trabajo futuro

Tres son los aspectos que nos interesa abordar en un futuro.

En primer lugar, sería enriquecedor realizar el mismo análisis que ha sido presentado aquí para los textos de tipo legal de ESZIC_ES y ESZIC_PT.

Por otro lado, sería interesante poder analizar el tipo de unidad que desempeña la función de sujeto explícito tanto para EP como para PB, con el fin de comprobar si este último presenta realmente más casos de sujetos prononimales abiertos, incluso sin antecedente humano (Barbosa, Duarte & Kato, 2003: 5).

Finalmente, sería esclarecedor poder analizar en detalle en qué contextos (sintácticos o discursivos) el EP se inclina por los sujetos nulos mientras que el PB prefiere la presencia de sujetos explícitos.

Referencias

Barbosa, P., Duarte, M. & Kato, M. A. (2003). Sujeitos indeterminados em PE e PB. *Boletim da Associação Brasileira de Linguística, 26*, 405-409.

Barbosa, P., Duarte, M. & Kato, M. A. (2005). Null Subjects in European and Brazilian Portuguese. *Journal of Portuguese Linguistics, 4*, 11-52.

Cavalcante, S. R. de O. & Duarte, M. (2008). The Subject Position in Brazilian Portuguese: the Embedding of a Syntactic Change. *University of Pennsylvania Working Papers in Linguistics*, *14*(2), article 8.

Chomsky, N. (1981). *Lectures on government and binding*. Berlín, New York: Mouton de Gruyter.

Duarte, M. (1993). Do pronome nulo ao pronome pleno: a trajetória do sujeito no português do Brasil. In I. Roberts & M. A. Kato (Eds.), *Português Brasileiro: Uma viagem diacrônica (Homenagem a Fernando Tarallo)* (pp. 107-128). Campinas: Editora da UNICAMP.

Duarte, M. (1995). *A Perda do Princípio "Evite pronome" no Português Brasileiro*. Ph.D. Dissertation. UNICAMP.

Kato, M. A. & Negrão, E.V. (Eds.) (2000). *Brazilian Portuguese and the Null Subject Parameter*. Frankfurt: Vervuert-Iberoamericana.

Kato, M. A. (2009). The partial pro-drop nature and the restricted VS order in Brazilian Portuguese. In M. A. Kato & E. V. Negrão (Eds.), *The Null Subject Parameter in Brazilian Portuguese*. Frankfurt: Vervuert.

Rello, L. (2010a). *Elliphant: A Machine Learning Method for Identifying Subject Ellipsis and Impersonal Constructions in Spanish.* Ph.D. Dissertation.

Rello, L. (2010b). From Zero to Hero: the Importance of Marking Ellipsis and Guidelines for its Annotation in Spanish. *Natural Language Processing & Human Language Technology, BULAG: Bulletin de Linguistique Appliquée et Générale, 35*.

Rello, L. & Gayo, I. (2011). Clasificación y Anotación de Sujetos en Portugués. Artículo presentado en el *XL Simposio Internacional y III Congreso de la SEL (Sociedad Española de Lingüística)*. Madrid, Spain, February, 7-10.

The deictic force of demonstrative determiners and definite articles in Spanish and Dutch. A perspective from a corpus of translated texts

Patrick Goethals
Faculty of Applied Linguistics, University of Ghent

Abstract

This paper explores the constructional meaning of NPs introduced by a demonstrative determiner and a definite article in Spanish and Dutch. First, I will briefly sketch a generic description of the meaning of the two constructions, focussing on three features: unique identifiability, deictic force and the informational value of the NP. Then, a contrastive analysis based on a bidirectional corpus of translated texts will show that the two paradigms differ cross-linguistically.The quantitative analysis of translation shifts clearly reveals a systematic difference. Through the qualitative analysis of a subgroup of examples, I will show that part of the translation shifts can be related to the fact that Dutch demonstrative determiners and definite articles have both less deictic force than their Spanish counterparts.

Keywords: demonstratives, definite articles, Spanish, Dutch, corpus of translated texts

Resumen:

En este estudio se describe el significado de los sintagmas nominales introducidos por un determinante demostrativo y un artículo definido en español y neerlandés. Empezaré por esbozar una descripción genérica del significado de las construcciones, haciendo referencia a tres características: la unicidad referencial, la fuerza deíctica y el valor informativo del sintagma nominal. Luego, un análisis basado en un corpus bidireccional de textos traducidos muestra que existe una diferencia contrastiva entre el neerlandés y el español que afecta a ambos paradigmas. El análisis cuantitativo de los cambios de traducción revela una diferencia sistemática. Mediante el análisis cualitativo de un subgrupo de ejemplos, intentaré demostrar que parte de los cambios de traducción se debe al hecho de que los determinantes demostrativos y los artículos definidos en neerlandés tienen menos fuerza deíctica que los paradigmas correspondientes en español.

Palabras clave: determinantes demostrativos, artículo definido, español, neerlandés, corpus de textos traducidos

This paper is a contribution to the contrastive analysis of demonstrative determiners and definite articles in Spanish and Dutch. The question that I wish to answer is whether the categories relate to each other in a similar way in both languages.

Existing studies on demonstratives mostly have adopted a monolingual or typological point of view. Relatively few details are known about specific contrastive differences. Do demonstratives have the same meaning effects in different languages? If there are differences, what are they, and how can they best be described?

1. The constructional meaning of demonstrative determiners and definite articles

In this section I will sketch the constructional meaning of Def NP and Dem NP by referring to three features: unique identifiability (a shared feature), deictic force and the informational value of the NP (two distinctive features).

Both Dem NPs and Def NPs are definite forms and refer to a discourse entity that is uniquely identifiable. For example, the discourse entity may be identifiable because the NP refers exophorically to a referent in the context (1), or anaphorically to an element in the previous discourse (2) or because the hearer and the speaker have a shared knowledge of the concept or entity (3). As we can see in (1-3), both Def NP and Dem NP occur in these contexts.

(1) Dame {el ~ ese} cenicero. (GRAE, §17.4b)

(2) Se preguntó si el suicidio tendría algo que ver con aquello. {El ~ Este ~ Ese} pensamiento lo estremeció.

(3) ¿Recuerdas {el ~ ese} museo que visitamos cuando estábamos en Madrid?

However, the discourse entity may also be identifiable exclusively on the basis of a restrictive definition in the NP. This means that there may be no anaphoric or exophoric link, and the concept may be entirely new for the hearer. In these cases, only Def NP is possible (4):

(4) En lo que sigue, examinaremos {el ~ ?ese} problema clasificatorio que plantean las formas demostrativas.

This means that the construction Def NP only conveys that the discourse entity can be identified on the basis of the information given in the NP (Epstein, 2002). It is up to the reader to infer whether the identifiable concept is coreferential with a prior discourse referent, whether it refers exophorically to the context, or whether it is indeed a new discourse referent. On the contrary, Dem NP not only marks the identifiability of the discourse entity, but it relates this to a deictic domain. The discourse referent is identifiable as a consequence of its being present in the deictic domain. This excludes its use in examples such as (4), where the discourse referent is totally new.

It also explains why Dem NPs cannot occur in the so-called associative anaphors (García Fajardo, 2006; GRAE, §17.4f). In (5), *la cocina* is a new referent, which is identified as the kitchen of *my* house through the association with the apartment-frame mentioned

in the preceding context. Dem NP is not possible because its use would imply that the referent is already present in the deictic domain, instead of being indirectly identifiable through an associative link with another concept:

(5) Ayer mis padres vieron por primera vez mi nuevo piso. Les gustaba mucho {la ~ ?esta ~ ?esa ~ ?aquella cocina}.

Figure 1 visualizes the constructional meaning of Dem NP and Def NP. The construction Dem NP instantiates a deictic center as the anchor of the deictic domain wherein the discourse entity is identified. The construction Def NP gives a description of the discourse referent (xxx), marks its identifiability (visualized with a square box), but does not per se evoke a deictic domain.

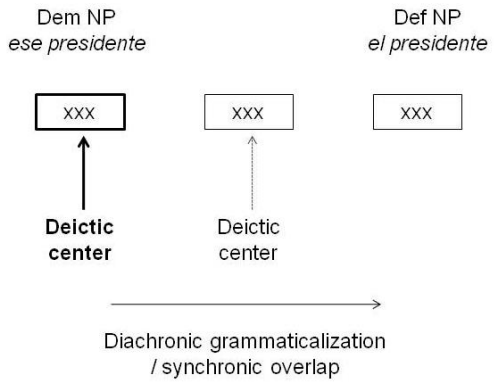


Figure 1. The deictic force of Dem NP and Def NP

It is important to note that the two poles are connected diachronically and synchronically. As is well-known, definite articles evolve diachronically from demonstrative determiners (Diessel, 1999:128-9). This diachronic evolution is a bleaching process that affects in particular the instantiation of the deictic domain: the demonstrative paradigm gradually loses deictic force. Also synchronically, Dem NP and Def NP show a considerable overlap: in contexts such as (1-3), they have the same referential denotation, and the difference between the two constructions would only consist in a more or less explicit evocation of the deictic domain.

Of course, in other contexts the difference in meaning between Def NP and Dem NP might lead to a different referential interpretation of the NP. In (6), Dem NP *esa literatura* refers to a type of literature defined in the previous context, whereas Def NP *la literatura* evokes the general concept of literature: Def NP marks that the concept of *literature* is identifiable as such.

(6) Los tiempos no están para {esa ≠ la} literatura (Pérez Reverte, *El Club Dumas*)

Thus far, I have focused on the marking of the deictic domain and the identification of the referent. However, as Maes & Noordman (1995) claim, [Dem NPs] "cannot be

described sufficiently in terms of their identificational capacity" (1995:259). Following these authors, Dem NPs also have a predicative informational function.

In Def NP, the NP has primarily an identifying function: the information must restrict the reference in such a way that the hearer will be able to identify the discourse referent in a unique way. Dem NP already marks the discourse entity is identifiable through its relation to the deictic center, and therefore the NP is not primarily assigned an identifying function, but rather a predicating function: the information in the NP is not primarily given to identify the referent, but rather to comment on it. Maes & Noordman (1995) call this the "modifying" or "predicating" function of Dem NPs.

This difference explains why in (7) Dem NP *esa película maravillosa* is more acceptable than Def NP *la película maravillosa*. In the Def NP construction, the postnominal adjective *maravillosa* is interpreted as a restrictive determiner that should help to uniquely identify the movie, but this restrictive and identifying funcion conflicts with its inherent evaluative meaning. The fact that this conflict does not arise with Dem NP shows that the construction is not expected to be identifying, but is open for introducing predicative elements, such as evaluative adjectives.

(7) Ayer vi {?la ~ esa} película maravillosa. (GRAE, 17.4c-d)

In sum, we have described the constructional meaning of Dem NP and Def NP as a combination of three features:

(a) both constructions convey a uniquely identifiable discourse referent;

(b) Dem NP includes a deictic domain in the scope of its constructional meaning domain;

(c) the NP has an identifying function in Def NP, and a predicative function in Dem NP.

2. A contrastive perspective on demonstratives and definite articles

From a contrastive point of view, it is an important challenge to verify whether the transition points between the exclusive and overlapping domains of Dem NP and Def NP are identical cross-linguistically. Methodologically, this is very complex.

The classic method of grammaticality judgments gives few results. First, they are very difficult to make, since in many contexts Dem NP and Def NP are both grammatical, although they can give rise to slightly different connotations. Moreover, when there is indeed a difference in acceptability, then closely related languages give the same results. If we would translate examples 1-7 above into Dutch, exactly the same grammaticality judgments would apply.

More results are obtained by research on comparable monolingual corpora (see particularly on French and Dutch). As is shown by Vanderbauwhede (in press a, b), this method can provide insight into the relative frequency of Dem NPs, and into the relative frequency of the different uses (e.g. direct and indirect anaphoric, cataphoric, exophoric, discourse-deictic or non-phoric uses). Yet, with this method it remains difficult to compare cross-linguistically the behavior of Def NP and Dem NP in very specific contexts: would the

other language use the same paradigm in this context or not? Therefore, it is useful to search complementary data. One of these methods is to use a corpus of translated texts and focus on translation shifts (Da Milano, 2004 for a typological study; Jonasson, 2001, 2002 for Swedish-French, Whittaker, 2004 for Norvegian-French and Vanderbauwhede, in press a for Dutch-French). When systematic trends in translation shifts are found, and these shifts occur in both translation directions, we can presume that they are due to contrastive differences between the languages (Goethals, 2007).

3. Translational shifts analysis: a quantitative approach

In this section, I will present data that come from a bidirectional corpus of translated texts (Spanish-Dutch and Dutch-Spanish[96]). I will investigate in which contexts Dem NPs are translated by Def NPs, or, vice versa, Dem NPs are used to translate Def NPs.

In the left column of table 1 the different realizations of Dem NP are listed: DIST (*die/dat*) and PROX (*deze/dit*) for Dutch, and the three forms AQUEL-, ES- and EST- for Spanish. Results are given for the two translation directions (NL→ES and ES→NL), especially the number that is translated by or is the translation of a Def NP. This means for example that in the subcorpus NL→ES, 67/404 cases of the Dutch distal Dem NP, which occur in the source text, are translated by a Def NP in the Spanish target text; 7/80 cases of *aquel-* (in the Spanish target text) are a translation of a Dutch Def NP.

Table 1. Translation shifts Dem NP – Def NP in the bidirectional corpus.

	Corpus NL → ES			Corpus ES → NL			Total		
Dem NP	total	↔ Def NP		total	↔ Def NP		total	↔ Def NP	
nl DIST	404	67	16,5%	405	77	19,0%	809	144	17,8%
nl PROX	98	10	10,2%	187	26	13,9%	285	36	12,6%
nl total	502	77	15,3%	592	103	17,4%	1094	180	16,4%
es AQUEL-	80	7	8,8%	82	5	6,1%	162	12	7,4%
es ES-	207	5	2,4%	200	9	4,5%	407	14	3,4%
es EST-	179	3	1,7%	155	8	5,2%	334	11	3,3%
es total	466	15	3,2%	466	22	5,0%	903	37	4,1%

The data reveal clear systematic tendencies. First of all, Dutch Dem NPs far more frequently correspond to a Spanish Def NP (16,4%) than is the case for Spanish Dem NPs (4,1%) (p-value 0). Importantly, the results do not depend on the translation direction: in the corpus NL-ES, 15,3% of the Dutch Dem NPs are translated by a Spanish Def NP, and, similarly, in the corpus ES-NL, 17,4% of the Dutch Dem NPs are translations of a

96 This corpus consists of Spanish and Dutch essayistic and literary texts and their translations. The corpus was developed at the Faculty of Applied Linguistics (Ghent). The total number of words is over 1,5 million, but, given the frequency of the demonstrative paradigm, for this research smaller samples were used. For a more detailed description of the corpus, see also Goethals 2007.

Spanish Def NP in the source text (this difference is not statistically significant: p 0.36). For the Spanish Dem NPs the results are also quite similar in both translation directions, with a non-significant difference (p 0.24).

This suggests that there is a contrastive difference between Dutch and Spanish: quite frequently what is expressed by a Dutch Dem NP is better expressed in Spanish by a Def NP. Following the outline of the constructional meaning of Dem NP and Def NP (section 2), this would mean that there is a difference regarding the deictic force and/or the informational value of the NP. For reasons of space, I will only focus on the factor of deictic force.

4. Qualitative analysis: deictic force in Spanish and Dutch Dem NPs and Def NPs

In this section, I will discuss several examples where the translation shift between Dutch Dem NP and Spanish Def NP can be related to the feature of deictic force. The overall hypothesis is that the Dutch paradigms have lost in a further degree their original deictic force than their Spanish counterparts. A translation shift can be necessary when a Dutch Def NP is 'not deictic enough' (while Spanish Def NP still would retain some of its original deictic force), or when a Spanish Dem NP is 'too strongly deictic' (while Dutch Dem NP would already be more bleached).

4.1. Dutch Def NP is 'not deictic enough'

In these examples, the asymmetry between Dutch Dem NP and Spanish Def NP seems due to a restriction on the use of Dutch Def NP: this construction does not seem to have the same degree of deictic force as Spanish Def NP and therefore the Spanish Def NP is translated by a Dutch Dem NP. This occurs in contexts that require a determiner with sufficient deictic force, i.e. when the (co)referential link between the NP and its referent is not self-evident. This can be the case in direct anaphoric, indirect anaphoric, discourse deictic or exophoric uses (I will give examples of direct and indirect anaphors).

A direct anaphor is illustrated in (8), where Def NP *la industria* is interpreted as a hyperonym that is coreferential with the concept *el chocolate* in the previous sentence. In Dutch, the use of Def NP would complicate this interpretation: the reader would have some difficulty in interpreting *de industrie* as a hyperonym for *the chocolate industry*, and not as the generic concept *the industry*. The use of Dem NP makes this anaphoric link self-evident.

(8) En las últimas décadas del siglo XIX se desató la glotonería de los europeos y los norteamericanos por el chocolate$_i$. El progreso de la industria$_i$ dio un gran impulso a las plantaciones de cacao en Brasil (Eduardo Galeano, *Las venas abiertas*)

In de laatste decennia van de 19de eeuw brak bij de Europeanen en de Noord-Amerikanen de chocoladevraatzucht los. De ontwikkeling van deze$_i$ {de$_{??}$} industrie was een grote stimulans voor de cacaoplantages in Brazilië.

In the corpus there are 14 examples of hyperonymic direct anaphors that show a Def-Dem asymmetry. In almost all cases (13/14), the Dem form occurs in the Dutch text. The opposite pattern is very exceptional. These data suggest that Def NP in Spanish has more deictic force than its Dutch counterpart, and activates more the anaphoric deictic domain, which favors the hyperonymic above the generic interpretation.

The next example is a different kind of non-evident direct anaphoric link. In (9), the NP is a metaphoric description of the antecedent. Apparently, in Spanish, Def NP is still sufficiently deictic to show the reader the anaphoric link with the antecedent, whereas in Dutch this seems less evident. In Dutch, the use of Dem NP makes the coreferential interpretation smoother.

(9) al inclinar el rostro, el cabello$_i$ se le había deslizado sobre la cara; tras la$_i$ cortina rubia observaba a su visitante con suspicacia (Pérez Reverte, *El Club Dumas*)

Toen ze haar hoofd naar voren boog, waren haar haren$_i$ voor haar gezicht gegraven. Vanachter dat$_i$ {het$_{?i}$} blonde gordijn nam ze haar bezoeker argwanend op.

Obviously, this type of examples is not frequent, but it is interesting to note that all asymmetric examples (three to be precise) had Dem NP in Dutch. The opposite pattern was not found.

Let us now consider indirect anaphors, which establish a coreferential relation between the NP and a prior description that exceeds the boundaries of an NP (i.e. a clause, a full sentence or even a sequence of sentences; Botley, 2006). In total we find 30 asymmetric indirect anaphors. Here again, the pattern Dutch Dem NP – Spanish Def NP (27) is far more frequent than the opposite pattern (3). For example, in (10) *el gesto* refers to the action of *levantar el vaso*. The Spanish Def NP is translated by the Dutch Dem NP *die beweging*, and, indeed, it seems counterintuitive to use Def NP (*de beweging*). The reason is that in Dutch the coreferential relation with the prior description is more easy to process with Dem NP than with Def NP.

(10) Makarova se había acercado por el otro lado de la barra [...]. La Ponte, a medio levantar el vaso$_i$, detuvo el$_i$ gesto mientras hacía una mueca instintiva de avidez profesional. (Pérez Reverte, *El Club Dumas*)

Makarova was achter de bar naar hen toe gekomen [...]. La Ponte, die zijn glas half omhoog had geheven$_i$, verstarde in die$_i$ {de$_{?i}$} beweging terwijl een instinctmatig trekje van beroepsmatige hebzucht op zijn gezicht verscheen.

4.2. Spanish Dem NP is 'too deictic'

Thus far, I have focused on examples in which Dutch Def NP seems less adequate than Spanish Def NP because of its more reduced deictic force. Now we will look at the other end of the continuum of the feature deictic force, namely examples in which the asymmetry NL Dem NP – ES Def NP seems due to the (too) strong deictic force of Spanish Dem NP. Particularly, it could explain a quite spectacular serie of translation shifts in one of the texts in the corpus. As examples (11-13) show, Dem NP *die jachtopziener* (*that game*

warden) is consistently translated by Def NP *el guardabosque*. It is important to know that the main story line is about the murder of a game warden, for which three boys have been convicted in the past. A letter has shown up which says that the boys were innocent. The three central items *the gamewarden*, *the boys* and *the letter* are constantly repeatedly throughout the story (for example, the NP *die/de jachtopziener* occurs 65 times). In the Dutch source text, in 15 references to the game warden a Dem NP is used. Interestingly, these cases are systematically translated in Spanish by Def NP[97]. The same phenomenon can be observed in the references to the three boys (see e.g. 12). I hypothesize that this is due to the lesser need to activate a deictic domain in which the referent is identified. Since the referents are the key topics of the story, their identification, and their topical status, is self-evident. In Dutch, the conflict between the Dem form and the topical status of the referents seems to be less severe than in Spanish: in Spanish, Dem NP seems to suggest that the referent is no longer topical, and needs a stronger device to be identified, or to regain its topical status. I think that we can say that Dutch Dem NP allows a more bleached form of deictic force, which would not consist in evoking a particular deictic center, but rather in evoking the overall narrative frame in which the referent is situated.

(11) 'In het dorp heeft iedereen zijn mond vol over die jachtopziener van Bidernais.' (Anke de Vries, *De Medeplichtige*)

Todos en el pueblo hablan del guardabosque de Bidernais.

(12) 'Niet zozeer over die jachtopziener zelf, maar over die jongens die zijn gegrepen.

Bueno, más que sobre el guardabosque hablan de los muchachos condenados.

(13) Je was toch met die jachtopziener verloofd?

Tú eras la novia del guardabosque?

5. Conclusion

I have tried to show that the use of a corpus of translated texts may help to describe subtle cross-linguistic differences. The quantitative analysis of translation shifts suggests that Dutch Dem NP has more evolved towards the pole of Def NP than its Spanish counterpart. In the qualitative discussion of the examples I have focused on one dimension of the constructional meaning of Dem NP and Def NP, namely their deictic force. The examples show that several translation shifts between Dem NP and Def NP can be related to this meaning dimension. The frequent asymmetric situation of Dutch Dem NP versus Spanish Def NP arises in contexts where Dutch Def NP would not have enough deictic force (unlike Spanish Def NP), and also in contexts where Spanish Dem NP would have too much deictic force (unlike Dutch Dem NP).

97 Since we work with fragments of the translated texts, not all these cases are included: 6 out of the 15 cases occur in the fragments on which the data in table 1 are based.

References

Botley, S. P. (2006). Indirect anaphora. Testing the limits of corpus-based linguistics. *International Journal of Corpus Linguistics, 11*(1), 73-112.

Da Milano, F. (2007). Demonstratives in parallel texts: a case study. *Parallel Texts: Using translational equivalents in linguistic typology. Special Issue of Sprachtypologie und Universalienforschung, 60*(2), 135-147.

Diessel, H. (1999). *Demonstratives: Form, Function, and Grammaticalization*. Amsterdam, Netherlands: Benjamins.

Eguren, L. (1999). Pronombres y adverbios demostrativos. Las relaciones deícticas. In I. Bosque & V. Demonte (Eds.), *Gramática Descriptiva de la Lengua Española* (pp. 929-972). Madrid: Espasa.

Epstein, R. (2002). The definite article, accessibility, and the construction of discourse referents. *Cognitive Linguistics, 12*(4), 333-378.

García Fajardo, J. (2006). La instrucción de contrastar en el demostrativo español. *Verba, 33*, 175-186.

Goethals, P. (2007). Corpus-driven Hypothesis Generation in Translation Studies, Contrastive Linguistics and Text Linguistics. A case study of demonstratives in Spanish and Dutch parallel texts. *Belgian Journal of Linguistics, 21*, 87-104.

Grae (2009). *Nueva gramática de la lengua española. Real Academia Española.* Madrid: Espasa.

Jonasson, K. (2001). Traduction et point de vue narratif. In O. Eriksson (Ed.), *Aspekter av litterär översättning från franska. Kollokvium vid Växjö universitet 11-12 maj 2000*. Växjö: Växjö University Press. (pp. 69-81).

Jonasson, K. (2002). Références déictiques dans un texte narratif. Comparaison entre le français et le suédois. Référence discursive dans les langues romanes et slaves. In M. Kesik (Ed.), *Actes du Colloque International de Linguistique textuelle, Lublin 24-30 septembre 2000*. Lublin: Wydawnictwo UMCS. (pp. 107-121).

Macías Villalobos, C. (2006). *El demostrativo en Miguel Delibes*. San Vicente de Raspeig: Taller Digital de Establecimiento de Textos Literarios y Científicos.

Maes, A., & Noordman, L. (1995). Demonstrative nominal anaphoras: a case of nonidentificational markedness. *Linguistics, 33*, 255-282.

Vanderbauwhede, G. and S. Verleyen (2010). The French and Dutch noun phrase in contrast: the case of the demonstrative determiner. *Linguisticae Investigationes, 33*(2), 267-284.

Vanderbauwhede, G. (in press a). The shifting of the demonstrative determiner in French and Dutch in parallel corpora: from translation mechanisms to structural differences. *Meta: Journal des Traducteurs, 56*(2).

VANDERBAUWHEDE, G. (in press b). Les emplois référentiels du déterminant démonstratif en français. Essai de systématisation. *Le Français Moderne* (2013)

WHITTAKER, S. (2004). Étude contrastive des syntagmes nominaux démonstratifs dans des textes traduits du français en norvégien et des textes sources norvégiens: stratégie de traduction ou translationese? *Forum, 2*(2), 221-240.

A corpus-based contrastive study between the English gerund and its Spanish counterparts

Mª Ángeles Gómez Castejón
University of Leuven (Belgium) / UNED

ABSTRACT

Most of the previous contrastive studies between the English gerund and its Spanish counterparts provide only a functional characterization and do not include a cognitive characterization. Furthermore, the traditional characterization of the English gerund, based on tense and aspect criteria, proves to be inadequate in establishing its meaning. This article proposes that it is important to include a cognitive analysis because this facilitates us in establishing the meaning of the English gerund as well as in establishing a hierarchy between this category and its Spanish counterparts from a conceptual point of view. In this sense, the use of a parallel corpus enables us to check in greater detail the cognitive relationship between them. As a complementary perspective, this article also provides a descriptive translation study.

RESUMEN

La mayoría de los estudios contrastivos anteriores entre el gerundio inglés y sus equivalentes españoles sólo proporcionan una caracterización funcional y no incluyen una caracterización cognitiva. Además la caracterización tradicional del gerundio inglés basada en criterios temporales y aspectuales no resulta adecuada para establecer su significado. Este artículo plantea que es importante incluir un análisis cognitivo porque este análisis nos permite establecer el significado del gerundio inglés así como establecer una jerarquía entre esta categoría y sus equivalentes españoles desde un punto de vista conceptual. En este sentido, el uso de un corpus paralelo nos permite comprobar con mayor detalle la relación cognitiva entre ellos. Como una perspectiva complementaria, este artículo ofrece además un estudio descriptivo de traducción.

Keywords: contrastive studies; English gerund; Spanish counterparts; functional characterization; cognitive characterization; corpus study; translation study

Palabras clave: estudios contrastivos; gerundio inglés; contrapartidas españolas; caracterización funcional; caracterización cognitiva; estudio a partir de un corpus, estudio de traducción

1. INTRODUCTION

A corpus-based contrastive study between the English gerund and its Spanish counterparts presents three areas of difficulty. Firstly, previous contrastive studies are limited in their analysis of these constructions; in particular, they merely provide a functional characterization. Secondly, the traditional characterization of the English gerund, in terms of tense and aspect criteria, is unable to coherently establish its meaning and nature. Finally, the use of corpora and translated texts is not fully approved by some authors.

To address our first problem, the fact that most previous studies provide only a functional characterization and not a cognitive characterization, we consider it necessary to indeed include a cognitive characterization of the English gerund and its counterparts. This description enables us to establish a hierarchy between them based on their coincidences and differences from a conceptual point of view counteracting this deficiency.

To follow, we also believe that a cognitive characterization will resolve the disparity of the English gerund analysis, both in terminology and criteria, which has previously made it difficult to establish the meaning of the English gerund. Its characterization has traditionally been established on tense and aspect criteria in terms of simultaneity/ imperfectivity respectively. This description is invalid as the English gerund not only expresses simultaneity, anteriority and posteriority but there are even cases where the temporal relation is not pertinent as in (1a). In relation to the aspectual readings (i.e. imperfective and perfective) the English gerund can express both readings as in (1b):

(1) a. Does an artist's life entail **sleeping** with anyone and everyone? (YOU 30)[98]

b. Just **thinking** about those years, Thelma having an affair with Harry almost right up to when she died, gives Janice a hollow Safe feeling. (RAB 195)

Finally, as far as the adoption of a bilingual parallel corpus and translated language are concerned, some authors still question their value for contrastive analysis: "We also have to concede that by matching texts and their translations in parallel corpora, we reduce the target language to a mirror image of the source language (Teubert, 1996: 250). Here, Teubert considers translated texts as *a mirror image of the source language* which implies that translated texts are an inadequate base for a rigorous contrastive analysis. We argue otherwise that parallel corpora are a suitable tool and that translated texts are a valid form of data when carrying out this type of analysis as explained in section 2.

After taking all this into account and in order to counteract the limitations presented above, we will provide a cognitive analysis first of the English gerund. This will constitute the base for the cognitive analysis between the English gerund and its Spanish counterparts. The whole analysis will be complemented by a translation study. Overall, both the cognitive analysis and the translation study validate the use of a parallel corpus.

98 All the examples belong to my corpus as explained in section 2. In all examples the source texts are provided in brackets after the example. A specific code has been used in which the first letters refer to the title of the source text and the numbers refer to the order within the corpus.

2. AIM AND EMPIRICAL DATA (PARALLEL CORPUS)

The main goal of this paper is to analyse and to compare the English gerund and its Spanish counterparts from a conceptual point of view and to provide translation tendencies which can help the translation of the English gerund into Spanish.

Regarding the use of a parallel corpus we believe that its use provides a suitable framework to carry out the objectives explained in the previous paragraph. The term bilingual parallel corpus refers to the corpus which includes original texts from one language (source language) with translation into another (target language). We agree with Mauranen (2002: 161) on the fact that a parallel corpus provides valid data for linguistic analysis and that translated texts should be considered authentic and contextualised data. In other words, it is a reflection of the everyday communication of a vast number of speakers:

This paper argues that corpora of translated texts constitute a valuable source of evidence for contrastive research, since they fulfil many of the criteria that have generally been seen as strengths in corpus study- for example language that has been used in its normal communicative contexts by a large number of users. (Mauranen, 2002: 161)

My corpus is an English-Spanish parallel corpus compiled specifically for my PhD. It consists of approximately 1000 pairs of English original texts and their Spanish translations. The works are described according to the usual parameters such as author, genre and the year of publication. The corpus includes 12 original texts of six authors (two per author) and their Spanish translations. All texts date from 1994 to 2002 and belong to narrative fiction:

Author	Genre	Publication	Original Text	Spanish Translation
D. Lodge	Novel	2002	*Thinks* (2001)	*Pensamientos secretos*(=PENSA)
		2001	*Therapy* (1996)	*Terapia* (= TERA).
D. Lessing	Novel	1996	*Love, again* (1996)	*De nuevo, el amor* (= AMOR)
		2006	*The sweetest dream* (2001)	*El sueño más dulce* (= SUEDUL)
N. Gordimer	Novel	1995	*None to accompany me* (1994)	*Nadie que me acompañe* (= NADIE).
		2005	*The pickup* (2001)	*El encuentro* (= ENCUEN)
J. M. Coetzee	Novel	2004	*Youth* (2002)	*Juventud* (= JUVEN).
		2000	*Disgrace* (1999)	*Desgracia* (= DESG).
J. Updike	Novel	1998	*In the beauty of the lilies* (1996)	*La belleza de los lirios* (= LILIE)
		2003	*Licks of love* (2000)	*Conejo en el recuerdo y otras historias* (= BELI).
R. Philip	Novel	1997	*Sabbath's theater* (1995)	*El teatro de Sabbath* (= TEASAB)
		2001	*The human stain* (2000)	*La mancha humana* (= MANHU)

Figure (1): English-Spanish Corpus

As illustrated in figure (2), the English corpus is the starting point of reference and this is compared with the corresponding Spanish translations:

English original texts:

D. Lodge

D. Lessing

N. Gordimer → **Spanish translations**

J. M. Coetzee

J. Updike

R. Philip

Figure (2): The directionality of the parallel corpus

While compiling my corpus, my priority was to include all the contexts in which the English gerund appeared chronologically without privileging any specific context. In doing so I aimed to get a truer reflection of the actual use of the English gerund.

3. COGNITIVE ANALYSIS: ENGLISH GERUND/SPANISH COUNTERPARTS

Before proceeding to the cognitive analysis itself, it seems necessary to present the meaning of the English gerund (3.1). Then we will analyse the Spanish counterparts in relation to the English gerund from a cognitive point of view (3.2.). Within this last section we will focus on the most frequent Spanish counterpart, the infinitive, and then we will present the third most frequent counterpart the *that-clause,* as shown in Figure (3):

Counterparts	Absolute Frequency	Relative Frequency
Infinitive	236	50%
Substantive	80	17%
That-clause	42	9%
Main verb	40	9%
Gerund	29	6%
Relative clause	22	5%
Zero	11	2%
Participle	7	1%
Version proposed by the Translator	7	1%
Total	474	100%

Figure (3): The distribution of the Spanish Counterparts

We have decided to present the infinitive and the *that-clause* because they provide contrasting perspectives as they reflect very different conceptualizations in relation to the English gerund. This justifies their being first and third most frequent counterparts respectively.

3.1. The English Gerund Characterization: integrating element

We have already explained that the characterization of the English gerund cannot be based on tense and aspect criteria, we argue that the cognitive approach enables us to establish a coherent characterization of the English gerund. Based on the cognitive approach observations, the progressive *–ing* form and the nominalization process are key elements in establishing the meaning of the English gerund.

Firstly, the *-ing* form construes a process holistically, therefore it makes the profiled relationship nonprocessual. Moreover, the *–ing* takes an internal perspective on this relationship which means that the *–ing* imposes an immediate temporal scope delimiting some internal portion of the overall relationship. Only this portion is profiled and construed as homogeneous (Langacker, 2008: 155).

Secondly, the nominalization process explains the fact that the English gerund is profiled collectively as part of an abstract entity or region:

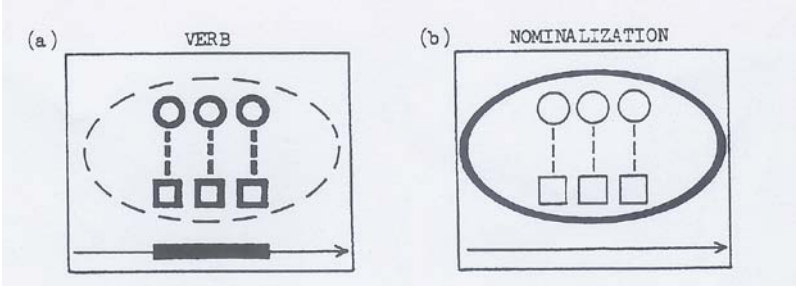


Figure (4): Verb and Nominalization (Langacker, 1991: 24)

Inherent in every verb there is an abstract region but it is only latent as depicted in (4a) with a broken line ellipse. In (4b) this latent region is completely profiled due to the nominalization process. In this paper the English gerund is described as an event seen in its entirety from a close perspective of the conceptualizer and interpreted as an abstract entity having no temporal internal structure.

3.2. Spanish Infinitive and That-clause: different degrees of conceptualization

As previously explained, we will present the infinitive and then the *that-clause* to provide contrasting conceptualizations with respect to the English gerund.

With regard to the Spanish infinitive, the high number of examples illustrating the correspondence between this counterpart and the English gerund can be explained in terms of their cognitive resemblance. The Spanish infinitive and the English gerund share a perfective aspect:

> In English, there is PERFECTIVE ASPECT when the verb form used reflects the fact that the speaker wants to refer to the actualization of a situation in its entirety, i.e; that he views the situation as if there were a temporally unstructured whole. This means that he does not refer to the situation as having an internal structure (with a beginning, middle and end). (Declerck, 2006: 99)

Having a perfective aspect implies that the Spanish infinitive and the English gerund are interpreted as an event construed as a whole without an internal temporal structure:

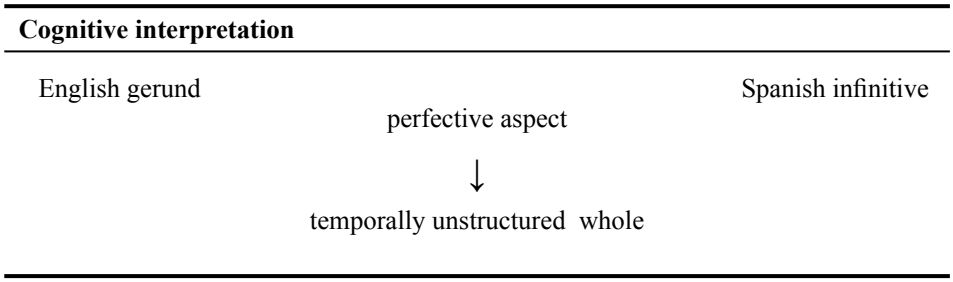


Figure (5): Cognitive interpretation: English gerund/Spanish infinitive

Moreover the Spanish infinitive and the English gerund express a subordinate relationship with the main verb. In particular the Spanish infinitive lacking time, number and person endings implies a neutral aspectuality and therefore needs to be within structures which provide this subject-verb agreement (Hernanz, 1999: 2201) favouring a subordinate relationship represented by an arrow:

(2) a. He does not like ← **drinking** (YOU 27)

b. No le gusta ← **beber** (JUVEN 41)

The conceptualization of the *that-clause* shows that although a particular translation and its original version may be equivalent, each language (i.e. target and source language) proposes structures and different images to encode the same situation. In this sense, the conceptualization of the *that-clause* differs from the English gerund in three ways: first of all, the *that-clause*, unlike the English gerund, has a sequential scanning or, in other words, the states are described individually. Secondly, the *that-clause* has its own grounding and finally it behaves like an independent clause.

Given its independent clause behaviour, it is understandable that the *that-clause* puts its own participants on a stage and therefore, the whole construction evokes two different scenes, represented by two different boxes, as seen in (3a):

(3) a. No te creas que lamento **lo que te he dicho**
b. I'm not sorry for **saying** so

The *that-clause* shows a syntactic and semantic autonomy or independence in relation to the rest of the construction; the presence of the complementizer "*that*" marks the transition between them at a syntactic level. The English gerund establishes otherwise a syntactic and semantic dependence in relation to the main verb in (3b); this is symbolized by the presence of a single box.

4. TRANSLATION STUDY: COMPLEMENTARY PERSPECTIVE

As we have presented the cognitive analysis, we will now provide a complementary study, in this instance a descriptive translation study. We will present the different contexts and factors which help to explain a particular counterpart. Some of these elements (i.e. the contexts and factors) can be better understood when taking into account the cognitive observations presented in section 3.

4.1. Spanish Infinitive

There are two main contexts for the translation of the Spanish infinitive. The first one is quite predictable within the Spanish system. In English we find the following sequence [V + E. gerund] **w**hich corresponds to [V + Spanish infinitive] in Spanish. As Hernanz (1999: 2277) points out the infinitive is the "expected" grammatical category when combined with a main verb (V) belonging to certain semantic classes (i.e. cause, emotion, communication, physical perception and cognitive perception verbs):

(4) a. I **have hated living** with you, hated every minute of it (YOU 436)
b**. He odiado vivir** contigo, cada minuto que he pasado aqui, (JUVEN 436)

Indeed, the use of the infinitive with the semantic classes as cited above shows a linear distance with the main verb. This is directly related to the concept of subordinate relationship explained in the subsection 3.2.

The second context needs to be explained. The sequence [Prepositional phrase + E. gerund] is translated into a [Verbal Periphrasis + Spanish infinitive]:

(5) a. […] or have influence with estate agents who **were wary about letting** to blacks; (NONE 664)

b. […] o tener influencia con agentes inmobiliarios que **estuviesen dispuestos a alquilar** pisos a negros; (NADIE 664)

In these types of constructions the Spanish system is defective in comparison to the English system: it imposes the presence of an infinitive after a preposition; a gerund can never be used in these constructions (*estuviesen dispuestos a **alquilando)**.

4.3. Spanish That-clause

To complete the translation study, I will cover the Spanish *that-clause*. There are three contexts in which the Spanish *that-clause* is presented. The first two are predictable from the Spanish system. Firstly when the English gerund appears with its subject, Spanish "rescues" the English gerund's subject as the subject of the *that-clause* (i.e. their/*ellos*):

(6) a. He craved **their saying it** (SAB 162)

b. Anhelaba **que (ellos) lo dijeran** (TEASAB 162)

In the second place when the English gerund is in a passive form, Spanish tends to favour active constructions and provides the *that-clause:*

(7) a. I like **being called** Tubby (THE 810)

b. Están convencidos de que me gusta **que me llamen Michelines** (TERA 810)

Finally, the third context needs to be explained as it seems quite unpredictable. When the main and complement clause share the same subject in English, the tendency is to provide an infinitive in Spanish (i.e. *Recuerdo decirle*). The use of a *that-clause* instead is explained by a deliberate separation of the two actions denoted by the main and complement verbs due to their different conceptual nature (i.e. *recuerdo*, cognitive perception nature, and *dije*, communicative nature, respectively):

(8) a. I remember **telling** her (THIN 697)

b. Recuerdo **que le dije** (NADIE 697)

5. General conclusions

The cognitive analysis has proven to be the coherent approach for our study for two significant reasons. Firstly it provides an explanation of equivalence between the English gerund and its Spanish counterparts based on their cognitive resemblance. This cognitive resemblance is directly related to the frequency of each counterpart and therefore this justifies the fact that the infinitive and the *that-clause* are the first and third most frequent counterparts respectively.

Secondly the cognitive analysis makes it possible to establish a valid and coherent characterization of the English gerund. Moreover the characterization of the *English* gerund from its nominal profile, as an abstract entity, is corroborated by the Spanish data: the most frequent counterpart is the infinitive which shares the abstract region's interpretation with the English gerund. Consequently, the translation study has proved to complement and reaffirm the cognitive analysis.

Moreover, the use of a parallel corpus has proven to be extremely useful for both our cognitive analysis and translation study.

References

Declerck, R. (2006): *The Grammar of the English Verb Phrase. Vol. I: The Grammar of the English Tense System*. Berlin / New York: Mouton de Gruyter.

Hernanz, L. (1999). El infinitivo. In I. Bosque & V. Demonte, (Eds.), *Gramática descriptiva de la lengua española* 2, 2196-2356. Madrid: Espasa Calpe.

Langacker, R. W. (1991). *Foundations of Cognitive Grammar*, Vol. 2, *descriptive applications*. Standford: Standford University Press.

Langacker, R. W. (2008). *Cognitive grammar: a basic introduction*. Oxford: Oxford University Press.

Mauranen, A. (2002). Will 'translationese' ruin a contrastive study? *Languages in contrast* 2 (2), 161- 185. Amsterdam – Philadelphia: John Benjamins.

Teubert, W. (1996). Comparable or Parallel Corpora? *International Journal of Lexicography* 9 (3), 238-264.

A contrastive structural analysis of Shakespeare's *Hamlet* versus Sumarokov's *Gamlet*: a corpus-based approach

Irina Keshabyan Ivanova
Universidad de Murcia

ABSTRACT

This article presents a new and novel investigation of the internal structural organisation of the two contrastive plays, that is, Shakespeare's Hamlet (1685) and Sumarokov's Gamlet (1787). The main aim is to compare the structures of the plays through the identification of the dimensions of structural variation linked to the cross-textual representation of the complexity of the relationships among all characters, with a particular emphasis on how the main characters Hamlet, Claudius, Polonius, Gertrude and Ophelia interact with each other as well as with all secondary characters. To this end, corpus-based techniques -in other words, computational quantification tools will be applied to the textual research. The study will also be based on the systematic qualitative analysis and comparison of the empirical data. On the whole, the key findings will show considerable distinctions between the structures of the plays per acts associated with their organisation of the social network of the characters that have mutual connections with each other.

Keywords: corpus-based techniques, quantification, qualitative, contrastive, structural organisation, social network, interact

RESUMEN

Este artículo presenta un enfoque nuevo y original del estudio de la estructura interna de las dos obras contrastivas -Hamlet (1685) de Shakespeare y Gamlet (1787) de Sumarokov. El objetivo esencial es comparar las estructuras de las dos obras a través de la identificación de las dimensiones de la variación estructural en lo referido a la representación inter-textual de la complejidad de las relaciones entre todos los personaje, con particular énfasis en las interacciones de los personajes principales, tales como Hamlet, Claudio, Polonio, Gertrudis y Ofelia, tanto entre sí como con todos los personajes secundarios. Para este fin, los métodos basados en corpus, es decir, los métodos computacionales y cuantitativos serán aplicados a la investigación de los textos. El estudio también será basado en un análisis sistemático cualitativo y comparativo de los datos empíricos. Así pues, los resultados principales van a poner de manifiesto las diferencias significativas entre las estructuras de las obras por actos en relación con la organización de la red social de los personajes que tienen conexiones mutuas entre sí.

Palabras clave: métodos basados en corpus, cuantitativo, cualitativo, contrastivo, estructura, red social, interacción

1. Introduction

Over the last two decades, the role and effectiveness of corpora and corpus-linguistic techniques has become more prominent. Cantos and Sánchez (2000: 1) explain the "tremendous growth in the compilation and use of corpora" due to "the increasing interest among linguists in studying language in use, rather than linguistic systems in the abstract", which "is primarily connected with the possibilities offered by corpora in machine-readable form, so-called computer corpora".

The main area of research of this investigation is the study of text by means of corpus-based techniques -in other words, by means of applying quantitative and analytical corpus-based methodologies to literary and textual research.

This paper provides a wide range of empirical data and in-depth qualitative and quantitative analyses and comparison of the internal structural organisation of the two contrastive plays, that is, the *Fourth Folio Edition* of *The Tragedy of Hamlet Prince of Denmark* (1685) by Shakespeare and *Gamlet* (1787) by the Russian playwright Sumarokov, translated into English by Richard Fortune in 1970. The investigation is based on the electronic collection of these texts, that is, on the computerised texts. The analysed texts are presented in Table 1.

Table 1. Texts used in the structural analysis.

Author	Title	Abbreviation
Shakespeare	*The Tragedy of Hamlet Prince of Denmark* (1685), the *Fourth Folio Edition*	SH
Sumarokov	*Gamlet* (1787), in Russian (for reference)	SG-R
	Hamlet (1970), translated into English by Richard Fortune	SG

For ease of reference, the *Fourth Folio Edition* of Shakespeare's *Hamlet* (1685) is referred to as *Hamlet* or SH. The Russian text is referred to as SG-R, whereas the English translation is referred to as *Gamlet* or SG. However, one should bear in mind that in this investigation Sumarokov's *Gamlet* in Russian and the English translation of the Russian text are used indistinctively, although the general parameters of structural variation are analysed between SH and SG and not between SH and SG-R.

In fact, the focus is on the formal aspects of the plays that could be easily located, extracted, computerized and quantified. The research question asks whether, and to what extent, the distribution patterns of the interactions of each main character, namely Hamlet, Claudius, Polonius, Gertrude and Ophelia, with all characters both main and secondary are similar or different in *Hamlet* versus *Gamlet*. The comparison is carried out per act: intra-play and inter-plays.

The main aim is to compare the structures of the plays through the identification of the dimensions of structural variation linked to the cross-textual representation of the complexity of the relationships among all characters, with a particular emphasis on how the main characters Hamlet, Claudius, Polonius, Gertrude and Ophelia interact with each other as well as with all secondary characters.

This paper is divided into four parts. Part 1 gives some general information about the texts used for the analysis, the goals, the area of research and the research question posed. Part 2 centres on the chosen variables and the quantitative tools which are applied to the analysis. Part 3 concentrates on the interpretation of the results obtained. And finally, Part 4 summarises the major findings and draws some conclusions based on the results obtained and the aims fulfilled.

2. Patterns of the interactions of the main characters and procedure of the quantitative analysis

The first task of the analysis here is to identify the salient co-occurrence structural patterns in the texts and to interpret them in empirical/quantitative terms. The interaction variables are used to reveal possible divergences in the structures of SH and SG. To this end, the total number of interaction variables is selected and quantified by examining the two text files directly. After, the extracted data are computerised, tabulated (intra-play), cross-tabulated (inter-plays) and presented in tables, graphs and schemes. Finally, these variables are compared in quantitative and qualitative terms per act: inter-plays.

The tools used for the computational quantification and presentation of the data in tables and graphs are SPSS V.15 and Excel (Office 2007). The tool applied for the design of the schemes is the computational programme Illustrator (Version *CS3*).

3. Data presentation and analysis of the distribution patterns of the interaction variables of the main characters intra-play and inter-plays

The stages of this investigation are presented in sub-sections 3.1-3.6 which focus on the distribution patterns of the interactions of the main characters Hamlet, Claudius, Polonius, Gertrude and Ophelia per acts where they appear inter-plays. With respect to Tables 3-13, it should be noted that these are extracts of the larger tables that are available online.[99] Moreover, the numbers in the last line show the total data corresponding to the latter tables. Furthermore, greater attention is paid to the data shown as a percentage as such data are considered more reliable for this kind of quantitative analysis.

99 The digitalised versions of these tables can be downloaded from: http://www.tesisenred.net/TDR-1028110-143814/index_cs.html

3.1. SH versus SG: interaction variables of Hamlet per act

Table 2 (Extract). SH versus SG: distribution patterns of the interactions of Hamlet per act I.

Hamlet with each main & secondary character	**Number of interactions**		**Differences (H vs M)**	**Each main & secondary character with Hamlet**	**Number of interactions**		**Differences (M vs H)**
SH vs SG	SH	SG	(SH-SG)	SH vs SG	SH	SG	(SH-SG)
Hamlet	2	1	1	Claudius	2	-	-
Claudius	2	-	-	Gertrude	3	13	-10
Gertrude	3	9	-6	Marcellus	8		
Marcellus	1			Horatio	41		
Horatio	26			Ghost	9		
Ghost	11			Both (Bar-Mar)	3		
Horatio-Marcellus	13			All (Bar-Mar-Hor)	1		
Armans		6					
Total	71	18	-5	**Total**	69	21	-10
%				%			
Hamlet	2.82	5.56	-2.74	Gertrude	4.35	61.90	-57.55
Gertrude	4.23	50.00	-45.77				
Total	7.05	55.56	-48.51	**Total**	4.35	61.90	-57.55

Table 3. SH versus SG: distribution patterns of the interactions of Hamlet per act II.

Hamlet with each main & secondary character	**Number of interactions**		**Differences (H vs M)**	**Each Main & secondary character with Hamlet**	**Number of interactions**		**Differences (M vs H)**
SH vs SG	SH	SG	(SH-SG)	SH vs SG	SH	SG	(SH-SG)
Hamlet	2	1	1	Claudius	4	-	-
Claudius	2	-	-	Polonius	8	0	8
Polonius	5	0	5	Gertrude	24	-	-
Gertrude	27	-	-	Ophelia	27	9	18
Ophelia	30	8	22	Rosincros	6		
Rosincros	4			Guildenstare	13		
Guildenstare	13			Players	2		
Players	2			Horatio	7		
Horatio	8			Both (Rosin-Guild)	1		
Total	104	12	28	**Total**	93	10	26
%				%			
Hamlet	1.92	8.33	-6.41	Polonius	8.60	0.00	8.60
Polonius	1.92	0.00	1.92	Ophelia	29.03	90.00	-60.97
Ophelia	4.81	66.67	-61.86				
Total	8.65	75.00	-66.35	**Total**	37.63	90.00	-52.37

Table 4. SH versus SG: distribution patterns of the interactions of Hamlet per act V.

Hamlet with each main & secondary character	**Number of interactions**		**Differences (H vs M)**	**Each main & secondary character with Hamlet**	**Number of interactions**		**Differences (M vs H)**
SH vs SG	SH	SG	(SH-SG)	SH vs SG	SH	SG	(SH-SG)
Claudius	4	0	4	Claudius	3	0	3
Polonius	-	1	-	Polonius	-	1	-
Gertrude	2	-	-	Gertrude	3	-	-
Ophelia	-	9	-	Ophelia	-	10	-
Clown	17			Clown	18		
Horatio	33			Horatio	27		
Laertes	15			Laertes	12		
Osrick	12			Osrick	14		
Total	87	12	4	**Total**	78	13	3
%				%			
Claudius	4.60	0.00	4.60	Claudius	3.85	0.00	3.85
Total	4.60	0.00	4.60	**Total**	3.85	0.00	3.85

The data in Table 2 show that Shakespeare appears to ascribe much more importance to Hamlet's socialisation with the secondary characters that belong to a lower social rank compared to Sumarokov who is especially drawn to family relationships between the mother and the son.

The data in Table 3 display that in Act III Shakespeare pays greater attention to socio-political relationships of Hamlet with Polonius. However, Sumarokov's Hamlet does not socialise with Polonius which may emphasise that this relationship is of no importance to Sumarokov. The link is particularly asymmetrical between Hamlet and Ophelia which probably highlights that the connection between them is closer in SG than in SH.

The data in Table 4 display that in Act V Shakespeare's interest is in family and political relationships between Hamlet (stepson-prince) and Claudius (father-king) as well as in the link between Hamlet and the people of a lower social status, namely Clown, Horatio, Laertes and Osrick. In contrast to Shakespeare, Sumarokov pays increasing attention to the personal relationship between Hamlet and Ophelia and to a lesser extent to political connections of Hamlet.

3.2. SH versus SG: interaction variables of Claudius per act

The data in Table 5 show that in Act II Shakespeare pays attention to the interactions of Claudius with the main and secondary characters compared to Sumarokov who centres only on the interactions of Claudius with the main characters such as Polonius and

Gertrude. The latter point seemingly points to the fact that Claudius is completely isolated from the secondary characters and lacks political importance in SG. At the same time, Shakespeare is possibly drawn to the political importance of the king Claudius.

Table 5. SH versus SG: distribution patterns of the interactions of Claudius per act II.

Claudius with each Main & secondary character	**Number of interactions**		**Differences (C vs M)**	**Each main & secondary character with Claudius**	**Number of interactions**		**Differences (M vs C)**
SH vs SG	SH	SG	(SH-SG)	SH vs SG	SH	SG	(SH-SG)
Polonius	8	6	2	Polonius	6	5	1
Gertrude	1	1	0	Gertrude	2	1	1
Rosin-Guild	1			Voltimand	1		
Total	13	7	2	**Total**	9	6	2
%				%			
Polonius	61.54	85.71	-24.17	Polonius	66.67	83.33	-16.66
Gertrude	7.69	14.29	-6.60	Gertrude	22.22	16.67	5.55
Total	69.23	100.00	-30.77	**Total**	88.89	100.00	-11.11

The data in Tables 6 and 7 display that in Acts III and V Shakespeare's Claudius socialises with both main and secondary characters whilst Sumarokov's Claudius only socialises with the main characters, namely Polonius and Ophelia.

Table 6. SH versus SG: distribution patterns of the interactions of Claudius per act IV.

Claudius with each main & secondary character	**Number of interactions**		**Differences (C vs M)**	**Each main & secondary character with Claudius**	**Number of interactions**		**Differences (M vs C)**
SH vs SG	SH	SG	(SH-SG)	SH vs SG	SH	SG	(SH-SG)
Hamlet	8	-	-	Hamlet	9	-	-
Claudius	1	0	1	Polonius	-	3	-
Polonius	-	2	-	Gertrude	6	-	-
Gertrude	8	-	-	Ophelia	3	1	2
Ophelia	3	1	2	Rosincros	3		
Rosincros	2			Laertes	20		
Laertes	20			Messenger	3		
Total	49	3	3	**Total**	44	4	2
%				%			
Claudius	2.04	0.00	2.04	Ophelia	6.82	25.00	-18.18
Ophelia	6.12	33.33	-27.21				
Total	8.16	33.33	-25.17	**Total**	6.82	25.00	-18.18

Table 7. SH versus SG: distribution patterns of the interactions of Claudius per act V.

Claudius with each main & secondary character	**Number of interactions**		**Differences (C vs M)**	**Each main & secondary character with Claudius**	**Number of interactions**		**Differences (M vs C)**
SH vs SG	SH	SG	(SH-SG)	SH vs SG	SH	SG	(SH-SG)
Hamlet	3	0	3	Hamlet	4	0	4
Claudius	1	0	1	Polonius	-	3	-
Polonius	-	1	-	Gertrude	1	-	-
Gertrude	2	-	-	Laertes	2		
Laertes	3			Osrick	1		
Osrick	1			Soldier		1	
Total	16	1	4	**Total**	8	4	4
%				%			
Hamlet	18.75	0.00	18.75	Hamlet	50.00	0.00	50.00
Claudius	6.25	0.00	6.25				
Total	25.00	0.00	25.00	**Total**	50.00	0.00	50.00

3.3. SH versus SG: interaction variables of Polonius per act

Table 8. SH versus SG: distribution patterns of the interactions of Polonius per act II.

Polonius with each main & secondary character	**Number of interactions**		**Differences (P vs M)**	**Each main & secondary character with Polonius**	**Number of interactions**		**Differences (M vs P)**
SH vs SG	SH	SG	(SH-SG)	SH vs SG	SH	SG	(SH-SG)
Hamlet	19	-	-	Hamlet	20	-	-
Claudius	6	5	1	Claudius	8	6	2
Polonius	1	0	1	Gertrude	3	0	3
Gertrude	0	1	-1	Ophelia	5	-	-
Ophelia	5	-	-	Reynoldo	13		
Reynoldo	13						
Players	3						
Total	59	6	9	**Total**	49	6	5
%				%			
Claudius	10.17	83.33	-73.16	Claudius	16.33	100.00	-83.67
Polonius	1.69	0.00	1.69	Gertrude	6.12	0.00	6.12
Gertrude	0.00	16.67	-16.67				
Claudius-Gertrude	10.17	0.00	10.17				
Total	22.03	100.00	-77.97	**Total**	22.45	100.00	-77.55

Table 9. SH versus SG: distribution patterns of the interactions of Polonius per act II.

Polonius with mach main & secondary character	**Number of interactions**		**Differences (P vs M)**	**Each main & secondary character with Polonius**	**Number of interactions**		**Differences (M vs P)**
SH vs SG	SH	SG	(SH-SG)	SH vs SG	SH	SG	(SH-SG)
Hamlet	8	0	8	Hamlet	5	0	5
Claudius	3	-	-	Claudius	4	-	-
Gertrude	3	-	-	Gertrude	1	-	-
Ophelia	0	13	-13	Ophelia	0	12	-12
Players	1						
Total	18	13	-5	**Total**	10	12	-7
Hamlet	44.44	0.00	44.44	Hamlet	50.00	0.00	50.00
Ophelia	0.00	100.00	-100.00	Ophelia	0.00	100.00	-100.00
Total	44.44	100.00	-55.56	**Total**	50.00	100.00	-50.00

The data in Tables 8 and 9 show that in Acts II and III Shakespeare's focus is on political and family relationships of Polonius who interacts with both main and secondary characters such as Hamlet, Claudius, Ophelia on the one hand, and Reynoldo, players, etc. on the other. By contrast, Sumarokov's only interest lies in political connections of the same character with Claudius and Gertrude.

3.4. SH versus SG: interaction variables of Gertrude per act

The data in Table 10 display that in Act I both Shakespeare and Sumarokov are drawn by the family relationship between Gertrude and Hamlet, although with preference to Shakespeare. At the same time, Sumarokov makes Gertrude socialise with Armans, that is, with a person of a lower social position.

Table 10. SH versus SG: distribution patterns of the interactions of Gertrude per act I.

Gertrude with each main & secondary character	**Number of interactions**		**Differences (G vs. M)**	**Each main & secondary character with Gertrude**	**Number of interactions**		**Differences (M vs. G)**
SH vs SG	SH	SG	(SH-SG)	SH vs SG	SH	SG	(SH-SG)
Hamlet	3	13	-10	Hamlet	3	9	-6
Hamlet-Armans		1		Armans		4	
Total	3	14	-10	**Total**	3	13	-6
%				%			
Hamlet	100.00	92.86	7.14	Hamlet	100.00	69.23	30.77
Total	100.00	92.86	7.14	**Total**	100.00	69.23	30.77

Table 11. SH versus SG: distribution patterns of the interactions of Gertrude per act II.

Gertrude with each main & secondary character	**Number of interactions**		**Differences (G vs. M)**	**Each main & secondary character with Gertrude**	**Number of interactions**		**Differences (M vs. G)**
SH vs SG	SH	SG	(SH-SG)	SH vs SG	SH	SG	(SH-SG)
Claudius	2	1	1	Claudius	1	1	0
Polonius	3	0	3	Polonius	0	1	-1
Gertrude	0	1	-1	Ratuda		5	
Guildenstare	1						
Rosin-Guild	2						
Ratuda		5					
Total	9	9	3	**Total**	1	7	-1
%				%			
Claudius	22.22	11.11	11.11	Claudius	100.00	14.29	85.71
Polonius	33.33	0.00	33.33	Polonius	0.00	14.29	-14.29
Gertrude	0.00	11.11	-11.11				
Claudius-Polonius	11.11	11.11	0.00				
Total	66.66	33.33	33.33	**Total**	100.00	28.58	71.42

The data in Table 11 show that in Act II Shakespeare is interested in Gertrude's relationships with the main characters such as Claudius and Polonius and secondary characters, namely Guildenstare and Rosincros, on a more or less the same level. As opposed to Shakespeare, Sumarokov is mostly drawn by Gertrude's relationship with the people of a lower social rank, that is, her confidante Ratuda.

3.5. SH versus SG: interaction variables of Ophelia per act

Table 12. SH versus SG: distribution patterns of the interactions of Ophelia per act III.

Ophelia with each main & secondary character	**Number of interactions**		**Differences (Oph vs. M)**	**Each main & secondary character with Ophelia**	**Number of interactions**		**Differences (M vs. Oph)**
SH vs SG	SH	SG	(SH-SG)	SH vs SG	SH	SG	(SH-SG)
Hamlet	27	9	18	Hamlet	30	8	22
Polonius	0	12	-12	Polonius	0	13	-13
Gertrude	1	-	-				
Ophelia	1	1	0				
Total	29	22	6	**Total**	30	21	9
%				%			
Hamlet	93.10	40.91	52.19	Hamlet	100.00	38.10	61.90
Polonius	0.00	54.55	-54.55	Polonius	0.00	61.90	-61.90
Ophelia	3.45	4.55	-1.10				
Total	96.55	100.00	-3.45	**Total**	100.00	100.00	0.00

Table 13. SH versus SG: distribution patterns of the interactions of Ophelia per act IV.

Ophelia with Each Main & Other Character	**Number of Interactions**		**Differences (Oph vs. M)**	**Each Main & Other Character with Ophelia**	**Number of Interactions**		**Differences (M vs. Oph)**
SH vs SG	SH	SG	(SH-SG)	SH vs SG	SH	SG	(SH-SG)
Claudius	3	1	2	Claudius	3	1	2
Polonius	-	4	-	Polonius	-	4	-
Gertrude	4	-	-	Gertrude	3	-	-
Ophelia	0	2	-2	Laertes	4		
Gertrude-Claudius	2	-	-	Flemina		2	
Clau-Gert-Laertes	5			Captain of the Guard		1	
Flemina		3					
Total	14	11	0	**Total**	10	8	2
%				%			
Claudius	21.43	9.09	12.34	Claudius	30.00	12.50	17.50
Ophelia	0.00	18.18	-18.18				
Total	21.43	27.27	-5.84	**Total**	30.00	12.50	17.50

The data in Table 12 display that in Act III Shakespeare's Ophelia primarily interacts with Hamlet compared to Sumarokov's Ophelia who mostly socialises with her father Polonius and to a lesser degree with Hamlet.

The data in Table 13 show that in Act IV Shakespeare is keen on the relationships of Ophelia with the main characters, namely Gertrude and Claudius, whilst Sumarokov places more importance on the interactions of Ophelia with her father Polonius.

Consequently, the previously examined and discussed data probably provide evidence of considerable differences in the distribution patterns of the interactions of the main characters Hamlet, Claudius, Polonius, Gertrude and Ophelia per acts where they coincide inter-plays.

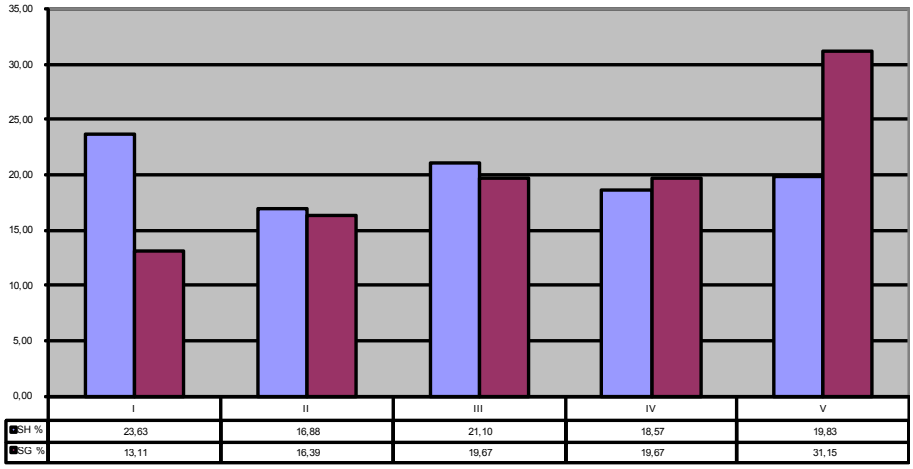
3.6. Summary of the distribution patterns of the lines of interaction among all characters per act: inter-plays

This stage of the analysis focuses on the results related to the distribution patterns of the lines of interaction among all characters per act: inter-plays. What I have done is count the lines of interaction among all characters and present them in schemes[100]. After, I have quantified these data as a percentage and displayed these in a graph.

100 For further information, see Keshabyan-Ivanova (2010).

To sum up, the data in Graph 1 show:

1. Significant quantitative dissimilarities with regard to the complexity of the relationships among all characters per Acts I and V: inter-plays. In these acts, all characters interact with each other more frequently per Act I in SH and per Act V in SG -in other words, the complexity of the relationships, that is, the interaction pattern among all characters increases progressively in SG whilst it fluctuates slightly from act to act and decreases towards the end of the play in SH.
2. Considerable quantitative commonalities in Shakespeare and Sumarokov's perspectives of the complexity of the social network among all characters per Acts II, III and IV: inter-plays.



Graph 1. SH versus SG: summary of the distribution patterns of the lines of interaction among all characters per acts I-V.

4. Conclusions

Whereas potentially SH and SG are similar plays, yet a wide range of qualitative variation and extensive quantitative distinctions are revealed with respect to the complexity of the relationships, that is, the interaction patterns among all characters, particularly among the main characters, namely Hamlet, Claudius, Polonius, Gertrude and Ophelia, with each other and with all secondary characters.

In fact, Shakespeare mostly links his main characters to the socio-political aspect of life, that is, he focuses on their interactions with both main and secondary characters to solve political problems. On the contrary, Sumarokov's main goal is to show family and personal contacts among humans, especially of a high social ranking, and to some extent deal with socio-political relationships of these characters within society. Thus, Sumarokov's perception of the main characters carries mostly family and personal associations and to a lesser degree political connotations.

To conclude, the major findings on the distribution patterns of the interactions presented in this paper shed light on considerable divergences in Shakespeare and Sumarokov's perceptions of the main and secondary characters. In so doing, theses data expose that these perceptions have taken Sumarokov to distort the structural organisation of social connections in Shakespeare's original play *Hamlet*.

References

Cantos, P., & Sánchez, A. (Eds.). (2000). An Introduction. *Cuadernos de Filolodía Ingesa*, (9)1, 1-3.

Keshabyan-Ivanova, I. (2010). *A Contrastive Structural and Lexical Study of Shakespeare's Hamlet and Sumarokov's Gamlet: A Corpus-Based Approach to Literature.* Retrieved from http://www.tesisenred.net/TDR-1028110-143814/index_cs.html

Sumarokov, A. (1970). Hamlet. In H., Jr., Nebel (Ed.), *Selected Tragedies of A. P. Sumarokov* (pp. 87-134). Evanston: Northwestern University Press (Original work published 1787).

Sumarokov, A. (1787). Gamlet. Tragediia. In N. Novikov (Ed.), *Complete collection of all the works in poetry and prose in 10 volumes of Aleksandr Petrovich Sumarokov* (Vol. 3, pp. 61-134). Moscow: Universitetskaia Tipografiia.

Shakespeare, W. (1685). *The Tragedy of Hamlet Prince of Denmark* (pp. 343-357). Retrieved from: http://adrastea.ugr.es/search~S1*spi?/.b1438681/.b1438681/1,1,1,B/l856~b1438681&FF=&1,0,,1,0

Hacia un enfoque empírico en la semántica a través de la traducción: estudio contrastivo del verbo *sentir*.

Jansegers Marlies & Enghels Renata
Universidad de Gante

Resumen. *Últimamente, la lingüística de corpus se ha revelado útil tanto para el estudio de aspectos morfosintácticos de la lengua, como para estudios de índole semántica. El presente estudio examina en qué medida dos tipos de corpus – uno paralelo y otro comparable – pueden ser complementarios para el estudio semántico de los cuasi-sinónimos entre lenguas. Investiga cómo tanto el método introspectivo como la combinación de varios tipos de datos empíricos pueden llevar a un refinamiento gradual de las hipótesis previas. En este proceso gradual, se insiste más particularmente en el papel fundamental de la traducción como eslabón entre el método introspectivo por un lado y el corpus comparable por otro lado. El estudio de caso concreto se centra en el verbo* sentir(e) *en español francés e italiano.*

Palabras claves: *semántica, polisemia, corpus paralelo, corpus comparable, verbos de percepción, sentir.*

Abstract. *In the last few decades, corpora have been proven useful for the study of both morphosyntactic and semantic aspects of language. This study examines to what extent two types of corpus – a parallel and a comparable corpus – can be complementary for the semantic study of near-synonyms between languages. We will investigate how the introspective method as well as the combination of several types of empirical data can lead up to a gradual refinement of previous hypotheses. In this gradual process, the fundamental role of translation will be emphasized as a link between the introspective method on the one hand and the comparable corpus on the other hand. The case study focuses on the verb* sentir(e) *in French, Spanish and Italian.*

Keywords: *semantics, polysemy, parallel corpus, comparable corpus, perception verbs, sentir*.

1. Introducción: hacia un enfoque empírico en la semántica

En las últimas décadas, la lingüística ha experimentado una transición desde los enfoques más tradicionales basados básicamente en la intuición hacia un enfoque más cuantitativo basado en métodos empíricos. El uso de corpus se ha revelado particularmente útil para el análisis de fenómenos morfosintácticos de la lengua, pero también dentro del campo de la semántica, han surgido voces pidiendo aproximaciones más objetivas del objeto de estudio. De hecho, contrariamente a la idea de que el mejor método para abordar la semántica es la intuición (Talmy, 2007), varios autores (cf. entre otros Glynn, 2010; Oster, 2010) sostienen que la introspección no es suficiente sino que representa meramente la primera etapa en lo que Geeraerts (2010) llama *el ciclo empírico* (*the empirical cycle*). Sin embargo, el uso de corpus para el estudio semántico conlleva ciertas dificultades metodológicas (cf. las distintas maneras para recoger los datos, la amplia gama de técnicas cuantitativas para el tratamiento de los resultados, etc. véase Glynn, 2010).

El presente estudio dedica especial atención a la pregunta en qué medida dos tipos distintos de corpus – uno paralelo y otro comparable – pueden ser complementarios para el estudio semántico de los llamados *cuasi-sinónimos* entre lenguas.

Intentaremos contestar a esta pregunta mediante el estudio concreto del verbo de percepción español *sentir* y sus equivalentes morfológicos en francés y en italiano. Como estos verbos derivan todos del mismo étimo latín (*sentire*) partimos de la hipótesis de que también desde el punto de vista semántico son cognados perfectos. Un análisis del tratamiento lexicográfico del verbo (sección 2) así como los resultados de ambos estudios de corpus (secciones 3 y 4) nos permitirán someter a prueba esta hipótesis.

2. Estudio lexicográfico

A fin de formarse una primera idea de la semántica de *sentir* en estas tres lenguas, conviene comparar el tratamiento lexicográfico del verbo[101]. A simple vista, la comparación de los diccionarios ofrece una imagen poco clara de los distintos significados del verbo.

Sin embargo, como se desprende de la tabla 1, si focalizamos por ejemplo en los verbos utilizados para parafrasear el significado de *sentir* – verbos de percepción, de cognición o de emoción – observamos algunas semejanzas entre las tres lenguas:

Tabla 1. Estudio lexicográfico: semejanzas

Tipo de verbo	DUE	PR	GDIU
Verbos de percepción	*percibir, experimentar*	*percevoir*	*percipere*
Verbos de cognición	*darse cuenta de, ser consciente de, creer, juzgar, opinar, sospechar, presentir, barruntar*	*avoir/prendre conscience, se rendre compte de, deviner, discerner, pressentir*	*avvertire, giudicare, informarsi, presentire*
Verbos de emoción	*ser afectado, ser capaz de impresionarse o emocionarse*	*éprouver, ressentir, être affecté, apprécier*	*provare, ammirare, apprezzare*

101 Por motivos prácticos se presentan aquí los resultados de un diccionario relevante por lengua estudiada, o sea el *Diccionario de Uso del Español* (DUE), *le Petit Robert* (PR) y el *Grande Dizionario Italiano dell'Uso* (GDIU); véase la bibliografía.

Por otro lado, también surgen una serie de divergencias. Así, el DUE menciona el significado específico de *lamentar* y la expresión fija *lo siento*, contrariamente a los diccionarios francés e italiano. Por su parte, el PR menciona explícitamente que el verbo *sentir* no se utiliza para la percepción auditiva, mientras que este significado parece presentarse en español y resulta muy elaborado en italiano. Finalmente, en comparación con las dos otras lenguas, la entrada del verbo en el GDIU es mucho más extensa y menciona algunos significados ausentes en el DUE y el PR como *sentire un medico* ('consultar a un médico') o el uso como interjección *senti*.

Este análisis lexicográfico revela pues que, aunque los verbos en las tres lenguas tengan algunos significados básicos en común, también deben existir diferencias semánticas entre ellos, lo cual nos incita a rechazar nuestra hipótesis inicial y establecer la alternativa de que los verbos *sentir(e)* en las tres lenguas son cuasi sinónimos pero no cognados perfectos.

Sin embargo, este tipo de análisis lexicográfico más bien introspectivo presenta algunas desventajas. Así, no permite medir en qué medida las lenguas difieren precisamente, es decir, si hay algunas que se aproximan más que otras en determinados dominios. Además, los diccionarios no facilitan la detección de los núcleos semánticos en cada lengua individual. Son precisamente este tipo de limitaciones que estimulan el uso de otros métodos que permiten dar cuenta del uso *real* de la lengua.

3. Equivalencia de sentir en un corpus paralelo

3.1 El uso de la traducción en la investigación lingüística

Es bien sabido que, mientras que el uso de corpus comparables ha sido generalmente aceptado en círculos lingüísticos, el uso de traducciones queda sujeto a discusión. En efecto, aunque varios autores han insistido en las distintas posibilidades relacionadas al uso de corpus paralelos (por ejemplo Johansson, 1999; Noël, 2003), el problema más frecuentemente citado es el llamado *translationese* (McEnery & Xiao, 2008: 22-23), ligado a la pregunta: "¿cómo se puede estar seguro de que durante la transposición del texto, el traductor no era consciente o inconscientemente influido por la lengua fuente?" (Van Hoecke & Goyens, 1990: 124). Más recientemente, varios lingüistas (como Gilquin, 2008 y Vanderschueren, 2010) han abogado por la combinación de un corpus paralelo y un corpus comparable. El estudio actual se sitúa en esta corriente, pero va más allá, insistiendo en el papel fundamental de la traducción como eslabón entre el método puramente introspectivo por un lado y el análisis de un corpus comparable por otro lado.

Los datos del corpus paralelo provienen de dos textos germánicos, a saber *Harry Potter and the Sorcerer's Stone* (J.K. Rowling) y *Män som hatar kvinnor* (S. Larsson) y sus respectivas traducciones en las tres lenguas romances [102]. La elección de los textos fuente se explica por el intento de limitar al máximo el riesgo de la *translationese*, que es particularmente frecuente entre lenguas 'hermanas' (Vanderschueren, 2010: 95). Además,

102 Para más información, véase la bibliografía. El corpus entero incluye aproximadamente 800 080 palabras: 268 800 en las traducciones españolas, 273 680 en la parte francesa y 257 600 en las traducciones italianas.

en vez de comparar el texto fuente y sus respectivas traducciones, hemos comparado las tres traducciones entre sí. Con el objetivo de examinar cómo un mismo contexto semántico en el texto fuente se traduce en las tres lenguas romances, se han extraído todas las ocurrencias del verbo *sentir* en cada lengua separada para buscar después los contextos correspondientes en las dos otras lenguas.

3.2 Resultados

El número diferente de ocurrencias de *sentir* en las tres traducciones sugiere que el verbo conoce un uso mucho más amplio en italiano (236 casos) en comparación con el español (127 casos) y el francés (82 casos). Además, en el corpus entero se encuentran solamente 17 ejemplos de correspondencia exacta entre las tres lenguas, que denotan casi todos una percepción física general (1):

(1a) Harry *sintió* como si se le helaran las entrañas. (HPESP: 211)

(1b) Harry *sentit* son sang se glacer. (HPFR: 141)

(1c) Harry *sentì* le budella congelarglisi dentro la pancia. (HPIT: 124)

En los demás ejemplos los verbos no se presentan como cognados perfectos. Y es más, el análisis detenido de los datos paralelos permite averiguar que las equivalencias del verbo *sentir* en las tres lenguas podrían dividirse básicamente en seis categorías – ilustradas más detalladamente abajo –, o sea: (1) correspondencia exacta, (2) traducción mediante un verbo derivado (como *ressentir*), (3) mediante un verbo de percepción, (4) un verbo de cognición, (5) una expresión emotiva o (6) correspondencia cero.

Empezando por el *sentir* español, el gráfico siguiente ofrece un resumen cuantitativo de las correspondencias en francés y en italiano:

Tabla 2. *Sentir* español y equivalentes

equivalentes	francés		italiano	
	#	%	#	%
sentir/sentire	41	32,3%	32	25,2%
ressentir	19	15%	-	-
verbo de percepción	2	1,6%	25	19,7%
verbo de cognición	4	3,1%	20	15,7%
expresión emotiva	29	22,8%	29	22,8%
sin correspondencia	32	25,2%	21	16,5%
total	127	100%	127	100%

De la tabla resulta que el número de casos correspondientes con el francés es algo más elevado que con el italiano, y sobre todo si se tiene en cuenta también las traducciones mediante el verbo derivado *ressentir*. En francés hay cierta competencia entre *sentir* y *ressentir* cuando el estímulo refiere a un estado de ánimo más abstracto, como *le bonheur, la crainte* etc. (2):

(2a) Había comenzado a *sentir* una punzada de miedo cada vez que mencionaban a Quien-tú-sabes. (HPESP: 108)

(2b) Il commençait à *ressentir* un frisson de crainte chaque fois qu'on lui parlait de Vous-Savez-Qui. (HPFR: 73)

La comparación de las traducciones mediante un verbo cognitivo también llama la atención, ya que contrariamente al francés, el número de traducciones mediante otro verbo cognitivo, como *pensare, capire,* en italiano es bastante alto. Esta discrepancia sugiere que en italiano el verbo *sentire* mismo acepta más difícilmente el significado de un verbo de cognición (3):

(3a) Comenzaba a *sentir* que nada podía sorprenderlo. (HPESP: 103)

(3b) Cominciava a *pensare* che tutto fosse possibile. (HPIT: 60)

Finalmente, en un número considerable de ocurrencias, el *sentir* español transmite cierto sentido emotivo de arrepentimiento (4a), imposible para el *sentir* francés e italiano; estas lenguas deben recurrir a otras expresiones como *désolé* o *scusatemi* (4b-c):

(4a) ¡Vas a despertar a los muggles! – Lo... *siento* –lloriqueó Hagrid, y se limpió la cara con un gran pañuelo–. (HPESP: 18)

(4b) Vous allez réveiller les Moldus! – *Dé... désolé*, sanglota Hagrid en sortant de sa poche un grand mouchoir […]. (HPFR: 10)

(4c) 'Sveglierai i Babbani!' '*S-s-s-scusatemi*...' singhiozzò Hagrid tirando fuori un immenso fazzoletto [...]. (HPIT: 10)

La comparación a partir del *sentir* francés arroja una imagen bastante similar, y corrobora que el grado de correspondencia entre el *sentir* francés y español es más alto que con el italiano:

Tabla 3. *Sentir* francés y equivalentes

equivalentes	español		italiano	
	#	%	#	%
sentir/sentire	37	45,1%	32	39,1%
verbo de percepción	12	14,6%	8	9,8%
verbo de cognición	4	4,9%	17	20,7%
expresión emotiva	-	-	-	-
sin correspondencia	29	35,4%	25	30,5%
total	82	100%	82	100%

Además, el hecho de que el traductor español recurra a menudo a otros verbos de percepción – especialmente *oler* – podría indicar que en esta lengua el verbo *sentir* se presta menos a la expresión de la modalidad olfativa (5b). El italiano también utiliza otros verbos en este campo, aunque en menor medida (5c):

(5a) Toute la maison *sentait* le chou et Mrs Figg passait son temps à lui montrer les photos de tous les chats qu'elle avait eus. (HPFR: 14)

(5b) Toda la casa *olía* a repollo y la señora Figg le hacía mirar las fotos de todos los gatos que había tenido. (HPESP: 24)

(5c) Harry detestava quella casa. *Puzzava* di cavolo e Mrs Figg lo costringeva a guardare le fotografie di tutti i gatti che aveva posseduto in vita sua. (HPIT: 14)

Finalmente, el análisis de correspondencia mutua que toma el *sentire* italiano como punto de partida arroja resultados un tanto diferentes:

Tabla 4. *Sentir* italiano y equivalentes

equivalentes	español		francés	
	#	%	#	%
sentir/sentir	29	12,3%	30	12,7%
verbo de percepción	131	55,5%	133	56,3%
verbo de cognición	12	5,1%	8	3,4%
expresión emotiva	-	-	-	-
sin correspondencia	64	27,1%	65	27,5%
total	236	100%	236	100%

En primer lugar, observamos que el número de correspondencias perfectas en español y en francés es bastante bajo (un 12,3% y un 12,7%) pero muy comparable en ambas lenguas. En segundo lugar, llama la atención que el número de traducciones por otros verbos de percepción es muy alto. Un estudio más detenido indica que se trata sobre todo de verbos auditivos como *oír/escuchar* en español (6b) y *entendre/écouter* en francés (6c), lo que sugiere al mismo tiempo un fuerte desarrollo del significado auditivo en italiano:

(6a) 'Ho *sentito* dire della sua famiglia' disse Ron cupo. (HPIT: 64)

(6b) '*Oí* hablar sobre su familia', dijo Ron en tono lúgubre. (HPESP: 110)

(6c) 'J'ai *entendu* parler de sa famille', dit Ron d'un air sombre. (HPFR: 75)

En suma, el análisis del corpus paralelo nos ha llevado a algunas conclusiones interesantes, que complementan y refinan los resultados del análisis lexicográfico: (1) permite medir y expresar cuantitativamente el grado de correspondencia entre los cuasi-sinónimos entre las tres lenguas, y (2) la naturaleza semántica de los verbos correspondientes en las otras lenguas permite verificar y designar con más precisión los núcleos semánticos inherentes al verbo. Además, como el número de concordancias perfectas es limitado, este estudio del corpus paralelo corrobora nuestra hipótesis resultante del análisis lexicográfico según la cual los verbos *sentir(e)* en las tres lenguas no son cognados perfectos.

Sin embargo, al mismo tiempo, este estudio de corpus paralelo no es del todo conclusivo e induce a su vez a otra pregunta esencial, o sea: *¿en qué medida estos núcleos se verifican en textos originales en las tres lenguas?*

4. Equivalencia de sentir en un corpus comparable

A fin de formarse una idea aún más precisa de la semántica de los verbos *sentir(e)* hemos compuesto un corpus de 1500 ocurrencias – 500 por lengua – que debe de ser representativo del uso del verbo en español, francés e italiano escrito actual. La mitad de los ejemplos proviene de obras literarias, la otra mitad de textos periodísticos[103]. La tabla siguiente visualiza las frecuencias de los significados del verbo en cada lengua:

Tabla 5. Sentir en los corpus comparables

	ESP		FR		IT	
	#	%	#	%	#	%
percepción fisica general	55	11%	88	17,6%	59	11,8%
percepción emotiva	313	62,6%	25	5%	47	9,4%
percepción cognitiva	58	11,6%	199	39,8%	30	6%
percepción específica	40	8%	58	11,6%	261	52,2%
uso interjección	-	-	-	-	50	10%
categoría X	34	6,8%	130	26%	53	10,6%
total	500	100%	500	100%	500	100%

Como se observa en esta figura, en general, reaparecen los mismos núcleos básicos señalados por los diccionarios y el corpus paralelo[104]. Sin embargo, llama la atención la presencia de una categoría llamada *X*, que reúne todas las ocurrencias del verbo que no se dejan clasificar según los núcleos descubiertos anteriormente. Obsérvese a título ilustrativo el ejemplo siguiente:

> (7) Debes de estar impaciente -dije, *sintiendo* el sabor a mala leche en mi propia voz, una voz insolente que no sabía de dónde venía. (CREA: Ruiz Zafón C., 2003)

Este ejemplo ilustra que el locutor, dándose cuenta de la riqueza semántica del verbo, puede establecer deliberadamente cierto juego entre los distintos significados posiblemente presentes. El verbo podría interpretarse como un uso metafórico de la percepción gustativa por la presencia del SN *sabor a mala leche*. No obstante, es bien sabido que la expresión *estar de mala leche* refiere a cierto estado de ánimo, o sea el mal humor del locutor, por lo que se añade también cierto matiz subjetivo. Finalmente, por la presencia del SP *en mi propia voz*, el locutor juega también con el sentido auditivo del verbo *sentir*.

En suma, en vez de considerar los núcleos semánticos del verbo *sentir* como separados o significados discretos, más bien vale reconocer la existencia de zonas transitorias y traslapos entre ellos y caracterizarlos en términos de continuidad.

103 Para más detalles sobre la composición del corpus, véase la bibliografia.
104 Por motivos prácticos estos resultados no se discutirán en detalle.

5. Conclusiones

El punto de partida de este estudio ha sido la hipótesis de que los verbos *sentir* en español, francés e italiano son cognados perfectos, lo cual ha sido verificado en tres etapas.

En primer lugar, el estudio lexicográfico (básicamente introspectivo) ha mostrado la necesidad de rechazar esta hipótesis inicial y establecer la hipótesis alternativa de que los verbos *sentir(e)* no son cognados perfectos. Sin embargo, al mismo tiempo, esta aproximación introspectiva también hace surgir dos problemas, o sea primero ¿cómo podemos medir las diferencias? Y en segundo lugar ¿cuáles son los núcleos semánticos principales del verbo?

Luego, el análisis de corpus paralelo se ha revelado fundamental para superar estas limitaciones del método introspectivo porque permite (1) expresar cuantitativamente el grado de correspondencia entre estos cuasi-sinónimos en las tres lenguas y (2) designar con más precisión los núcleos semánticos inherentes al verbo. Sin embargo, hemos observado que este estudio del corpus paralelo genera a su vez otras preguntas. Así, el estudio de un corpus comparable permite refinar otra vez los resultados previos, apuntando hacia la existencia de continua entre los núcleos semánticos en vez de categorías discretas.

6. Bibliografía

Corpus:

Astrologo, M. (1998). *Harry Potter e la Pietra Filosofale*. Milano: Adriano Salani Editore [HPIT].

Atilf: *Base Textuelle Frantext*. www.frantext.fr.

Camilleri, A. (2005). *La luna di carta*. Palermo: Sellerio editore.

Dellepiane, A. (1999). *Harry Potter y la piedra filosofal*. Barcelona: Emecé Editores S.A. [HPESP].

Giorgetti Cima C. (2007). *Uomini che odiano le donne.* Venezia: Marsilio.

Grumbach L., & De Gouvenain, M. (2006). *Les hommes qui n'aimaient pas les femmes*. Arles: Actes Sud.

Il Corriere della Sera: archiviostorico.corriere.it/searchresultsArchivio.jsp.

Larsson, S. (2005). *Män som hatar kvinnor.* Stockholm: Norstedts Förlag.

Le Monde: www.lemonde.fr.

Lexell M. & Ortega Román J. J. (2008). *Los hombres que no amaban a las mujeres*. Barcelona: Ediciones Destino.

Lo Gatto, A. (2004). *L'intreccio di universi paralleli.* www.latelanera.com/files/ebook054.pdf

Ménard, Jean-François (1998). *Harry Potter à l'école des sorciers*. Paris: Gallimard Jeunesse [HPFR].

Real Academia Española: *Corpus de Referencia del Español Actual*. www.rae.es [CREA].

Rowling, J. K. (1997). *Harry Potter and the Philosopher's Stone*. London: Bloomsbury.

Obras citadas:

De Mauro, T. (1999). *Grande dizionario Italiano dell'uso*. Torino: Utet [GDIU].

Geeraerts, D. (2010). The doctor and the semantician. In D. Glynn & K. Fischer (Eds.), *Quantitative methods in cognitive semantics: corpus-driven approaches* (pp. 63-78). Berlin / New York: Mouton de Gruyter.

Gilquin, G. (2008). Causative Make and Faire: A Case of Mismatch. In M. Gómez González, J. L. Mackenzie & E. M. González Álvarez (Eds.), *Current Trends in Contrastive Linguistics: Functional and Cognitive Perspectives* (pp. 177-201). Amsterdam: John Benjamins.

Glynn, D. (2010). Corpus-driven Cognitive Semantics. Introduction to the field. In D. Glynn & K. Fischer (Eds.), *Quantitative methods in cognitive semantics: corpus-driven approaches* (pp. 1-41). Berlin / New York: Mouton De Gruyter.

Johansson, Stig (1999). Towards a multilingual corpus for contrastive analysis and translation studies. In L. Boris (Ed.), *Parallel corpora, parallel words. Selected papers from a symposium on parallel and comparable corpora at Uppsala University* (pp. 47-59). Amsterdam: Rodopi.

McEnery, T., & Xiao, R. (2008). Parallel and comparable corpora: what is happening? In G. Anderman & M. Rogers (Eds.), *Incorporating corpora. The Linguist and the Translator* (pp. 18-31). Clevedon: Multilingual Matters Ltd.

Moliner, María (1998[2]): *Diccionario de uso del español*. Madrid: Gredos [DUE].

Noël, D. (2003). Translations as evidence for semantics: an illustration. *Linguistics 41,* 757-785.

Oster, U. (2010). Using corpus methodology for semantic and pragmatic analyses: What can corpora tell us about the linguistic expression of emotions? *Cognitive Linguistics, 21,* 727-763.

Robert, P., Rey, A., Rey-Debove, J. (2008). *Le Nouveau Petit Robert. Dictionnaire alphabétique et analogique de la langue française. Nouvelle édition du Petit Robert de Paul Robert.* Paris: Dictionnaires Le Robert [PR].

Talmy, L. (2007). Foreword. In M. González-Márquez, I. Mittelberg, S. Coulson & M.J. Spivey (Eds.), *Methods in Cognitive Linguistics*. Amsterdam / Philadelphia: John Benjamins.

VANDERSCHUEREN, C. (2010). The use of translations in linguistic argumentation. A case study on Spanish and Portuguese subordinate clauses introduced by para. *Languages in Contrast, 10,* 76-101.

VAN HOECKE, W. & GOYENS, M. (1990). Translation as a witness to semantic change. *Belgian Journal of Linguistics, 5,* 109-131.

A corpus-based study on the use of narrative in English and Spanish youth conversation

Monica Palmerini and Serenella Zanotti

Roma Tre University, Italy

Abstract

In this paper we present the first results of a corpus-based analysis of narrative in Spanish and English youth language. The study focuses on conversational narrative, i.e. on natural occurring narratives that emerge in young people's everyday interactions. Our analysis is based on two Bergen Corpora of teenage informal conversation: the *Corpus of London Teenage Language* (COLT), and the Madrid subcorpus of the *Corpus Oral de Lenguaje Adolescente* (COLAm). In this paper we direct our analysis to some specific aspects of youth narrative: namely story openers and closers, and quotation strategies. The analysis highlights remarkable similarities in narrative construction in the two linguistic communities and illustrates how the study of both youth language and conversational narrative can benefit from the use of comparable corpora.

Key words: corpus linguistics, conversational narrative, youth language, Spanish, English.

Resumen

En este artículo presentamos los primeros resultados de un análisis basado en corpora de narraciones en el lenguaje juvenil español e inglés. La investigación se centra en la narración conversacional, es decir, en las narraciones espontáneas que emergen en las interacciones cotidianas de los jóvenes. El análisis se basa en dos corpora juveniles de conversación informal elaborados en la Universidad de Bergen: el *Corpus of London Teenage Language* (COLT) y el subcorpus de Madrid del *Corpus Oral de Lenguaje Adolescente* (COLAm). En esta contribución se analizan algunos aspectos específicos de la narración: es decir, los recursos de apertura y de cierre de la narración y las estrategias de cita. El estudio subraya considerables semejanzas en la construcción de la narración en las dos comunidades lingüísticas y pone de relieve cómo la investigación sobre el lenguaje juvenil así como sobre la narración conversacional se puede beneficiar del uso de corpora comparables.

Palabras clave: lingüística de corpus, narración conversacional, lenguaje juvenil, español, inglés.

Introduction[105]

Pioneered by Labov and Waletsky's seminal work (1967; Labov 1972; 1997), narrative analysis, i.e. the study of oral narratives of ordinary people, developed on the basis of elicited, interview-style stories. In more recent times, a new direction of research has focused attention on the simplest and most fundamental context where narrative surfaces, i.e. spontaneous informal conversation (Wolfson, 1982; Polanyi, 1985; Toolan, 2001: 146-182; Norrick, 2000 and 2007; Langellier & Peterson, 2004). These studies have pointed out that, rather than a de-contextualized phenomenon, narrative in its everyday dimension is to be regarded as "a conversational strategy for accomplishing some interactional end" (Norrick, 2000: 1-2).

In this paper *conversational narrative* will be approached combining two different perspectives: a sociolinguistic one, as the analysis will concentrate on youth language; and a contrastive one, as it will offer a comparison of the use of narrative in British English and Peninsular Spanish youth talk. The overall approach envisaged is ultimately corpus-based, since the analysis has been carried out on and through two comparable corpora of youth language that have both been constructed at the University of Bergen: the *Corpus of London Teenage Language* (COLT), and the Madrid subcorpus of the *Corpus Oral de Lenguaje Adolescente* (COLAm).[106]

Studies carried out over the last decade (Briz, 2003; Rodríguez, 2002; Androutsopoulos & Georgakopoulou, 2003; Stenström & Jørgensen, 2009; Jørgensen, 2010 and Bucholtz, 2011) have demonstrated the interest of youth language as a site of innovation and identity construction and paved the way for further research from a wide range of perspectives.

Single-language and contrastive studies have been carried out on the two Bergen corpora, which have investigated different aspects of youth language, with special reference to discourse markers, intensifiers, taboo words, tags, and phatic talk (Stenström, Andersen and Hasund, 2002; Stenström, 2005a, 2005b, 2006a, 2006b, 2006c; Jørgensen, 2008b, 2008c, 2009). However, none of these studies has specifically addressed the issue of narrative. More generally, it should be noted that, while a general model for the analysis of conversational narrative has been offered (see for instance Toolan, 2001; Blum Kulka, 1993; Norrick, 2000), very little has been done in the field of narrative in youth conversation, with the notable exception of Cheshire, 2000, 2003 and Cheshire & Williams, 2002.

In this paper we intend to take a further step in this direction, presenting the first findings of a corpus-based investigation on how adolescent speakers in two of the most spoken and influential languages in the world, English and Spanish, use and construct narrative in conversation.

At this early stage of our research we have focused our attention on two main aspects:

1. *story opening and closing* (see Cheshire, 2000: 251 and Norrick, 2007: 132), as we intended to investigate the dynamics between narrative and non-narrative space and study

105 This article is the result of a joined effort between the two authors. It should be specified, however, that Monica Palmerini is responsible for part 3 and Serenella Zanotti for part. 2. The other sections are to be ascribed to both authors.

106 Both corpora are available on the internet at http://torvald.aksis.uib.no/colt/cwb and http://www.colam.org.

the way young speakers mark the boundaries between "narrated world" and "commented world" (Weinrich, 1964);

2. *reported speech*, i.e. quotation strategies and other devices used by young speakers to mark the boundary between their own and the other people's voices (see Tannen, 1989 and Mayes, 1990).

1. The data: corpora and sample selection

The starting point and an inescapable prerequisite for our project was the availability of two comparable corpora of oral informal youth language, namely the above mentioned COLT and COLAm. These corpora appear particularly suitable to stimulate research on the communicative style of the young and to carry out sociolinguistic surveys, since they are complemented by important background information on the speakers.

COLT is a collection of spontaneous conversations among London teenagers recorded in 1993 by student recruits aged 13 to 17. COLAm is a the Madrid subcorpus of the larger COLA project, aimed at building a corpus of informal youth language from various capitals of Spanish speaking countries (to date, it includes Argentina and Chile). It was compiled in Madrid in 2002-2004 on the COLT model using more sophisticated equipment (Jørgensen 2008a). Both corpora contain approximately half a million words. Thus they are perfectly comparable in terms of corpus design, method of data collection, and number of words, although they differ in chronological context, there being a gap of nine years between the recordings.

The contrastive analysis that is being presented here has been conducted on a sample of 15 conversations from each corpus, casually selected on the basis of containing narrative signals such as *tell me*, *you know what*, and *you know when*. The total number of words for both samples is around 50.000, thus amounting to approximately 1/10 of the total corpus extension.

2. The COLT data

2.1. *Narrative density*

The sample relating to the London teenagers consists of 15 conversations recorded by 9 recruits (5 girls and 4 boys). The conversations are variable in length, ranging from 617 to 10.798 words. As shown by table 1, some of the recruits contribute to the sample with more than one conversation.

Table 1. Narrative densit

Recruit	Conversations per recruit	Sex	Age	Words per conversation	Narrative units per conversation
Josie	4	F	14	1145 10.025 1080 7331	4 27 3 14
Cassie	1	F	15	4741	7
Craig	1	M	13	617	4
Sarah	2	F	13	9956 1176	4 2
Skonev	1	M	12	2586	9
Alex	1	M	14	3398	4
Caroline	1	F	14	3313	4
Jack	3	M	16	2719 6436 3839	1 12 3
Catriona	1	F	16	10.798	18
TOT.	**15**			**69.160**	

Narrative density and length vary from conversation to conversation and from recruit to recruit. Table 2 presents the data regarding the number of narratives per teller.

Table 2. Narrative units per teller

Recruit	Narrative units	Tellers
Josie 1	4	Josie
Josie 2	27	Josie
Josie 3	3	Josie
Josie 4	14	Josie (9) W34 (2) (Josie collaborates and joins in) W2 (3)
Cassie	7	Cassie (4) W26 (3)
Craig	4	Craig (2) W2 (2)
Sarah 1	4	Sarah (2) W2 (1) Sarah + W4+W5+W2 (1)
Sarah 2	2	W6 (2)
Skonev	9	W2 + W3 (1) W3 (2) W3 + W2 (2) W2 (2) Skonev + W2 +W3 (2) Skonev (2)
Alex	4	W 23 (1) W22+W23 (1) W31 (1) W26+W23 (1)
Caroline	4	W7 (1) Caroline (2) W2 (1)
Jack 1	1	Jack (1)
Jack 2	12	W4+W8+W9 W8 (4) W8+W7(1) W4+Jack (1) Jack (1) W7 + W1 W9 (1) W4 + W7 (1)
Jack 3	3	Jack (1) W13 (1) W7 (1)
Catriona	18	Catriona (9) W4 (8) W7 (1)

It is evident that some of the speakers are particularly active and prolific as narrators. The data also reveals that group or polyphonic narratives are typical of male friendship groups, whereas single-teller narratives are the norm among girls[107].

2.2. *Story openers*

Narratives can be prompted in conversation by other speakers, who may invite other participants to tell a story (e.g. "*right, Eleni and Lucinda, go on you explain the [story]*", "*Okay tell me what happened last night*"). However, elicited narratives are quite rare in our sample, the most recurrent situation being one with tellers struggling to take the floor.

Boundaries between narrative and non-narrative space are signalled in a number of ways by means of story openers and story closing devices. The opening structures that occur more frequently in the sample can be grouped into 6 main categories:

1. opening formulae, which include structures aimed at emphasizing the teller's subjectivity (e.g. *I'll tell you what happened*), questions containing a request for telling rights (e.g. *Did you hear about*, *(Do) you know what*) and markers of shared reminiscence (e.g. *Do you remember when*, *You know how, You know, You know when*) (see Cheshire, 2000: 251);
2. temporal clauses (e.g. *and yesterday, one day, last year*, etc.);
3. left dislocation of thematic element (e.g. *like, my dinner last night yeah*);
4. discourse markers, which include connectors (*so, well, right, okay*, etc.: e.g. *so, and he goes*), conative signals (*listen to this; watch this*), and markers of discontinuity (typically realized by a negation: e.g. *No, I was, I was taking the trombone*);
5. presentative structures (e.g. *there was this*, *there is this*, etc.: e.g. *There's this girl and she comes home*);
6. zero opener, which entails that the narrative opens up either with a so called abstract,[108] that may take the form of a plot summary, or with an announcement making claims about the interest of what will follow (see Toolan, 2001: 154). Sometimes the lack of a story opener is due to a delay in the moment of recording or to a material break in the tape.

In the case of chained narratives, which are stories that follow one another, the absence of a story opener can be explained with the fact that the narrative space is not interrupted and teller 2 simply takes over the telling from teller 1. This is often the case in recruit Skonev's conversation.

2.3. *Story closing signals*

Narrative is normally closed by tellers themselves, who normally give explicit evaluations on the narrated events so as to orient the reception of the tale by the other participants (e.g.

107 This is in line with Jenny Cheshire's findings (Cheshire, 2000).

108 Labov and Waletsky (1967) recognise six essential components in the internal structure of a narrative: Abstract, Orientation, Complicating action, Evaluation, Resolution and Coda.

I couldn't stop laughing; *That was wicked*). However, tellers may also leave evaluation to the other participants. Another aspect worth mentioning is that the end of the narrative often coincides with a climactic segment that is typically followed by laughter, which is itself one of the most recurrent signals of story closing. Often the climactic segment consists of a gesture or sound performed by the teller, which seals the narrative and opens space for new stories (e.g. *there's me wisht gone!*).

Other recurrent closing strategies include:

a. temporal shifts (e.g. … *but ever since I've never been able to handle it*);

b. recapitulation (e.g. *That was, that was what erm Dempsey done to Jane and Mussy*);

c. framing (that is, the use of a structure parallel to that of the opener);

d. use of discourse markers (e.g *you know?*, *you know what I mean*?);

e. phrase-closing tags (e.g. *she wasn't interested in* and stuff).

2.4. *Quotation strategies*

The part played by reported speech in the narratives of London teenagers is worth noting[109]. In particular, *narratives of saying* (Toolan, 2001: 158), that is stories that focus on what was said, are preponderant and even in *narratives of action* the quotation of the other people's words is essential and almost never missing.

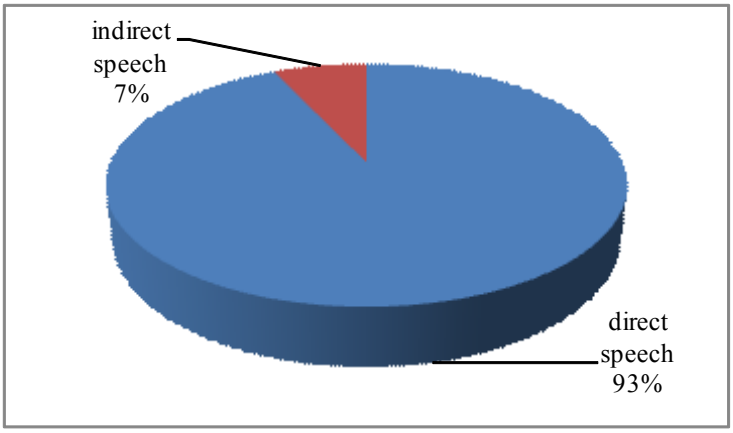


Figure 1. Quotation strategies

As clearly indicated by the graph, the strategy of quotation that is predominant in the sample is direct speech, with indirect speech confined to a mere 7.4%. Hybrid forms of reported speech are also documented.

109 This aspect is discussed in detail in Stenström, Andersen and Hasund (2002, ch. 5).

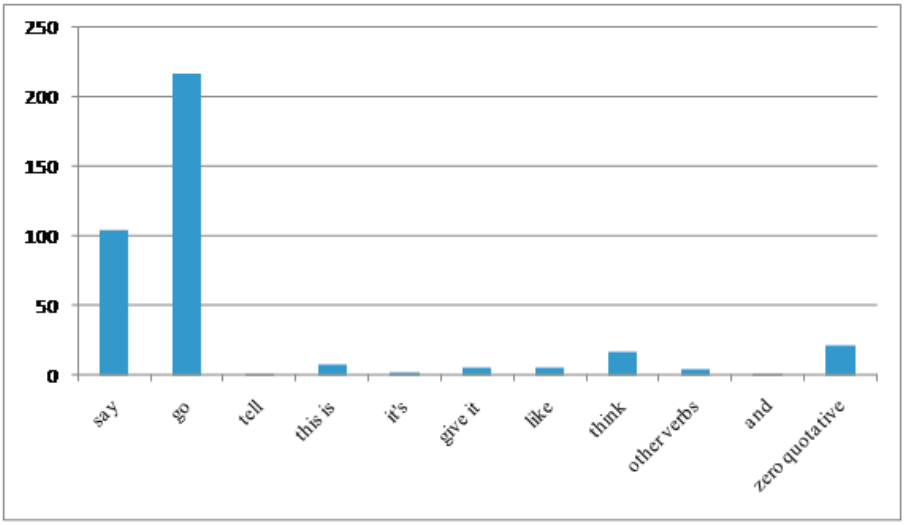


Figure 2. Quotatives

As figure 2 shows, the quotative that is most largely employed in introducing direct discourse in London teenage talk is the verb *go*. In particular, *go* seems to be the preferred quotative in combination with mimicking, sound effects and interjections, whereas *say* seems to be restricted to bracketing purely verbal utterances (see Stenström, Andersen and Hasund, 2002: 123-4 and Levey, 2003: 27).

Indeed, the quotative system of London teenagers is extremely varied and includes an array of forms, the most notable among them being *give it*, *this is /there's + subject* and *like*.[110] The occurrences of quotative *give it* in COLT are restricted to the conversations recorded by two recruits, Josie and Jock: e.g.

- Jane *gives it* I'll shoot you with my machete.
- my sister *gives it* <mimicking> Sam! Let go of my hair!
- And she *gives it* <nv> mimicking licking sound </nv>
- and when you <unclear> she just *gives it*, no, leave them alone!...
- She *gives it*, well one of them's called Kevin.
- and now he *gives it* { nv } whining { /nv} { mimicking girlie voice } Oh yes Paul { /}

According to Harris (2006: 114), *give it* may be an instance of an Indian language retention.

110 For *like* in British youth talk see Andersen 2001 (ch. 5) and Stenström, Andersen and Hasund (2002: 108-9). See also Macaulay (2001) for quotative *like* in Glasgow adolescent language, Tagliamonte & D'Arcy (2004) and Tagliamonte & Hudson (1999) for Canadian English, and Winter (2002) for Australian English. Buchstaller (2001) discusses *like* as a "mimesis marker" in American English.

The presentative structures *there's / this is + subject* is one of the most innovative traits of London teens' talk (see Fox & Cheshire, 2007). In the COLT, these structures are used to introduce both a verbal quotation and a gesture/facial expression, as illustrated by the following examples:

There's me + quote

There's me don't you dare th= she stopped.

There's me, have you <unclear> he said <mimicking

There's me + gesture

There's me!... Cos every time I hear a spider I

There's me!... Got up. Didn't even go, bother go

This is + subject

he goes, this is for you *this is me*, thanks. <laughing>

This is my brother and sister, ah! <laughing>

this is Jane to me the other day , throw your kitten off my floor

This is me + gesture

and he grabbed hold of me bum! *This is me...*

As the above examples highlight, the performative quality of teenagers' narrative is particularly evident in the COLT corpus.[111] London teenage speakers commonly resort to quotation in their narratives; they also seem to privilege multimodal enactments involving verbal as well as non verbal resources, including sound effects and gestures (see Levey 2003: 27). More often than not reported utterances are expressed with strong emotion and mimicry, as indicated by the number of occurrences of the label "mimicking" in the transcripts of the conversations.

3. The COLAm data

3.1. *Narrative density*

As opposed to the wide literature available for English, it should be noted that there are not so many studies addressing the analysis of conversational narrative in Spanish (Silva-

111 See Stenström, Andersen and Hasund (2002: 110-14). On narrative as performance see Goffman (1986) and Toolan (1988: 165-9).

Corvalán, 1987; Baixauli, 2000; Shiro, 2007) and, more specifically, in Spanish youth talk. This issue appears to have been tackled in a rather unsystematic way in contributions which deal with general features of youth language and colloquial Spanish as well as in studies devoted to other specific aspects of the language of the young, mainly discourse markers (see Briz, 1998, 2003; Rodríguez González 2002; Stenström, 2009).

Table 2. presents the basic information on the Spanish sample from COLAm, which includes recordings by 11 Madrid teenage recruits (mostly girls). The length of the 15 conversations[112] considered for the analysis varies, ranging from approximately 1.000 words to almost 7.000, while the number of narrative units per conversation oscillates between one and five. As far as narrative density is concerned, the most narrative-oriented conversation is the medium length one (MALCE2_12a around 3.000). As to the use of narrative function in the conversations, two speakers, MAORE2J01[113] and MAORE2J02, appear to be especially active as narrators in their interactions[114].

Table 2. Narrative density

Conversation	Words per conversation	Narrative units per conversation
MABPE2a	2709	3
MABPE2b	1161	2
MAESB2_01c	1724	1
MAESB2_06a	2953	3
MALCC2_08	6047	4
MALCE2_04a	3434	2
MALCE2_04b	2481	1
MALCE2_12	3486	3
MAORE2_07b	5712	3
MAORE2_08	6917	4
MAORE2_10	3575	2
MAORE2_12a	3160	5
MAORE2_12b	2472	4
MAORE2_12c	2195	4
MALCE2_01	1116	1
TOT.	**49.142**	**42**

112 In the table we keep the labels used to identify the conversation files in the COLAm corpus.
113 The speakers in COLAm are indicated through the conversation labels.
114 The internal structure of the COLAm narratives (single-teller, chained, polyphonic, etc.) and gender differences in storytelling styles have not been considered here but will be the object of another paper..

3.2. *Story openers*

The story opening devices used by Madrid teenagers to mark the switch from non narrative to narrative share the function of aiming at grabbing the listeners' attention. The following main types of story openers have been identified:

1. opening formulae, i.e. clauses focusing on new or shared knowledge. These opening devices typically take the form of interrogative structures (e.g. ¿s*abes lo que me ha dicho?; pero ¿qué te iba a decir?*; *¿te acuerdas la vez que...?*; *fue a ti a la que te llamé el otro día la casa de piluca*) or negative structures (e.g. *no viste la pelea que tuvimos)*. Quite interestingly, young speakers often signal the transition into storytelling with the negative indefinite pronoun *nada* 'nothing' (e.g. *bueno y entonces nada…, y nada…*) (cf. Stenström, 2009). Another interesting point is that the first person subject pronoun and the possessive *mi* are often used as story openers, especially when several tellers struggle to take the floor (e.g. *yo tenía un pollito…yo mira a ver he tenido un conejo; en mi casa también llega un animal y se muere*).

2. discourse markers: these linguistic units, which do not have a syntactic function but rather a pragmatic one, also typically occur in story opening. They can be of two kinds: a) *phatic signals*, which include forms such as *oye, mira*, *tía*, *ah*, aimed at establishing communication with the audience (Jorgensen, 2008b; Jorgensen & Martínez 2010) (e.g. tía *os he contado lo de mi abuela*), or exclamative signals like ¡*ay!* (e.g. ¡ay!, *el otro día leí lo del Doñana*); b) c*onnectors*, marking continuity (*pues, bueno, claro) (e.g.* claro *es que sí últimamente estoy comiendo tostadas),* the copulative conjunction *y*, the conjunction *que* (e.g. que *el zenith se tiró a la hermana pequeña del chollín*) and other devices signalling discontinuity (*pero*).

3*)* presentative structures, i.e. focalizing constructions used to introduce the relevant narrative sequence. The typical device here is *es que*, often preceded by a discourse marker (e.g. *ah bueno es que yo el otro día vi a uno*; *es que no he desayunado*).

4) thematic opening: this opening strategy consists in the left dislocation of some thematic element of the story, which helps to introduce the main subject of the conversation. It can be represented by a) a temporal expression (*ayer, el otro día*); b) a locative expression (e.g. en el bus *estabamos todos los tíos*); c) an expression pointing at a specific referent (e.g. el mensaje *que tengo escrito para mandárselo, Marcos*).

5) zero opener, usually with the direct enunciation of the so called "abstract" (see note 3) of the story, without any introductory element. In our sample this strategy is mainly based on declarative structures (e.g. *casi nos pegamos*; *mi madre me hizo engordar*; *bueno llegó la facture*; *me escribió un mensaje*).

6) an additional category is that of re-opening the story after an interruption, which is marked by discourse markers such as *bueno* and *pues*, or by explicit attempts on the part of the speaker to continue his/her telling (e.g. *ah pues ayer...*; *y nada...*; *bueno entonces nada...*; *bueno pues eso tía*; *bueno te sigo contando*).

3.3. *Story closing signals*

The narrative sequence usually ends immediately after reaching a climactic point in the narrative, typically dramatized and "performed" to hold the listeners' attention. Frequently the closing of a narrative sequence combines the telling of events with an evaluative component, with the aim of influencing the audience's interpretation of the story (e.g. y *ella por favor dímelo y yo mírala tío*).

The main phenomena relating to narrative closing can be grouped into the following categories:

1. pure narrative sequence, often in direct speech (e.g. *y se lo dieron gratis, salimos y la pillamos juntos*);
2. evaluation, either by the tellers themselves (e.g. *que rayada chaval, es que lo borré sin querer*) or by the listeners, who sometimes produce metanarrative comments (e.g. *como si fuese* [...] *pero es como una película*);
3. temporal shifts, marking the switch from the past experience to the present (e.g. *y desde entonces no he vuelto a comer*);
4. shifts of point of view (e.g. *y bueno yo flipando*);
5. gestures, sound effects and laughter (e.g. *y yo a los tres cuartos de hora estoy* así <cara ridícula>);
6. constructions including the indefinite pronoun *nada* (*y nada, pues nada,* etc.) (e.g. *y* nada *en plan así y la pobre Raquel se pasó toda la puta fiesta sin un niño sabes*; *y* nada *y eso...pero bueno*).

3.4. *Quotation strategies*

The construction of narrative by Madrid teenagers is achieved through a massive use of quotations, thus making the narration resound with their own or other people's voices. As the diagram below shows, the analysis of the sample highlighted the heavy preponderance of direct speech, with 311 occurrences vs 47 instances of indirect speech.

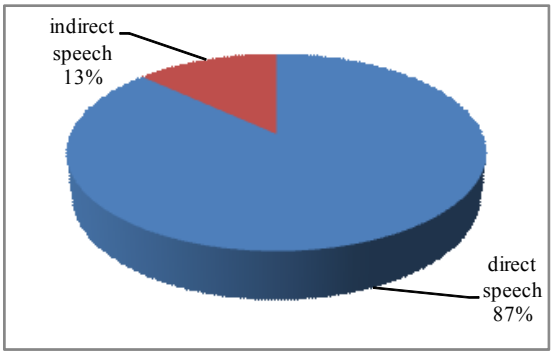


Figure 3. Quotation strategies

The following chart summarizes the most frequently encountered quotatives in Madrid conversations. The most commonly employed quotative is the verb *decir* 'to say', used mainly in its conjugated form (e.g. *y le* digo *creo que el chico se ha lanzado pero no sé; y* dije *y* dije*...no te da miedo; me* dice *es que no tengo no tengo presupuesto*), but documented also in its non finite, gerundive, progressive form (*y mi madre* diciendo y*o que voy a hacer por Serrano con los huevos*). Often the verb *decir* appears in particular expressive constructions after the verb *coger* 'to get' and *salir* 'to go out' (e.g. *y entonces* coge *piluca* y dice *mamá es para mí*). Another widespread quotative use involves personal subject pronouns (especially *yo*) and a few nominal expressions which, with the appropriate intonation, introduce direct discourse (*él/ella, el tío/la tía, el otro/la otra, el padre/la madre*): e.g. *y* yo *vale vale*; *y* la otra *ay muchísimas gracias*. On the other hand, the verb *poner* 'to put' seems to be a well established means to quote text messages (see Reyes, 1994; Maldonado González, 1999; Benavent, 2000). Also represented in the sample is the zero quotative, where direct speech appears without any signalling device.

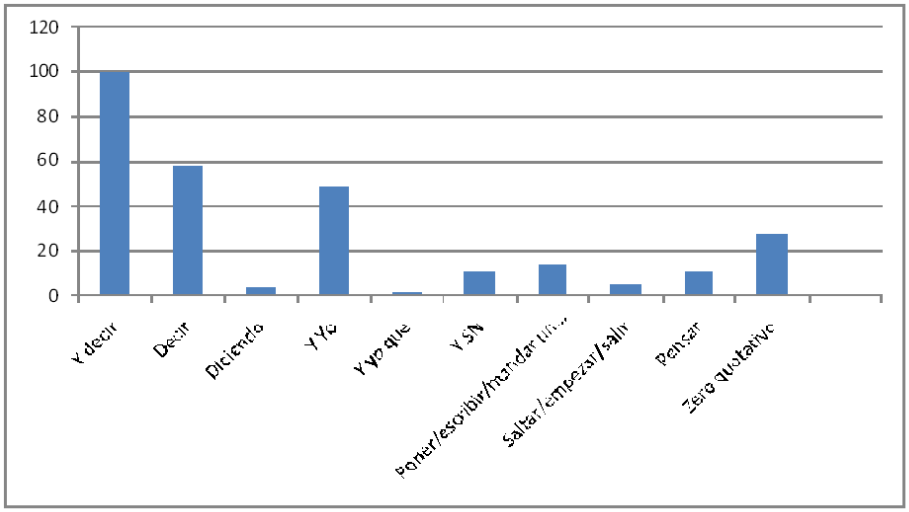


Figure 4. Quotatives

The stories told by the Spanish adolescents in the sample seem to primarily aim at recreating a vivid, enthralling representation of past events and situations as well as verbal exchanges. This aspect manifests itself in the frequent use of sound effects, gestures and a particular way of speaking, often in combination with direct speech (e.g. *para que al salir* <yuuuun> *se cayese para abajo; me dice* <imitando> *mejórate no se qué; me dice* <uaaaa> *apaga el teléfono apaga el ordenador; bajando a la piscina* así). The multimodality of these narratives is signalled in the corpus transcriptions by several descriptive labels such as <risa> *ja ja ja*", <imitando>, <ironía>, <cara ridícula> (e.g. *y yo a los tres cuartos de hora estoy así* <cara ridícula>. Non verbal sounds such as mobile ring tones are also often reproduced in the narratives (*y suena* pí*, primero* pirili *y luego* pirili pirili*, ya está* pí).

4. Contrastive analysis

A preliminary comparison of the English and Spanish data suggests that some fundamental properties of youth talk can be considered universal. It seems clear that storytelling represents an important social activity for young speakers. As our data suggest narrative has a central role in teenagers' conversation and contributes to building in-group solidarity and sense of shared values. The activity of storytelling seems to be a means for the young to reinforce their image as group members and to have their points of view, values and models of behaviour confirmed.

The prominence of direct speech appears to be a general trait of this conversational style and can be related to a number of factors such as:

- difficulty in handling complex syntax (see Levey, 2003);
- young speakers' preference for vivid reconstructions of events, whereby narrative inevitably turns into performance;
- the use of quotation as a powerful and subtle evaluative device. Indeed, research in the field of adolescent talk has pointed out its essentially "emotional and implicit nature" (Levey, 2003: 28). It can thus be argued that teenagers resort to quotation as a means of implicitly expressing evaluation.

It appears the worlds constructed through narrative by the London and the Madrid teenagers are essentially verbal in nature, with *narratives of saying* having equal if not greater representation in the corpora compared to *narratives of action*. It was also evident that the quotative system of teenage speakers is not only varied but sometimes also quite innovative. It has been suggested that "new quotative expressions fill a niche caused by a change in narrative style across the generations" (Fox & Cheshire, 2007) and that some structures are "elected when the narrative is 'performed' and the speaker adopts the stance of one of the participants in the event being constructed" (Fox & Cheshire, 2008).

Another point of contact between the two young linguistic communities can be found in the exploitation and the manipulation of suprasegmental traits, such as pitch, tempo and voice quality, for communicative effect and as a means of expressing evaluation.

Conclusions

In this paper we have offered a corpus-based contrastive investigation of the conversational narrative of London and Madrid teenagers. The analysis has shown some interesting similarities between teenagers' storytelling in the two languages. We believe that corpus-linguistics has the potential to open up new frontiers in this field of study. In particular, we argue for the construction of multilingual corpora of youth language so as to stimulate further contrastive research. This would lead to an awareness of specificities and similarities among youth varieties across languages in an increasingly global scenario.

Although most of the work is still to be done in terms of both methodological approach and data collection/analysis, it is quite clear that the study of conversational narrative may contribute to a better understanding of the complex phenomenon that is youth language.

Bibliography

Androutsopoulos, J. & Georgakopoulou, A. (Eds). (2003). *Discourse Constructions of Youth Identities*. Amsterdam: John Benjamins.

Andersen, G. (2001). *Pragmatic Markers and Sociolinguistic Variation: A Relevance-Theoretic Approach to the Language of Adolescents*. Amsterdam: John Benjamins.

Baixauli, I. (2000). Las secuencias de historia en la conversación coloquial. In A. Briz & Grupo Val.Es.Co. (Eds.), *¿Cómo se comenta un texto coloquial?* (pp. 81-107). Barcelona: Ariel.

Benavent Payá, E. (2000). La polifonía en la conversación coloquial: el caso del relato dramatizado. In *Actas del IV Congreso de Lingüística General* (Cádiz, 3-6 marzo de 2000), (vol. II, pp. 215-225). Cádiz: Servicio de Publicaciones de la Universidad de Cádiz.

Blum Kulka, S. (1993). 'You gotta know how to tell a story': Telling, tales and tellers in American and Israeli narrative events at dinner. *Language in Society*, 22, 361-402.

Briz, A. (2003). La interacción entre jóvenes. Español coloquial, argot y lenguaje juvenil. In M. T. Echenique Elizondo & J. Sánchez Méndez (Eds.), *Lexicografía y Lexicología en Europa y América* (pp. 141-154). Madrid: Gredos.

Bucholtz, M. (2011). *White Kids: Language, Race and Styles of Youth Identity*. Cambridge: Cambridge University Press.

Buchstaller, I. (2001). 'He goes and I'm like'. The new quotatives revisited. Paper presented at NWAVE 30, Raleigh, North Carolina, October 2001.

Cheshire, J. (2000). The Telling or the Tale? Narrative and Gender in Adolescent Friendship Networks. *Journal of Sociolinguistics*, 4(2), 234–62.

Cheshire, J. (2003). Social dimensions of syntactic variation: the case of *when* clauses. In D. Britain & J. Cheshire (Eds.), *Social Dialectology* (pp. 245-261). Amsterdam: Benjamins.

Cheshire, J. & Fox, S. (2008). Performed narrative: The pragmatic function of 'this is me' and other quotatives in London adolescent speech. Paper presented at 17th Sociolinguistics Symposium, Amsterdam 3-5 April 2008.

Cheshire, J. & Williams, A. (2002). Information structure in male and female adolescent talk. *Journal of English Linguistics*, 30, 217-238.

Fox, S. & Cheshire, J. (2007). *This is me/this is him:* The quotative system of London adolescents. Paper presented at NWAV 36, University of Pennsylvania, October 11-14, 2007.

Goffman, E. (1986). *Frame Analysis: An Essay on the Organization of Experience*. Boston: Northeastern University Press.

Harris, R. (2006). *New Ethnicities and Language Use Houndmills*. Basingstoke: Palgrave.

Jørgensen, A. M. (2008a). COLA: un Corpus Oral de Lenguaje Adolescente, *Anejos a Oralia,* 225-235.

Jørgensen, A. M. (2008b). La funcion fática de los vocativos en la conversación juvenil de Madrid y Londres. In A. Briz et al. (Eds.), *Cortesía y conversación: de lo escrito a lo oral. III Coloquio Internacional Programa EDICE* (pp. 1-14). Valencia: Universidad de Valencia.

Jørgensen, A. M. (2008c). *Tío* y *tía* como marcadores en el lenguaje juvenil de Madrid. In I. Olza Moreno, M. Casado Velarde & R. González Ruiz (Eds.), *Actas del XXXVII Simposio Internacional de la Sociedad Española de Lingüística (SEL)* (pp. 387-396), Pamplona: Servicio de Publicaciones de la Universidad de Navarra.

Jørgensen, A. M. (2009). *En plan* used as a hedge in Spanish teenage language. In A. B. Stenström & A. M. Jørgensen (Eds.), pp. 95-11.

Jørgensen, A. M. & Martínez, J. A. (2010). Vocatives and phatic communion in Spanish teenage talk. In Jørgensen, J. N. (Ed.), pp. 193-210.

Jørgensen, A. M. & Stenström, A. B. (2008). ¿Una cuestión de cortesia? Estudio contrastivo del lenguaje fático en la conversación juvenil, *Special Issue of Pragmatics*, 18(4).

Jørgensen, J. Norman (Ed.) (2010). *Love Ya Hate Ya: The Sociolinguistic Study of Youth Language and Youth Identities*. Cambridge: Cambridge University Press.

Labov, W. (1972). The Transformation of experience in narrative syntax. In *Language in the Inner City*. *Studies in the Black English Vernacular* (pp. 354-396). Philadelphia: University of Pennsylvania Press.

Labov, W. (1997). Some further steps in narrative analysis. *Journal of Narrative and Life History* 7(1-4), 395-415.

Labov, W. & Waletsky, J. (1967). Narrative analysis: oral versions of personal experience. In Helms, J. (Ed.), *Essays on the Verbal and Visual Arts* (pp. 12-44). Seattle: University of Washington Press.

Langellier, K. & Peterson, E. (2004). *Storytelling in Daily Life: Performing Narrative*. Philadelphia: Temple University Press.

Levey, S. (2003). Reported dialogue and pragmatic particles in the narratives of preadolescents. *World Englishes*, 22, 305–321.

Macaulay, R. (2001). 'You're Like 'Why Not?' – The quotative expressions of Glasgow adolescents. *Journal of Sociolinguistics* 5(1), 3-21.

Maldonado González, C. (1999). Discurso directo y discurso indirecto. In I. Bosque y V. Demonte (Eds.), *Gramática descriptiva de la lengua española* (vol. 3, pp. 3551-3595). Madrid: Espasa.

Mayes, P. (1990). Quotation in Spoken English. *Studies in Language*, 14, 323-363.

Norrick, N. (2000). *Conversational Narrative: Storytelling in Everyday Talk*. Amsterdam: John Benjamins.

Norrick, N. (2007). Conversational storytelling. In Herman, D. (Ed.), *The Cambridge Companion to Narrative* (pp. 127-141). Cambridge: Cambridge University Press.

Polanyi, L. (1985). *Telling the American Story*. Norwood: Ablex.

Rodríguez González, F. (Ed.) (2002). *El lenguaje de los jóvenes*. Barcelona: Ariel.

Schiffrin, D. (1994). *Approaches to Discourse*. Cambridge, MA: Wiley-Blackwell.

Shiro, M. (2007). El discurso narrativo oral en la vida cotidiana: géneros y procesos, en A. Bolívar (Ed.), *Análisis del discurso* (pp. 121-143). Caracas: Colección Minerva, Manuales Universitarios.

Silva-Corvalán, C. (1987). La narración espontánea: estructura y significado. In E. Bernárdez (Ed.), *Lingüística del texto* (pp. 265-292). Madrid, Arco/Libros.

Stenström, A. B. (2005a). 'He's well nice - Es mazo majo'. London and Madrid girls' use of intensifiers. In. S. Granath, J. Millander & E. Wennö (Eds.), *The Power of Words. Studies in Honour of Moira Linnarud*. Karlstad: Karlstad University.

Stenström, A. B. (2005b). 'It is very good eh, Está muy bien eh'. Teenagers' of use of tags – London and Madrid compared. In K. Mc Cafferty, T. Bull and K. Killie (Eds), *Contexts Historical, Social, Linguistic. Studies in Celebration of Toril Swan*. Pieterlen: Peter Lang.

Stenström, A. B. (2006a). Taboo words in teenage talk: London and Madrid girls' conversations compared. *Spanish in Context*, 3, 116-138.

Stenström, A. B. (2006b). The Spanish discourse markers *o sea* and *pues* and their English correspondences. In K. Aijmer & A. M. Simon-Vandenbergen (Eds.), *Pragmatic Markers in Contrast.* Amsterdam: Elsevier.

Stenström, A. B. (2006c). The Spanish pragmatic marker *pues* and its English equivalents. In A. Renouf & A. Kehoe (Eds.), *The Changing Face of Corpus Linguistics*. Amsterdam/New York: Rodopi.

Stenström, A. B. (2009). Pragmatic markers in contrast. Spanish *pues nada* and English *anyway*. In A. B. Stenström & A. M. Jørgensen (Eds.), pp. 137-159.

Stenström, A. B., Andersen, G. & Hasund, I. K. (2002). *Trends in Teenage Talk*. Amsterdam: Benjamins.

Stenström, A. B. & Jørgensen, A. M. (Eds.) (2009). *Youngspeak in a Multilingual Perspective*. Amsterdam: Benjamins.

Tagliamonte, S. & D'Arcy, A. (2004). *He's like, she's like*: The quotative system in Canadian youth. *Journal of Sociolinguistics*, 8(4), 493-514.

Tagliamonte, S. & Hudson R. (1999). *Be like et al.* beyond America: the quotative system in British and Canadian youth. *Journal of Sociolinguistics*, 3(2), 147-172.

TANNEN, D. (1989). *Talking Voices: Repetition, Dialogue, and Imagery in Conversational Discourse*. Cambridge: Cambridge University Press (2nd ed. 2007).

TOOLAN, M. J. (2001). *Narrative: A Critical Linguistic Introduction*. London and New York: Routledge (2nd edition).

WEINRICH, H. (1964). *Tempus: Besprochene und Erzählte Welt*. Stuttgart: Kohlhammer.

WINTER, J. (2002). Discourse quotatives in Australian English: Adolescents performing voices. *Australian Journal of Linguistics*, 22(1), 5-21.

WOLFSON, N. (1982). *The Conversational Historical Present in American English Narrative*. Cinnarminson: Foris Publications.

'Well' in Spanish translations: evidence from the P-ACTRES parallel corpus

Noelia Ramón
University of León

The English adverb well *is multifunctional, carrying meanings related to manner, degree or intensification. In addition,* well *is often grammaticalized into a discourse particle, especially in dialogue, and this requires a particularly careful treatment in the case of translations, as discourse particles do not carry easily definable meanings. Previous studies on the English particle* well *have shown that the translation into other languages is far from straightforward. The translations of* well *have been studied in the cases of Norwegian, Swedish, Dutch, German and Italian, and this paper will expand the analysis to Spanish. The study will focus on the translations of* well *in the P-ACTRES English-Spanish parallel corpus, which will provide the empirical material for the analysis. The aim is to provide an inventory of translation solutions in Spanish for the various functions of* well *in English original texts, in particular with regard to its use as a discourse marker.*

Key words: discourse markers, translation, parallel corpus.

El adverbio inglés well *es multifunctional y expresa significados relacionados con manera, grado o intensificación. Además, esta partícula se gramaticaliza a menudo para convertirse en un marcador discursivo, especialmente en diálogos, y esto requiere un tratamiento muy cuidadoso en las traducciones, ya que los marcadores discursivos no expresan significados claramente delimitados. Estudios previos sobre el adverbio ingles* well *han encontrado que las traducciones a otras lenguas no son en absoluto simples. Se han estudiado las traducciones de* well *a lenguas como noruego, sueco, alemán, neerlandés e italiano, y en este trabajo pretendemos expandir este análisis al español. El estudio se centra en las traducciones de* well *en el corpus paralelo inglés-español P-ACTRES, de donde se extrajo el material empírico para el análisis. El objetivo es obtener un inventario de posibles traducciones de* well *al español, en especial en el caso de su uso como marcador discursivo.*

Palabras clave: marcadores discursivos, traducción, corpus paralelo.

1. Introduction

The English form *well* is multifunctional, conveying a variety of lexical meanings, as well as grammaticalized meanings as a discourse particle, especially in dialogue. This last meaning requires a particularly careful treatment in translations. Previous studies (Aijmer & Simone-Vandenbergen, 2003; Johansson, 2006) have shown that the translation of *well* into other languages is far from straightforward.

This study will focus on the translations of *well* in the English-Spanish parallel corpus P-ACTRES, which will provide the empirical material for the analysis. This corpus contains about 2.5 million words of contemporary English texts and their corresponding translations into European Spanish. The aim of the study is to provide an inventory of translation solutions available in Spanish for *well* in English original texts, in particular with regard to its use as a discourse marker.

2. Theoretical background

Pragmatic or discourse markers are short words or phrases, particularly frequent in spoken communication, which do not add any propositional content to the utterance they are included in, but rather express the speaker's attitude towards the listener, negotiate background assumptions or express other types of interpersonal or textual meanings that contribute to the overall texture and coherence of discourse. Important studies on discourse markers in the past 20 years or so include Schiffrin (1987), Jucker & Ziv (1998), Lenk (1998), Andersen & Fretheim (2000), Fischer (2000), Aijmer (2002). Discourse markers behave syntactically like interjections and tend to occur in sentence-marginal positions or parenthetically.

As far as form is concerned, the most frequent items in English spoken conversation (BNC/spoken) are *yeah, oh, no, well, but, just know, mm, yes, like* and *cos*. The form *well* is one of the English discourse markers that has attracted most attention from scholars, from an intralinguistic perspective (Bolinger, 1989; Carlson, 1984; Chafe, 1986; Fraser, 1990; Halliday & Hasan, 1976; Schourup, 2001), as well as from a contrastive perspective, analyzing translations of this particle into Swedish and Dutch (Aijmer & Simon-Vandenbergen, 2003) into Norwegian and German (Johansson, 2006) or into Italian (Bazzanella & Morra, 2000). This paper will focus on the discourse marker *well* and the Spanish translations of this particle found in the P-ACTRES parallel corpus.

There has been a clear process of grammaticalization and *well* "has lost most of its original meaning in its evolution from a lexical adverb to a discourse particle." (Aijmer & Simon-Vandenbergen, 2003: 1126). The elusiveness of the meanings expressed by well is shown in the following statement: "If a foreign language learner says *five sheeps* or *he goed*, he can be corrected by practically every native speaker. If, on the other hand, he omits a *well*, the likely reaction will be that he is dogmatic, impolite, boring, awkward to talk to etc, but a native speaker cannot pinpoint an 'error'." (Svartvik, 1980: 171)

I will follow Aijmer & Simon-Vandenbergen's (2003) approach as to the two main types of functions that may be expressed by *well* as a pragmatic marker in English. It contributes

to the interpersonal function of language in the form of a politeness marker, establishing some kind of respect towards the addressee's face and recognizing the need to renegotiate meaning shared by both; positive appraisal as well as counter-expectation are here the core functions. It contributes also to the textual function as a boundary marker and topic introducer, as it usually occurs in sentence-initial position.

Moreover, I have included another different use of *well* as a discourse marker in those cases where it followed a modal verb to form an idiomatic pattern as in (1). Examples such as this one show a high degree of grammaticalization of the particle well and are thus better treated as pragmatic markers too:

(1) *Extreme free variation may **well** have been the result, which it certainly is not. (ETG1E.s45)*

3. Methodology

The empirical data used for the analysis in this paper were extracted from the English-Spanish parallel corpus P-ACTRES compiled at the University of León, Spain. P-ACTRES contains original English texts and their corresponding Spanish translations. This corpus includes written material from a variety of different registers (fiction, non-fiction, newspapers, magazines & miscellanea) published in the year 2000 or later, thus representing the contemporary stage of the English language, and the corresponding translations published in the European variety of Spanish. Today P-ACTRES comprises nearly 2.5 million words, approximately 1.2 million words per language.

The English source texts and their corresponding translations into Spanish are aligned at sentence level and can be searched with the Corpus Work Bench browser. Table 1 shows the number of words in each subcorpus.

Table 1: Contents of the English-Spanish Parallel Corpus.

	ENGLISH	SPANISH	TOTAL
Books – fiction	396,462	421,065	817,527
Books – non-fiction	494,358	553,067	1,047,425
Newspapers	115,502	137,202	252,704
Magazines	119,604	126,989	246,593
Miscellanea	40,178	49,026	89,204
TOTAL	1,166,104	1,287,349	**2,453,453**

All the cases of *well* as an adverb were extracted from the corpus, together with their corresponding Spanish translations. The various syntactic functions were identified and the cases classified as pragmatic markers were checked for their translations. The working hypothesis is that the multiple functions of *well* will be translated in a number of different ways in Spanish, thus highlighting the polysemic nature of this particle.

4. Results and discussion

There were 814 instances of *well* as an adverb in P-ACTRES. The structural analysis of all of these instances revealed the results shown in Table 2 below:

Table 2: Classification of all cases of well in P-ACTRES.

STRUCTURE	CASES	PERCENTAGE
as well (as)	369	45.3%
Well as main adverb /adjective	184	22.6%
Well as modifier of adjective	124	15.2%
Well as discourse marker	94	11.5%
Well as modifier of adverb	43	5.2%
TOTAL	**814**	**100%**

The vast majority of cases (45%) of the English form *well* occurred as part of the conjunction *as well (as)*. In 22.6% of the cases, *well* was the main adverb or adjective in its clause, in 15% of cases it was a modifier of an adjective, and in 5% of cases it was used as a modifier of another adverb. Finally, 11.5% of cases, a total of 94 cases, were labelled in the analysis as pragmatic uses of *well* as a discourse marker. This study will focus on the translations into Spanish of these uses of the English form *well*.

These 94 occurrences were further classified into two subgroups: on the one hand, the instances where *well* appears alone in sentence initial position, followed by a comma and in dialogue predominantly (57 cases); on the other hand, the cases of *well* where the adverb was so highly grammaticalized that it cannot be considered to carry out another function but to reinforce the epistemic modality indicated by the modal verb it follows (37 cases). This second group includes cases where *well* is not grammatically peripheral or marginal, but rather has become fused with the rest of the sentence and are therefore to be considered on the boundary between proper adverbs and discourse markers. I will claim in this paper that pragmatic markers tend to form collocations and patterns in text, precisely because of this textual function.

4.1. Well *as a single pragmatic marker*

The 57 instances included 9 cases from the non-fiction subcorpus, 1 case from the press subcorpus and the remaining 47 cases (82% of the total) appeared in the fiction corpus imitating spontaneous conversation. The analysis of the 57 instances of *well* as a single-word discourse marker and the corresponding translations into Spanish revealed the list of 13 different translations shown in Table 3 below, with their corresponding frequency of occurrence:

Table 3: Translational options of single discourse marker well in Spanish.

WELL AS A SINGLE ITEM	CASES
Bueno	30 – 52.6%
Pues	5 – 8.7%
Bien	5 – 8.7%
Omission	4 – 7.01%
En fin	3 – 5.2%
Vaya	2 – 3.5%
Pues bien	2 – 3.5%
Entonces	1 – 1.7%
En realidad	1– 1.7%
De acuerdo	1– 1.7%
Efectivamente	1– 1.7%
Claro está	1– 1.7%
Vamos a ver	1– 1.7%
TOTAL	**57**

The results show that there is a wide range of different translations, something which confirms previous studies on the translations of discourse markers (Aijmer & Simon-Vandenbergen, 2003). One form in particular, *bueno*, accounts for over half the cases (2), followed by *pues* (3) and *bien* (4), with less than 10% of occurrences each.

(2) *'Well*, it was pretty ghastly, by all accounts. (FIK1E.s320)

- *Bueno, la cosa fue un auténtico horror, al decir de todo el mundo*. (FIK1S.s323)

The use of the Spanish adjective *bueno*, the semantico-functional equivalent of *well*, shows that this particular adjective indicating positive evaluation has acquired a similar status in Spanish as a discourse marker.

(3) *'Well*, hang onto them. (FCJ1E.s612)

- *Pues no los dejes escapar.* (FCJ1S.s597)

In contrast, the use of *pues* adds a clearly causal meaning that is not explicit in the English original text, but may be inferred from the context.

(4) "Oh, yeah, *well* I hope to hear 'em some of these days. (EDB1E.s7)

- *Ah, bien.* (EDB1S.s10) *Espero escucharlas un día de estos*. (EDB1S.s11)

Bien is the Spanish adverb that most closely represents the representative meaning of *well* as a manner adverb, but, as mentioned above, it is the corresponding adjectival form *bueno*, not the adverbial form, which is most commonly grammaticalized into a pragmatic marker.

Omissions occur in only 7% of cases, and the remaining options occur so infrequently that further data would be needed to get a clearer picture. The last two cases in the list (see example 5) are what previous authors have called *routines*, i.e., short fixed phrases or clauses (*claro está, vamos a ver*) with similar pragmatic and textual meanings in Spanish to the ones encoded by the English form *well*.

(5) *'Well*, did she or didn't she? (FWM1E.s333)

Vamos a ver, ¿la cerraba o no la cerraba? (FWM1S.s329)

4.2. Well *as a pragmatic marker used after a modal verb.*

The second group of uses of *well* as a discourse marker include those cases where the particle closely follows an epistemic modal and acts as an idiomatic reinforcement of that particular modal. What we find in these cases are clear collacational patterns, although the number of instances is so low that it is difficult to draw relevant conclusions. Out of the 37 cases, the modal *may* occurred in 18 of cases, *could* in 10 and *might* in 9 cases. As for the registers, only 3 cases were found in fictional texts and all the others either in essays or press texts.

The 9 different translations into Spanish found in our corpus are listed in Table 4:

Table 4: Translational options of well after a modal verb.

***WELL* AFTER A MODAL VERB**	**CASES**
Omission	19 - 51.3%
Muy bien	5 – 13.5%
Perfectamente	4 – 10.8%
Bien	3 – 8.1%
Muy posible	2 – 5.4%
Igualmente	1 – 2.7%
Con toda seguridad	1 – 2.7%
Modulation	1– 2.7%
Acaso	1– 2.7%
TOTAL	**37**

In this case it is remarkable that the translational option most frequently taken is the actual omission of the pragmatic marker. In fact, over half the cases were not translated (6). This ties in with previous studies that show that the elusive nature of the meanings expressed by pragmatic markers makes them difficult to convey in another language and easy to omit in translations, as the propositional content is not affected at all.

(6) This might *well* have been fatal in a real operation... (ELAR1E.s240)

En una acción real el desenlace hubiese resultado fatal [...] (ELAR1S.s224)

When *well* is translated, it is generally and adverb that is chosen to express this meaning: *muy bien* (13% of cases), *perfectamente* (10%) or *bien* (8%), as in the examples below.

(7) If American automakers do not innovate quickly enough, in another decade you may *well* be driving a superefficient Chinese-made car. (RLA1E.s187)

Si los fabricantes de automóviles de otros países no innovan rápidamente, dentro de una década pudiera muy bien ocurrir que el lector conduzca un coche de bajísimo consumo de manufactura china. (RLA1S.s194)

(8) Men in British uniform acted suspiciously and may *well* have been spies. (EHJ1E.s2-19)

Hombres vestidos con uniformes del Ejército británico actuaban de forma sospechosa por lo que perfectamente podrían haber sido espías. (EHJ1S.s224)

(9) Because many are unwieldy and meticulously fashioned, they may *well* have been used to impress and woo. (EHF1E.s354)

Dado que muchas de ellas eran difíciles de manejar y sin embargo habían sido talladas meticulosamente, bien pudieron utilizarse para impresionar y cortejar al amante. (EHF1S.s342)

The first three translational options refer to the propositional content of the English particle *well*, but there are also two cases of *muy posible*, a combination that suggests the epistemic nature of the meaning conveyed here:

(10) If Freya was a machine, and the Germans were using it to defend their borders, it might *well* be in Denmark. (FFK2E.s691)

Si Freya era una máquina, y los alemanes la estaban utilizando para defender sus fronteras, era muy posible que se encontrara en Dinamarca. (FFK2S.s695)

The remaining options found in the corpus occur only once.

5. Conclusions

This paper has studied the translations into Spanish of the pragmatic marker *well* as it appears in the P-ACTRES parallel corpus. All the instance of this form were extracted and analyzed. The cases were divided into two clearly differentiated groups: single discourse markers in sentence-initial position in dialogue, and collocational combinations with modal verbs with an epistemic meaning of reinforcement.

In the case of single discourse markers we find a large number of different translational options, although over half the cases corresponded to one single adjective, *bueno*, the functional equivalent of *good*. Other minor options included *pues* or *bien*, whereas only 7% of cases were actually omitted in translations.

As for the patterns of *well* following modals, there were also many possible options, but half the cases were actually omitted in Spanish. The most frequent options were the functional equivalents of *well bien, muy bien* and *perfectamente*.

Different uses of a polyfunctional item such as *well* provide very different translational patterns. Pragmatic markers in dialogue tend to be viewed by translators as important part of the discourse and a translation is provided, whereas the case of the collocational patterns with modals is not viewed as so essential, so *well* is mostly omitted in these contexts.

References

Aijmer, K. (2002). *English Discourse Particles. Evidence from a Corpus*. Amsterdam/ Philadelphia: John Benjamins.

Aijmer, K., & Simon-Vandenbergen, A.M. (2003). The discourse particle *well* and its equivalents in Swedish and Dutch. *Linguistics* 41(6), 1123-1161.

Andersen, G., & Fretheim, T. (Eds.). (2000). *Pragmatic Markers and Propositional Attitude*. Amsterdam/Philadelphia: John Benjamins.

Bazzanella, C., & Morra, L. (2000). Discourse Markers and the Indeterminacy of Translation. In I. Korzen & C. Marello (Eds.), *Argomenti per una linguistica della traduzione. On linguistic aspects of translation. Notes pour une linguistique de la traduction*. (pp. 149-157). Alessandria: Edizioni dell'Orso.

Bolinger, D. (1989). *Intonation and its Uses. Melody in Grammar and Discourse*. London: Arnold.

Carlson, L. (1984). *'Well' in Dialogue Games: A Discourse Analysis of the Interjection 'well' in Idealized Conversation*. Amsterdam/Philadelphia: John Benjamins

Chafe, W. (1986). Evidentiality in English Conversation and Academic Writing. In W. Chafe & J. Nichols (Eds.) *Evidentiality. The Linguistic Coding of Epistemology*. (pp. 261-272). Norwood, NJ: Ablex.

Fischer, K. (2000). *From Cognitive Semantics to Lexical Pragmatics. The Functional Polysemy of Discourse Particles*. Berlin: Mouton de Gruyter.

Fraser, B. (1990). An Approach to Discourse Markers. *Journal of Pragmatics* 14(3), 383-395.

Halliday, M.A.K., & Hasan, R. (1976). *Cohesion in English*. London: Routledge.

Johansson, S. (2006). How well can "well" be translated? On the English discourse particle "well" and its correspondences in Norwegian and German. In K. Aijmer &

A.M. Simon-Vandenbergen (Eds.), *Pragmatic Markers in Contrast*. (pp. 115-137). Amsterdam: Elsevier.

Jucker, A.H., & Ziv, Y. (Eds.). (1998). *Discourse Markers. Description and Theory*. Amsterdam/Philadelphia: John Benjamins.

Lenk, U. (1998). *Marking Discourse Coherence. Functions of Discourse Markers in Spoken English*. Gunter Narr: Tübingen.

Schiffrin, D. (1987). *Discourse Markers*. Cambridge: CUP.

Schourup, L. (2001). Rethinking *well*. *Journal of Pragmatics* 33(7), 1025-1060.

Svartvik, J. (1980). *Well* in Conversation. In S. Greenbaum, G. Leech & J. Svartvik (Eds.), *Studies in English Linguistics for Randolph Quirk*. (pp. 167-177). London: Longman.

Translating Research Articles from Spanish into English: A Corpus-based Comparative Analysis of the Genre[115]

Cristina Toledo Báez
Universidad de Málaga

Abstract: This paper aims to prove whether research articles on the domain of Information and Technology Law published in Spanish share the Introduction-Method-Results-Discussion (IMRD) structure used in most articles written in English. More specifically, we focus on the section 'introduction' in order to study whether most articles have either the Create a Research Space (CARS) model (Swales, 1990) or the Open a Research Option (OARO) model (Swales, 2004). In previous studies with small corpora (Toledo Báez, 2009 and 2010), the results showed that the introductions CARS are much more frequent in English than in Spanish and the OARO structure is the most common in the Romance language. However, we need to prove correct this hypothesis with the intratextual comparative analysis of our bilingual, specialized, virtual, and representative comparable corpus consisting of a collection of 280 research articles on electronic commerce, 140 in Spanish and 140 in English.

Keywords: research articles, bilingual corpora, contrastive analysis, CARS, OARO, translation.

Resumen: El objetivo principal de este trabajo es comprobar si los artículos de investigación del Derecho de las Tecnologías de la Información y las Comunicaciones publicados en lengua española presentan la típica estructura anglosajona Introducción-Método-Resultados-Discusión (IMRD). Nos centraremos en las introducciones de los artículos en aras de analizar si la mayor parte de los mismos presenta el modelo Create a Research Space (CARS) (Swales, 1990) o el modelo Open a Research Option (OARO) (Swales, 2004). En estudios anteriores realizados con corpus de menor tamaño (Toledo Báez, 2009 y 2010) los resultados obtenidos mostraron que las introducciones con CARS son más frecuentes en lengua inglesa que en lengua española y que el modelo OARO es más utilizado en español que en inglés. Estimamos conveniente corroborar dicha hipótesis mediante el análisis contrastivo de un corpus bilingüe, especializado y representativo compuesto por 240 artículos de investigación, 140 en español y 140 en inglés.

Palabras clave: artículos de investigación, corpus bilingües, análisis contrastivo, CARS, OARO, traducción.

115 The research reported on this paper has been carried out in the framework of R&D national project *Ecosistema: espacio único de sistemas de información ontológica y tesauros sobre el medio ambiente. Ecoturismo* (Reference no. FFI2008-06080-C03-03).

1. Introduction

Translating research articles from any language into English is of paramount importance in the scientific community. However, before translating, it is necessary to ascertain the superstructure of both source text and target text according to the genre conventions in each language.

This article aims to prove whether research articles on the domain of Information and Technology Law published in Spanish share the Introduction-Method-Results-Discussion (IMRD) structure used in most articles written in English. More specifically, we focus on the section 'introduction' in order to study whether most articles have either the *Create a Research Space* (CARS) model (Swales, 1990) or the *Open a Research Option* (OARO) model (Swales, 2004). In previous studies with small corpora (Toledo Báez, 2009 and 2010), the results showed that the introductions CARS are much more frequent in English than in Spanish and the OARO structure is the most common in the Romance language. However, we need to prove this hypothesis with the intratextual comparative analysis of our bilingual, specialized, virtual, and representative (Corpas Pastor and Seghiri Domínguez, 2010/in press) comparable corpus consisting of a collection of 280 research articles on electronic commerce, 140 in Spanish and 140 in English.

2. Superstructure of the introduction section in research articles

Swales' model (1990: 140) called *Create a Research Space* (CARS) is the most accepted study about the text organization of the introduction in research articles. This CARS model encompasses three moves related to three aims: first, the need to re-establish in the eyes of the discourse community the significance of the research field itself; second, the need to 'situate' the actual research in terms of that significance; third, the need to show how this niche will be occupied and defended.

Table 1. A CARS model for article introductions (Swales, 1990: 141)

Move 1: Establishing a territory
Step 1: Claiming centrality
Step 2: Making topic generalization(s)
Step 3: Reviewing items of previous research
Move 2: Establishing a niche
Step 1A: Counter-claiming
Step 1B: Indicating a gap
Step 1C: Question-raising
Step 1D: Continuing a tradition
Move 3: Occupying the niche
Step 1A: Outlining purposes
Step 1B: Announcing present research
Step 2: Announcing principal findings
Step 3: Indicating RA structure

Even though this model was highly successful, Swales (2004: 244) proposed a new model called *Open a Research Option* (OARO), which is based on the work by various authors from different disciplines, i.e., Mauranen (1993) from Economics, Ahmad (1997) from Science, Clyne (1985) from Sociology, Fredrickson and Swales (1994) from Linguistics and Burgess (2002) from Languages and Literature. The OARO model is the following:

Table 2. The OARO (Open a Research Option) model

0.	*[Attracting the Readership]* Optional opening
1.	*Establishing Credibility* (one or more of the following four):
	a. Sharing background knowledge b. Justifying need for research per se c. Presenting interesting thoughts d. Introducing general goal
2.	*Offering a Line of Enquiry*
	a. Discussing current problems b. Expressing interest in an emerging topic
3.	*Introducing the Topic*

The main differences between the CARS and the OARO models are described on this table:

Table 3. Schematic comparison of two modules (Swales, 2004: 245)

OARO (*open a research option*)	**CARS (*create a research space*)**
Nontantagonistic stance	More antagonistic
Mostly softer fields	Mostly harder fields
Small discourse communities	Large discourse communities
Mostly non-Anglophone cultures	Mostly Anglophone cultures
Unsectioned/unconventional	Conventional (IMRD)

3. Contrastive study of introduction sections in research articles in Spanish and English

In this section we will focus on the introduction section in order to test whether most introduction sections from our corpus have either the CARS or the OARO models. In previous studies with small corpora (Toledo Báez, 2009 and 2010), the results showed that the introductions CARS are much more frequent in English than in Spanish and the OARO structure is the most common in the Romance language. However, we need to prove this hypothesis with the comparative analysis of our bilingual and comparable corpus consisting of a collection of 280 research articles on electronic commerce, 140 in

Spanish and 140 in English. All the articles belong to the Information Technology Law, which has its own discourse and its main feature is the combination of Law and Computing. We have denominated this discourse with a neologism according to its origin: *legal-technological discourse* (Toledo Báez, 2009). In the case of the Spanish articles, they all have been extracted from the specialized journal *Revista de Contratación Electrónica*[116]; however, the English articles have two different sources: the *Journal of Information, Law and Technology*[117] and the *International Journal of Law and Information Technology*[118]. We will now focus on the features of the introduction sections in each language.

3.1. Introduction sections in Spanish research articles

The comparative analysis of the 140 introduction sections in Spanish has shown that we find examples of OARO structures but also of CARS model even though we are working in a soft field. We will choose one example of each type of structure from two articles[119]:

Table 4. A CARS introduction section in a Spanish article

Moves and steps	**Introduction section**
Move 1: Establishing a territory Step 1: Claiming centrality	La denominación derecho informático es la forma de designar a aquella rama del Derecho encargada del estudio de los contratos cuyo objeto está constituido por bienes y servicios informáticos.
Step 2: Making topic generalization(s)	Ésta es la opinión mayoritaria de la doctrina a la hora de encontrar una referencia al derecho informático. Esos bienes y servicios informáticos, en la práctica, aparecen interrelacionados constituyendo un objeto único de contrato.
Step 3: Reviewing items of previous research	Desde una perspectiva histórica, el derecho y la informática son dos conceptos que aparentemente, sobre todo años atrás, distaban mucho el uno del otro.

116 <http://vlex.com/source/rce-59>
117 <http://www2.warwick.ac.uk/fac/soc/law/elj/jilt/>
118 <http://ijlit.oxfordjournals.org/>
119 The two articles are the following: Viguera Revuelta, R. (2008). Los contratos informáticos. *Revista de la contratación electrónica*, *97*, 61-92. Perales Viscasillas, M. P. (2006). Publicidad y formación del contrato: convención de UNCITRAL sobre la utilización de las comunicaciones electrónicas en los contratos internacionales. *Revista de la contratación electrónica, 72*, 59-76.

Move 2: Establishing a niche Step 1A: Counter-claiming	Sin embargo, con el transcurso del tiempo ambos conceptos fueron adquiriendo notas comunes hasta el punto de surgir el concepto derecho informático.
Step 1B: Indicating a gap	No obstante, las relaciones sociales y económicas generadas como consecuencia de las modernas tecnologías de la información y las comunicaciones han generado diversos problemas […]
Step 1C: Question-raising	[…] la necesidad de una regulación jurídica de los derechos y obligaciones como consecuencia de la contratación de estos bienes, las responsabilidades derivadas de la transferencia electrónica de fondos o de datos, la validez probatoria de los documentos generados por medios electrónicos o informáticos, […]
Step 1D: Continuing a tradition	Por todo ello el derecho informático está llamado a ir dando pasos para resolver los conflictos derivados de esta relación.
Move 3: Occupying the niche Step 1A: Outlining purposes Step 1B: Announcing present research	En este punto es donde se plantea si la rama derecho informático podría llegar a tener entidad suficiente como para constituir, por sí misma, una rama del Ordenamiento.
Step 2: Announcing principal findings	En mi opinión[116], siguiendo en este punto al profesor Hernández Gil, lo que de verdad se plantea no es que el Ordenamiento Jurídico vaya a ordenar nuevas realidades, sino que el Derecho mismo va a experimentar, en cuanto objeto de conocimiento, un cambio, una mutación, derivada de un modo distinto de ser elaborado, tratado y conocido.

Let us compare it with an OARO structure:

120 "En mi opinión" is the only subjective element that would not be suitable for a harder field, albeit it is quite common in Arts and Social Sciences.

Table 5. An OARO introduction section in a Spanish article

Moves and steps	**Introduction section**
Attracting the Readership	El 23 de noviembre de 2005 la Asamblea General de las Naciones Unidas adoptó la Convención sobre la utilización de las comunicaciones electrónicas en los contratos internacionales[1].
Establishing Credibility Sharing background knowledge	Convención que es fruto del trabajo realizado por la Comisión de las Naciones Unidas para el Derecho Mercantil Internacional (CNUDMI/UNCITRAL) que así aprobó el proyecto de Convención en su 38º período de sesiones celebrado en Viena del 4 al 15 de julio de 2005[2].
Justifying need for research per se	Pese a que normalmente las Convenciones requieren de un tiempo relativamente dilatado para su adhesión internacional, lo cierto es que la Convención -electrónica- está recibiendo un amplio consenso.
Presenting interesting thoughts	La recién aprobada Convención no regula exhaustivamente todas y cada una de las cuestiones que pueden surgir en el marco de la contratación electrónica.
Introducing general goal	Se trata de regular por medio de una Convención o Tratado internacional aquellas cuestiones que se consideran importantes y trascendentes […]
Offering a Line of Enquiry Discussing current problems	Asimismo, conviene resaltar el paralelismo de la Convención con otros textos internacionales […] La comparación entre ambos nos permite concluir que se presentan como proyectos complementarios, por lo que nada impide que las partes acuerden la aplicación de los E-terms 2004, siendo la ley aplicable la Convención.
Expressing interest in an emerging topic	En este sentido, el Grupo de Trabajo sobre comercio electrónico observó las diferencias entre la labor realizada por la CCI, que revestía la forma de asesoramiento en materia contractual a particulares, y su propia labor en lo que atañe a la Convención, que es de carácter legislativo.
Introducing the Topic	En definitiva, se concluía que no existían contradicciones sustanciales entre los dos instrumentos[4], tal y como veremos a continuación.

Comparing manually the 140 introduction sections, we have checked that 53 out of 140 Spanish introductions share the CARS structure and 87 the OARO structure. Consequently, it is proved correct the hypothesis by Toledo Báez (2009 and 2010) that OARO structure is the most common in the Romance language. Furthermore, we can also add that this structure is the most common in Law and Legal Sciences, which is also proven true in Toledo Báez (2009 and 2010).

3.2. *Introduction sections in English research articles*

Regarding articles in English, we also see that both CARS and OARO structures are used. We will choose two articles[121] in order to illustrate the two structures.

Table 6. A CARS introduction section in an English article

Moves and steps	**Introducción**
Move 1: Establishing a territory Step 1: Claiming centrality	In September 2003, a study1 (henceforth, 'the Dumortier study') requested by the European Commission on the legal and practical issues concerning the implementation of the Electronic Signatures Directive2 (henceforth, 'the Directive') was completed.
Step 2: Making topic generalization(s)	The study team under the supervision of Dumortier discovered that although the broad lines of the Directive have been respected in its implementation by the Member States, a number of issues have nevertheless been identified as problematic.
Step 3: Reviewing items of previous research	These problems can mainly be attributed to a misinterpretation of the Directive's wording […] The conclusion drawn by the study team was the following: […]
Move 2: Establishing a niche Step 1A: Counter-claiming	Given the above-mentioned situation, this article is aimed at providing a contribution to the electronic signatures debate in Europe by offering an in-depth analysis and re-interpretation of the system of liability for Certification Service Providers3.
Step 1B: Indicating a gap	[…] Normally, there is no contractual relationship between CSPs and third parties (e.g., the recipient of a digitally signed message or another CSP), who rely on the validity of certificates.

121 The two articles are the following: Balboni, P. (2004). Liability of certification service providers towards relying parties and the need for a clear system to enhance the level of trust in electronic communication. *Information and Communications Technology Law, 13(3)*, 211-242. Van Der Hof, S. (2003). European Conflict Rules Concerning International Online Consumer Contracts. *Information and Communications Technology Law, 12(2)*, 165-178.

Step 1C: Question-raising Step 1D: Continuing a tradition	Therefore, there is a strong need for clear liability rules in the relationship between CSPs and third parties, while the liability issues between CSPs and users basically can be regulated by the contractual agreement.
Move 3: Occupying the niche Step 1A: Outlining purposes	The aim of this article is to underline the importance of a clear and effective system of liability for CSPs in the process of building trust in electronic communication […]
Step 1B: Announcing present research	The two main points as regards the re-interpretation of the CSPs liability system are the following.
Step 2: Announcing principal findings	First, the concept of 'relying party' has to be extended to the signatory. […] Moreover, after balancing the pros and cons of the system recommended by the European legislator, another practicable system for the liability of CSPs will be proposed.

Let us compare it with an OARO structure:

Table 7. An OARO introduction section in an English article

Moves and steps	**Introducción**
Attracting the Readership	In January 2003, the European Commission launched a Green Paper on the conversion of the Rome Convention of 1980 on the law applicable to contractual obligations into a Community Instrument and its modernisation.2 [..]
Establishing Credibility Unidad 1: Descripción de antecedentes comunes	At present, the effect of the special conflict rules concerning consumer contracts under the Rome Convention of 1980 on online consumer contracts is not completely clear.
Justifying need for research per se	Thus, a revision has to at least provide legal certainty in this respect; whether the special conflict rules concerning consumer contracts should explicitly extend to online consumer contracts […]

Presenting interesting thoughts	In the European Union, Community law provide special conflict rules for certain consumer contracts that aim at protecting consumers—being considered the weaker party—against socially and/or economically stronger businesses and professionals.
Introducing general goal	Two situations need to be distinguished in this respect […]
Offering a Line of Enquiry Discussing current problems	first, the situation where parties have agreed upon a choice of law clause in, for example, the general terms and conditions accompanying the contract (see Section 3); and, second, the situation where such a clause does not exist or is invalid (see Section 4).8
Expressing interest in an emerging topic	This article concludes with a summary of the main points raised (Section 5).
Introducing the Topic	This article deals with existing special conflict rules for consumer contracts in Europe, which will be assessed in the light of electronic commerce.

Comparing manually the 140 introduction sections, we have checked that 74 out of 140 articles have the CARS structure and 66 share the OARO structure. Even though the difference is lower than in Spanish articles, it is also proven correct Toledo Báez's hypothesis: in English and in Anglo-Saxon cultures, CARS structure is more frequent than the OARO structure.

4. Conclusions

In this paper we have confirmed the hypothesis established in Toledo Báez (2009 and 2010) with a corpus of 280 research articles on electronic commerce, 140 in Spanish and 140 in English: CARS introductions are more frequent in English than in Spanish, albeit the difference is quite low, just 8%. However, the OARO structure is the most common in the Romance language and the difference is 34%, higher than the difference with CARS introductions. Consequently, the hypothesis mentioned above has been proven correct but we need more studies with more complex and representative corpora in order to extrapolate the results. Analyzing the results with statistical methods such as the student's T-test or the chi-square test would be of interest.

Nevertheless, it is important to note that these differences regarding the superstructure of the introduction section in research articles have an impact on the translation of research articles from Spanish into English. As a result, two questions rise here: Is it advisable to keep the superstructure found in the source language or, on the contrary, is it preferable

to adapt the superstructure of the source language to the most common superstructure in the target language? In this context, we propose to keep the superstructure of the source language when translating research articles because that superstructure belongs, on one hand, to a discourse community, either Spanish or English, and, on the other hand, to a specific domain, in this case Information and Technology Law.

5. References

Ahmad, U. K. (1997). Research Article Introductions in Malay: Rhetoric in an Emerging Research Community. En A. Duszak (Ed.), *Culture and Styles of Academic Discourse* (pp. 273-303). Berlin: Mouton de Gruyter.

Burgess, S. (2002). Packed Houses and Intimate Gatherings: Audience and Rhetorical Structure. En J. Flowerdew (Ed.), *Academic Discourse* (pp. 197-215). Harlow: Longman.

Clyne, M. (1985). *Language and Society in the German-Speaking Countries*. Cambridge: Cambridge University Press.

Corpas Pastor, G. and Seghiri Domínguez, M. (2010/in press). *El uso de corpus para la traducción de textos especializados: la cuestión de la representatividad*. Frankfurt am Main: Peter Lang.

Fredrickson, K. M. and Swales, J. M. (1994). Competition and Discourse Community: Introductions from *Nysvenka Studier*. En B. L. Gunnarsson, P. Linell and B. Nordberg (Eds.), *Text and Talk in Professional Contexts* (pp. 9-22). Uppasala: ASLA.

Mauranen, A. (1993). *Cultural Differences in Academic Rhetoric: A Textlinguistic Study*. Frankfort am Main: Peter Lang.

Swales, J. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Swales, J. (2004). *Research Genres: Explorations and Applications*. Cambridge and New York: Cambridge University Press.

Toledo Báez, M. C. (2009). *El resumen automático como recurso documental para la traducción de artículos de investigación del ámbito jurídico-tecnológico (español-inglés-francés)*. Málaga: Servicio de Publicaciones de la Universidad de Málaga.

Toledo Báez, M. C. (2010). *El resumen automático y la evaluación de traducciones en el contexto de la traducción especializada*. Frankfurt am Main: Peter Lang.

Variación lingüística y corpus

La reforma feminista del español en los anuncios de prensa. Un estudio basado en corpus.[122]

Merceces Bengoechea y José Simón
Universidad de Alcalá

Resumen

La publicidad abarca una serie de textos e imágenes altamente llamativos y con enorme visibilidad que constituyen un paisaje cultural con características propias. El presente trabajo se centra en el estudio del paisaje publicitario de 2007. El principal objetivo de la investigación ha sido averiguar hasta qué punto se ha utilizado el lenguaje no sexista en los anuncios publicados durante ese tiempo. Un elemento esencial en nuestro estudio ha sido el corpus creado con nuestras muestras. En el presente artículo expondremos los postulados de partida, la metodología de trabajo y los resultados de la primera fase de nuestro estudio, correspondiente a los anuncios insertados en el País *durante el mes de octubre de 2007.*

Palabras clave

lingüística de corpus; paisaje publicitario; feminización del lenguaje; políticas lingüísticas.

Abstract

Advertising covers a range of texts and images highly appealing and with great visibility which constitute a cultural landscape with characteristics of its own. This paper focuses on the advertising landscape corresponding to 2007. The main research objective was to determine to what extent non-sexist language was used in ads at that time. An essential element in our study was the corpus we created with our samples. In this article we will discuss our basic tenets, working methodology and the results of the first phase of our study, for the ads appeared in el País *during the month of October 2007.*

Keywords

corpus linguistics; adscape; language feminization; language policies.

122 Agradecemos al Instituto de la Mujer (I+D+I 37/06), así como al Ministerio de Ciencia e Innovación (Proyecto FEM2009-10976) el apoyo y la financiación que nos han prestado, sin la cual no habríamos podido llevar a cabo el presente estudio. Así mismo, queremos agradecer a Verónica González Araujo su inestimable ayuda en la captación de datos para nuestro corpus.

1. Introducción

Veinte años después de que se publicaran las primeras recomendaciones de uso no sexista del español (Council of Europe 1986; Departamento de la Dona 1987) acometimos un proyecto cuyo objetivo principal era estudiar los efectos que estas recomendaciones han tenido en el panorama lingüístico de nuestro país. Como parte de este proyecto, junto a otros estudios (Bengoechea 2008, 2009, 2011; Bengoechea y Simón 2010, 2011, de próxima aparición), decidimos investigar hasta qué punto el lenguaje de la publicidad se hacía eco de estas recomendaciones.

Creemos que el éxito o fracaso de las políticas lingüísticas no sexistas solo se podrá medir de forma escalonada, a lo largo de los años, mediante el estudio de los cambios observables en las diversas manifestaciones del discurso. En consecuencia, diseñamos un estudio longitudinal en el que iríamos efectuando calas a intervalos de tres años: la primera, correspondiente al año 2007, nos proporcionaría una visión del panorama tras dos décadas de recomendaciones; la segunda, correspondiente al 2010, los resultados estimables 3 años después de promulgarse la ley de igualdad 3/2007, que vino a convertir en norma lo que antes eran meras recomendaciones.

Dado que la publicidad nos llega por muy diversos conductos, a la par que tiende a ser muy reiterativa, restringimos nuestra población a los anuncios insertados en el diario *el País* durante el mes de octubre de 2007. La elección responde a que en aquellas fechas éste era el diario con mayor tirada nacional (y hoy día continúa siéndolo). Previamente efectuamos un sondeo, durante una semana, que nos permitió comprobar que la publicidad que buscábamos, la de carácter general, se repetía en el resto de diarios de alcance nacional publicados en Madrid, por lo que, a efectos de nuestro corpus, entendimos que incluir otros diarios en nuestra muestra no nos iba a aportar mayor diversidad.

Simultáneamente, se recogió toda la publicidad recibida por vía postal en un domicilio madrileño de clase media. Con estas muestras se compiló un primer sub-corpus cuyo estudio sirvió como piloto para el estudio de la publicidad en prensa y cuyos resultados aparecen publicados en Bengoechea y Simón (2010).

2. Presupuestos

En nuestro estudio, adoptamos el concepto de paisaje lingüístico (*linguistic landscape*) que puede definirse como lo han hecho Landry and Bourhis:

> The language of public road signs, advertising billboards, street names, place names, commercial shop signs, and public signs on government buildings combines to form the linguistic landscape of a given territory, region, or urban agglomeration. The linguistic landscape of a territory can serve two basic functions: an informational function and a symbolic function (1997: 25).

La prensa nacional es uno de los objetos que vienen a componer el paisaje lingüístico de un determinado país. Los anuncios publicados en ella constituirían signos del espacio público, cada uno de los cuales se pueden tomar como una unidad de análisis, en el sentido

definido por Backhaus: "any piece of text within a spatially definable frame" (2006: 55). También constituye un signo del espacio público la publicidad que nos llega diariamente al buzón de nuestro domicilio. Ambos tipos de textos forman parte del paisaje publicitario (*adscape*) que, a su vez, forma parte del paisaje lingüístico.

El concepto de paisaje lingüístico se convierte, pues, en una de las claves fundamentales a la hora de analizar las sociedades multilingües desde una perspectiva sociolingüística, en tanto que:

- Refleja el poder relativo y el estatus de las diversas variedades verbales en un contexto sociolingüístico específico.
- Ayuda a construir el contexto y el espacio sociolingüísticos.
- Revela cómo se posiciona ideológicamente una sociedad respecto a sus variedades verbales.
- Mide el estatus de que gozan en un momento las diversas variedades (en nuestro caso, las variedades sexista y no sexista del español).
- Mide la eficacia de las políticas lingüísticas llevadas a cabo por gobiernos e instituciones.

Puesto que ciertas áreas del Estado español son teóricamente monolingües, parecería carecer de sentido el estudio de su paisaje lingüístico para evaluar el estatus de las lenguas. Sin embargo, uno de nuestros supuestos de partida es que el español no sexista es una variedad del español[123], todavía minoritaria, pero en constante proceso de resituarse frente a la variedad hegemónica, el español de la Norma dictada por la Real Academia Española. En consecuencia, consideramos que podríamos llegar a conocer el estatus de que gozan en un momento dado ambas variedades estudiando el paisaje lingüístico español.

La variedad sexista del español se caracteriza fundamentalmente por los siguientes rasgos:

- Impera la perspectiva androcéntrica (ejemplos: "El pueblo saharaui ha levantado el campamento y se ha adentrado en el desierto con sus mujeres e hijos"; "toda la población debemos contribuir a combatir el cambio climático y despojarnos de la corbata y la chaqueta del traje").
- Las mujeres son denominadas profesionalmente mediante términos sexuados en masculino ("ella es arquitecto").
- Se impone el uso del masculino con valor genérico - Las expresiones sexuadas en masculino se constituyen en representativas del sexo femenino y del masculino ("los alumnos del colegio", por "las alumnas y los alumnos del colegio").

Para combatir estos usos secularmente arraigados en nuestra sociedad, gobiernos e instituciones han arbitrado una serie de recomendaciones que se cifran principalmente en:

- Feminizar la lengua y las profesiones: *jefa*, *interventora*, en vez de *jefe* e *interventor* para referirse a mujeres.

123 Algo que defendemos en Bengoechea (2011).

- Utilizar los términos masculinos sexuados para denotar únicamente a varones y renunciar al masculino pretendidamente genérico. Para ello se sugiere el uso de dobles formas, bien sea
    - Con duplicación: *las niñas y los niños*,
    - Con la barra: *los/as niños/as*,
    - Con el guión: *las, -os niñas, -os,* o
    - Con la arroba: *l@s niñ@s*,

  (en lugar de englobar a ambos sexos en "los niños").
- Neutralizar la lengua mediante sustantivos colectivos (*el profesorado* en lugar de "los profesores") o metonímicos (*el turismo*, en lugar de "los turistas").
- Utilizar recursos gramaticales, tales como construir las frases con el verbo en segunda persona del plural o del singular. Así, en lugar de decir "los interesados deben acudir", se sugiere "Si te interesa, debes acudir", "si os interesa, debéis acudir", o "si le(s) interesa, debe(n) acudir".

3. Corpus

Como ya hemos indicado, nuestro corpus de publicidad para el año 2007 se compone de dos subcorpus diferentes: Por una parte, la **publicidad postal** recibida en un domicilio madrileño de clase media durante el mes de octubre de 2007. El análisis de esta pequeña muestra (consta únicamente de 55 anuncios diferentes**)** nos sirvió de piloto para el diseño de una estructura de datos adecuada al objeto y los objetivos de nuestro estudio, así como para extraer unas primeras conclusiones. Por otra, 702 **anuncios de prensa** aparecidos en el País, durante el mismo período, que constituyen el sub-corpus que ahora presentamos.

Tanto en un caso como en el otro, muchos de los anuncios se repitieron a lo largo de esas semanas, por lo que, junto a la muestra se registró el número de repeticiones. El sub-corpus que nos ocupa recoge el total de 1538 anuncios (tokens) aparecidos en el País, correspondientes a 702 anuncios diferentes (tipos).

Teniendo en cuenta que el volumen de datos era reducido y dado que queríamos garantizar el acceso, no sólo al texto de los anuncios, sino a sus facsímiles, sin necesidad de recurrir a la copia impresa, optamos por compilar el corpus sobre una base de datos Access. Para corpus de grandes dimensiones es preferible recurrir a alguno de los estándares habituales, pero en este caso, basándonos en nuestra experiencia anterior con Nombra.en.red (Bengoechea y Simón 2008), nos inclinamos por una opción sencilla que nos reportaba importantes ventajas: por una parte, rapidez y flexibilidad de manejo y de programación; por otra, la posibilidad de realizar consultas en SQL, unida a la de automatizar y programar tareas y sucesos, bien mediante el lenguaje VBA que incorpora, bien mediante cualquiera de los integrados en Visual Studio. Todo ello nos permitió crear una sencilla, a la vez que robusta, interfaz de asistencia en las tareas de anotación y consulta, así como una serie de consultas SQL para tabular los datos y computar recuentos.

A estas ventajas se suma el hecho de que todos los elementos que integran el corpus (muestras, anotaciones, facsímiles, consultas y cálculos) hayan podido registrarse en un único archivo, lo que nos ha permitido trabajar en distintas ubicaciones sin necesidad de crear una aplicación en red a tal efecto.

No obstante, con el fin de facilitar la compatibilidad entre plataformas, la portabilidad y la consulta on-line con arquitectura cliente-servidor (que esperamos poder hacer pública pronto), se han exportado sendas réplicas a XML y servidores SQL (**sqlite** y **mySQL)** a las que se accede a través de cualquier navegador utilizando tecnología AJAX.

Dado que el elemento clave de un corpus son las muestras, con el fin de poder acceder a ellas de forma flexible, los anuncios se escanearon y se introdujeron en la base de datos en tres formatos diferentes:

- Imagen escaneada (jpg)
- Documento pdf
- Texto plano

Los dos primeros permiten representar en pantalla facsímiles de las muestras, para su anotación o revisión. El tercero facilita las tareas de anotación, así como la recuperación de información a través de consultas SQL. Al estar vinculadas las tres representaciones, se pueden efectuar búsquedas en el texto y acceder de forma automática (o mediante un simple clic) a los facsímiles.

Esta representación múltiple posibilita estudios desde otras perspectivas que podrán llevarse a cabo en el futuro, bien por el mismo equipo, bien por otras personas. Por ejemplo, sería muy interesante estudiar los roles asignados a hombres y mujeres a través de las imágenes que acompañan a los anuncios.

En la tabla principal de nuestra base de datos se registra toda la información correspondiente a cada una de las muestras: tanto los datos de identificación, como los relativos a las imágenes, pdf y el propio texto (una vez escaneado, reconocido mediante OCR y revisado). En una serie de tablas auxiliares se recogen las etiquetas utilizadas en la anotación de los distintos campos: descriptores, categorías gramaticales y recursos léxicos, tipo de anunciante, objetos anunciados, entre otros. Por último, en otra serie de tablas separadas se recogen las distintas anotaciones, conforme a taxonomías que mencionaremos más adelante. Con estas tablas, mediante una serie de relaciones y consultas, se han realizado varias combinaciones diferentes a las que se accede a través de diversos formularios (configurados como asistentes), diseñados para facilitar la anotación de los distintos rasgos, así como su posterior revisión y análisis. La figura 1 muestra el asistente de revisión en el que se pueden apreciar los distintos campos y los menús desplegables con categorías cerradas que han asistido en las tareas de anotación y revisión.

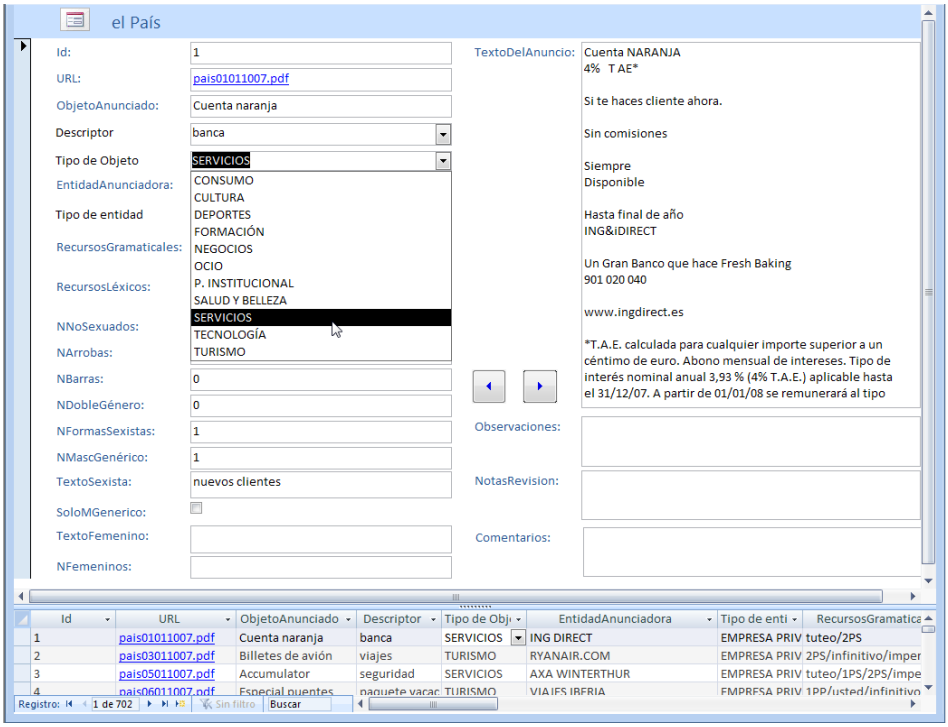


Figura 1. Asistente de revisión

El corpus, así configurado, registra un total de 702 anuncios diferentes (con sus repeticiones) de características muy diversas. El tamaño de las muestras, por lo que respecta al texto, es muy variado y va desde una única palabra (en varios anuncios) hasta las 820 de otro. El total de palabras en el corpus asciende a 71.079 y el tamaño medio de las muestras es de 101,25 palabras, con una desviación típica de 88,49.

4. Resultados

En nuestro corpus se han anotado diversos rasgos, entre ellos algunos accesorios, tales como:

- El tipo de productos ofertados (enseñanza, servicios, salud, belleza, hogar, consumo, etc.), lo que nos ha permitido comprobar que existe cierta correlación entre el producto, el sector de población al que va dirigido y el registro de lenguaje utilizado.
- El tipo de anunciante (empresa privada, ente público, ONG, particulares). También se ha puesto de manifiesto una cierta correlación entre el tipo de anunciante y la variedad lingüística empleada.
- El número de palabras en el texto.
- La sección o suplemento en que aparece el anuncio.

- La página en que se inserta.

No obstante, puesto que el objetivo era comprobar hasta qué punto los anuncios se hacían eco de la variedad no sexista de la lengua, tienen mayor interés aquellos rasgos que pueden ser interpretados como indicadores. El primero de éstos fue el carácter, sexuado o no, del anuncio. Partimos de las cinco categorías básicas que se muestran en la tabla 1.

Tabla 1. Categorías básicas

Categoría	Descripción
A – Texto neutro	Texto que no incluye referencias personales
B – Texto no sexista con referencias personales	Texto que incluye referencias personales, pero les da un tratamiento neutro (no sexuado)
C – Texto redactado en femenino	Texto que contiene expresiones sexuadas femeninas
D – Texto redactado en masculino y femenino	Texto que incluye referencias sexuadas en femenino y masculino utilizado como genérico.
E – Texto redactado en masculino	Texto que contiene expresiones masculinas sexuadas utilizadas como genéricas

Los resultados globales, conforme a este parámetro se resumen en la figura 2.

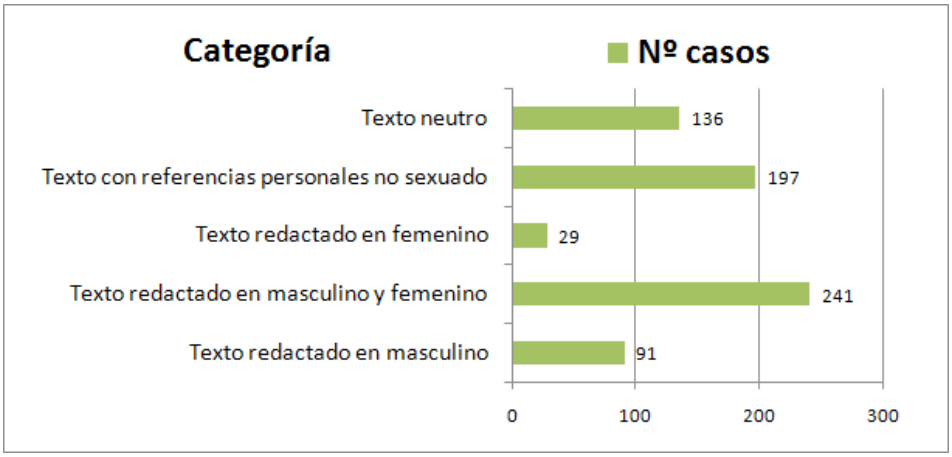


Figura 2. Distribución de frecuencias respecto al carácter sexuado del texto

A la vista de estos resultados, cabe destacar:

i. La baja incidencia de textos redactados en femenino.

ii. El relativo balance entre textos no sexuados (A+B = 333) y textos sexuados (C+D+E=361)

iii. El número de textos que incluyen el masculino en alguna medida (D+E) prácticamente iguala al de textos no sexuados.

Otro de los rasgos que se anotaron en el corpus y que se ha valorado como indicativo de la penetración de las recomendaciones antisexistas, es el uso de recursos gramaticales que "evitan" las referencias sexuadas. Los resultados se muestran en la figura 3.

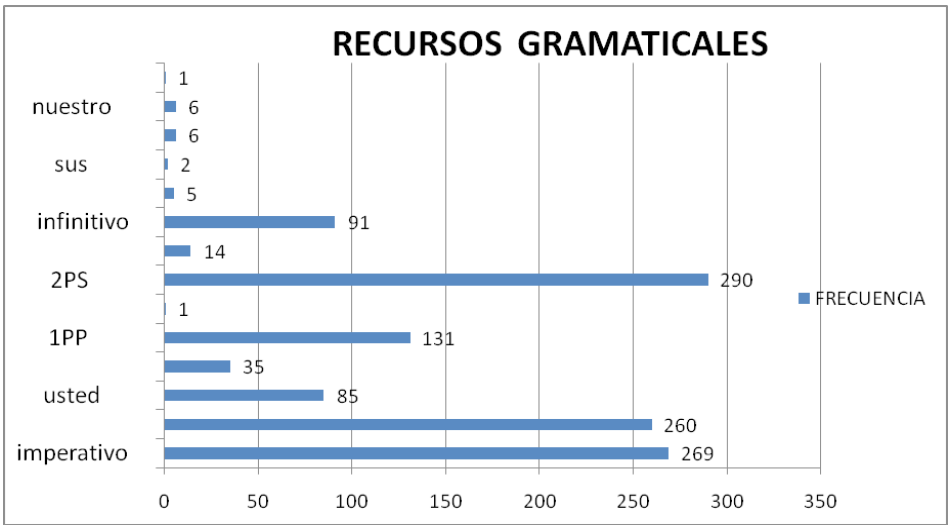


Figura 3. Distribución de frecuencias respecto a los recursos gramaticales empleados

Igualmente, se han anotado, por considerarlos altamente indicativos, los términos colectivos o metonímicos recomendados para evitar el uso de un masculino con valor genérico. Se han encontrado un total de 75. En la tabla 2 se muestran los 10 más frecuentes, junto al masculino genérico al que podrían estar sustituyendo. El símbolo Ø indica que no aparecen precedidos de determinante.

Tabla 2. Términos no sexuados más frecuentes

Término	Frecuencia	Masculino genérico equivalente
persona(s)	88	El hombre / los hombres (¿?)
Ø particulares	18	Los particulares
quien/quién/quienes	16	El / los que (¿?)
Ø profesionales	15	Los profesionales
Ø menores	21	Los menores / los niños
público	15	Los hombres (¿?)
nadie	7	Ninguno
Ø turista	7	El turista
Ø estudiante(s)	8	El/los estudiante/s
alguien	5	Uno
persona(s)	88	El hombre / los hombres (¿?)

5. Conclusiones

Nuestra investigación ha puesto de manifiesto que la variedad sexista del español es, con diferencia, la que impera en el paisaje publicitario lingüístico del periódico estudiado. En 2007 la variedad no sexista todavía se limitaba a una serie de productos (dirigidos a la mujer) o de ámbitos muy restringidos.

El rasgo que mejor caracteriza el paisaje que constituye nuestra muestra es la cantidad relativamente elevada de textos híbridos que mezclan ambas variedades: sexista y no sexista Este hecho, lejos de venir a señalar el escaso impacto de las políticas lingüísticas antisexistas, pondría de relieve que éstas han tenido un cierto efecto, pues de lo contrario habrían predominado los textos de la variedad sexista, el paradigma en todos los posibles paisajes lingüísticos hasta hace poco. Simultáneamente, la alternancia entre ambas variedades nos indica claramente que nos encontramos en un periodo de transición.

Cabe destacar, también, que cuanto más extenso es el texto, menor posibilidad parece existir de que se mantengan criterios no sexistas de forma consistente, lo que viene a reforzar la idea de que nos encontramos en un período de cambio y, ello a su vez, que el cambio se está produciendo.

En cualquier caso, la elección de elementos no sexistas en el mensaje publicitario tiene consecuencias en el paisaje lingüístico-mediático y en las impresiones de la gente que habita ese paisaje sobre la importancia cultural y simbólica de la variedad no sexista. Probablemente además desempeña la función simbólica de contribuir de la forma más directa a la identidad social, positiva o negativa, de las mujeres. Este es un aspecto en el que habrá que seguir indagando y nuestro corpus, una vez completado con las muestras de 2010, será una valiosa herramienta para poder llevar a cabo éste y otros estudios.

6. Referencias

Backhaus, P. (2006). Multilingualism in Tokyo: A look into the linguistic landscape. En Durk Gorter (ed.), *Linguistic Landscape: A New Approach to Multilingualism* (pp. 52-66). Clevedon: Multilingual Matters.

Bengoechea, M. (2008). Lo femenino en la lengua: sociedad, cambio y resistencia normativa. Estado de la cuestión. *Lenguaje y Textos*, 27, 37-68.

Bengoechea, M. (2009). Sexismo (y economía lingüística) en el lenguaje de las noticias: Inercias e incorporaciones igualitarias. En Pilar Fernández Martínez e Ignacio Blanco Alfonso (coords.), *Lengua y televisión* (pp. 32-62). Madrid: Fragua Comunicación.

Bengoechea, M. (2011). Non-sexist language policies of Spanish: An attempt bound to fail? *Current Issues in Language Planning,* 12 (1), 35-53.

Bengoechea, M. y Simón, J. (2008). Primeros resultados del estudio del Corpus Nombra. en.red. En *Actas del XXIV Congreso Internacional de AESLA: Aprendizaje de lenguas, uso del lenguaje y modelación cognitiva: perspectivas aplicadas entre disciplinas* (pp. 1439-1446). Madrid: UNED.

Bengoechea, M. y Simón, J. (2010). Gender Identity in Words for Professional Titles in Textbooks. En Rosa María Jiménez Catalán (ed.), *Gender perspectives on vocabulary in foreign/second language education* (pp. 188-211). London: Palgrave Macmillan.

Bengoechea, M. y Simón, J. (2011). El paisaje publicitario de 2007, tras veinte años de políticas lingüísticas antisexistas. En Fabienne Baider et Daniel Elmiger (eds.), *Intersexion: Languages romanes, langues et genres.* München: Lincom Europa. (en prensa).

Bengoechea, M. y Simón, J. (de próxima aparición). University students' attitude to non-sexist Spanish.

Landry, R. y Bourhis, R. Y. (1997). "Linguistic landscape and ethnoliguistic vitality: an empirical study". *Journal of Language and Social Psychology,* 16 (1), 23-49.

La lingüística forense y el uso de los corpus lingüísticos

Jordi Cicres
Universitat de Girona
Institut Universitari de Lingüística Aplicada – Universitat Pompeu Fabra

*En este artículo se discute acerca del uso de las distintas clases de corpus lingüísticos en la lingüística forense, tanto desde el punto de vista de la investigación como el de la práctica profesional. Además de la comparación entre el corpus de textos (orales o escritos) dubitados e indubitados, el perito debe de utilizar también corpus de referencia que le permitan decidir acerca de la rareza o idiosincrasia de las variables presentes en los corpus dubitado e indubitado. La definición de estos corpus de referencia es altamente compleja (y no siempre es posible, tanto por las dificultades técnicas como por la disponibilidad de tiempo). Sin embargo, estos corpus permiten calcular, para algunos parámetros, ratios de verisimilitud (*likelihood ratios*) dentro del marco bayesiano, con lo que se dispone de información muy valiosa que le permite llegar a conclusiones más fiables en los dictámenes.*

Lingüística forense, corpus dubitado, corpus indubitado, corpus de referencia, ratios de verisimilitud

This article discusses the use of different kinds of linguistic corpora in forensic linguistics, from the point of view of both research and professional practice. In addition to the comparison between the corpora containing known and unknown oral or written texts, the expert should also use reference corpora in order to analyze the rarity or idiosyncrasy of the variables present in the known and unknown samples. The design of these reference corpora is highly complex (and not always possible, because of technical problems and shortness of time). However, these corpora allow to calculate, for some parameters, likelihood ratios within a Bayesian framework, which provide invaluable information, which in turn leads to more reliable conclusions on the expert reports.

Forensic linguistics, unknown corpus, known corpus, reference corpora, likelihood ratios

1. Introducción

En las últimas décadas, el estudio interdisciplinar de la lengua y el derecho ha llevado al nacimiento de una nueva disciplina, la lingüística forense (cultivada por lingüistas y juristas), que se encarga del estudio de la interficie entre el lenguaje y el derecho. Gibbons y Turell (2008) establecen los tres grandes ámbitos de la lingüística forense: el lenguaje jurídico (*the language of the law*), el lenguaje judicial (*the language of the court*) y el lenguaje evidencial (*language as evidence*).

Estos grandes ámbitos se concretan en estudios y propuestas que tienen como fin último mejorar la administración de la justicia. Por ejemplo, se analiza la legibilidad de documentos (leyes, normas, contratos, pólizas de seguros, etc.) con el fin de hacerlos más claros y accesibles a los ciudadanos, así como se estudia el lenguaje judicial (la lengua de los jueces, testigos, inculpados, policía, demandantes/querellantes, etc.) con los mismos propósitos. También dentro del lenguaje judicial tiene el interés el ámbito de la traducción e interpretación en contextos de apoyo multilingües, así como el análisis de los interrogatorios.

Sin embargo, es el uso de evidencia lingüística (fonética, fonológica, morfo-sintáctica, semántica, pragmática, discursiva) y pruebas periciales basadas en la comparación forense de textos el ámbito más conocido. Dentro de sus actuaciones se encuentran comparaciones de voces y textos que pueden ayudar a la justicia a identificar hablantes, determinar o atribuir la autoría de textos, elaborar perfiles lingüísticos a partir de muestras de voz o de escritura, detectar y analizar casos de posibles plagios, o dictaminar en casos de litigio de marcas registradas y patentes.

La lingüística forense, pues, es eminentemente una disciplina aplicada (*problem-based*: trabaja para dar respuesta a preguntas o problemas concretos). Sin embargo, su gran complejidad de análisis viene determinada básicamente por la variación inherente en el lenguaje. No existen dos voces iguales, ni se pueden producir –independientemente– textos iguales. Por estos motivos, se hace imprescindible el uso de corpus lingüísticos de distinta naturaleza. La tarea fundamental del lingüista forense cuando compara textos orales o escritos consiste en un doble análisis cualitativo y cuantitativo. En ambos casos, el elemento subyacente del análisis es la búsqueda de las variables que presenten una variación intra-hablante/escritor pequeña y una variación inter-hablante/escritor grande. Posteriormente, se puede calcular la ratio de variación entre hablantes/escritores diferentes respecto a la variación dentro del hablante/escritor (entre otros, Rose 2002; Coulthard 2004; Turell, 2010*b*).

2. Los corpus de análisis

2.1. Los corpus dubitados

Los corpus dubitados están formados por textos (orales o escritos) cuya autoría desconoce la autoridad judicial (jueces y magistrados). Si se trata de grabaciones de audio, habitualmente presentan unas condiciones de calidad acústica deficiente (ya sea porque

han sido obtenidas vía telefónica, con su consiguiente filtro frecuencial[124], bien porque se han obtenido en ambientes acústicos ruidosos, o bien porque se ha empleado instrumental de grabación de baja calidad). Este hecho, al que se añade a menudo la corta duración de los fragmentos en los que aparece la voz que hay que analizar, dificulta el proceso de comparación forense de voces.

Por su lado, si el material dubitado está compuesto por textos, éstos pueden ser de distinta naturaleza (cartas, notas de suicidio, mensajes SMS, apuntes de Facebook, e-mails, etc.). Sin embargo, el foco de análisis es estrictamente lingüístico (el perito lingüista no entra a valorar la caligrafía, tarea que compete a peritos calígrafos, ni los aspectos tecnológicos que tienen que ver con el proceso de elaboración o transmisión del mensaje).

2.2. Los corpus indubitados

Los corpus indubitados están formados por los textos (o grabaciones) cuya autoría es conocida. Sirven de base para compararlos con los textos que conforman el corpus dubitado. Sin embargo, su obtención puede llevarse a cabo en un entorno colaborativo o no colaborativo. Si el sospechoso colabora, se pueden obtener muestras escritas u orales de calidad y duración suficiente para encontrar un número aceptable de realizaciones de las variables lingüísticas que se pretenden estudiar. En caso contrario, se deben de buscar otros textos o grabaciones que puedan ser considerados como indubitados (grabaciones públicas, e-mails enviados desde la cuenta del sospechoso, etc.). Si no es posible su obtención, el trabajo del perito lingüista es imposible.

Además, hay que tener en cuenta que el registro utilizado en los textos indubitados es habitualmente muy distinto al que se refleja en los dubitados. La escritura o habla en entornos relajados (incluso a veces a partir de lecturas) presenta diferencias considerables en multitud de aspectos con respecto a las muestras dubitadas (habitualmente obtenidas en situaciones de estrés). Por ejemplo, en los corpus orales, los parámetros referidos a la frecuencia fundamental de la voz (cuyo correlato auditivo es el tono) difieren considerablemente dependiendo del nivel de estrés, el volumen de la voz, y el tipo y nivel de ruido del entorno (por causa del conocido como efecto Lombardo[125]).

2.3. Los corpus de referencia

Los corpus de referencia son corpus representativos de un grupo de población y tipología de acto comunicativo concreto. Son utilizados por los lingüistas forenses con la finalidad de comparar la ratio de aparición de las distintas variantes en la población general con

124 La transmisión telefónica filtra las frecuencias inferiores a los 300 Hz y las superiores a 3400 Hz (Künzel, 2001; Rose, 2003), con lo que la pérdida de información acústica es considerable; también introduce cambios significativos en algunos parámetros acústicos, tales como los formantes vocálicos, en comparación con grabaciones obtenidas vía microfónica directa (Künzel, 2001, 2002; Byrne y Foulkes, 2004).

125 El efecto Lombardo –explicado en French (1998) en el contexto forense¬– da cuenta, entre otros aspectos, del aumento del nivel de F0 y del volumen en contextos ruidosos. French (1998: 63) señala que "[i]n general, the findings are that, in noisy conditions, one would expect to find a reduced rate of speaking (measured either in terms of syllable production or relative vowel duration), a higher average frequency for the first formants of vowels and a higher average pitch". La razón de este efecto es que los hablantes aumentan el esfuerzo vocal en ambientes ruidosos con el fin de asegurar una comprensión óptima.

respecto a las muestras dubitadas e indubitadas. Estos corpus aportan datos de referencia acerca de la distribución poblacional (*base-rate knowledge of population distribution*) de las variables analizadas. En este sentido, tiene mucho más interés que tanto en la grabación dubitada como indubitada se observe una coincidencia en un parámetro muy raro en la población general que una coincidencia en un parámetro muy común. Por ejemplo, el poder identificador de un tono de voz muy agudo en la voz masculina de los corpus dubitado e indubitado toma relevancia cuando se observa que este tono es muy raro en la población masculina general. Un tono de voz cercano al promedio de la población masculina general, por el contrario, tendría una importancia en la identificación del hablante mucho menor.

El principal problema con el uso de corpus de referencia en lingüística forense es su definición para ser utilizados en esta disciplina. Todos los corpus lingüísticos se crean con una finalidad, y su principal efecto es la tipología de los textos que lo conforman y su tamaño. Por ejemplo, si se requiere investigar el uso de unas piezas léxicas en las obras de Cervantes, se va a crear un corpus con la totalidad de las obras cervantinas y se crearán colocaciones de las piezas para su análisis cualitativo y cuantitativo. En este ejemplo, el corpus de referencia incluye la totalidad de las obras del autor. Sin embargo, en lingüística forense el objetivo de los corpus de referencia es ver la distribución poblacional de las variables analizadas en las grabaciones o textos dubitados e indubitados. Por ello, lo ideal sería disponer de corpus de referencia del total de realizaciones lingüísticas (orales y escritas) de toda la población para su comparación. Esta tarea es, obviamente, imposible.

La principal dificultad es, entonces, definir cuál es la población de referencia que se quiere tomar en cuenta para el análisis. De nada sirve comparar los escritos de estudiantes unviersitarios jóvenes bilingües (utilizados como corpus de referencia) si los textos dubitados e indubitados se corresponden a una persona sólo con estudios primarios, vieja y monolingüe. O de nada sirve comparar los parámetros acústicos de una población española peninsular si las grabaciones forenses se corresponden a un inmigrante cuya lengua materna no es el español.

Aún así, el uso de corpus de referencia continúa siendo esencial como herramienta útil para contextualizar (aunque de modo inexacto) los datos obtenidos del análisis de las muestras lingüísticas dubitadas e indubitadas. En el ámbito de la fonética forense, se disponen de datos poblacionales suficientemente fiables para algunas variables. Se obtienen de corpus como Vocastel, Ahumada, Locupol, etc. En lenguaje escrito se disponen de corpus muy extensos y de ámbito general (CREA, CTILC, etc.) y otros más específicos.

3. Los corpus lingüísticos y el idiolecto

El idiolecto es el conjunto de formas lingüísticas (fonéticas, fonológicas, morfológicas, sintácticas, pragmáticas, discursivas) que cada hablante, de forma individual, utiliza en su uso del lenguaje. En palabras de Burridge y Mulder (1998: 302), se trata de la "variation within a language that is associated with individual speakers". En el plano fonético, Nolan (1994: 331) añade que "even within a narrowly defined dialect community, individuals will have their own preferred detailed pronunciations of particular words. The combination

of number of such preferred alternative pronunciations yields an overall pronunciation which is idiosyncratic, [and] that is, an individual's idiolect". La presuposición de la existencia del idiolecto (que es imposible de demostrar experimentalmente) es de gran interés para las aplicaciones de la lingüística forense, en el sentido que aporta las bases para individualizar el uso lingüístico de cada persona.

La aplicación del estudio del idiolecto a los análisis periciales de textos orales o escritos permite identificar el conjunto de las formas lingüísticas idiosincrásicas presentes en los textos analizados, lo que permite al perito concluir que las muestras dubitada e indubitada presentan rasgos compatibles con la hipótesis de que han sido producidas por una misma persona (lo que apoyaría a una identificación del hablante o del escritor) o no. Son numerosos los casos forenses en los que el análisis del idiolecto reflejado en los textos ha permitido aportar evidencia lingüística clave en la resolución de casos (por ejemplo, Svartvik, 1968; Coulthard, 1994; Turell, 2010*a*).

Sin duda, uno de los casos más famosos en los que la evidencia lingüística ha tenido un papel fundamental es el de Derek Bentley, condenado a muerte en 1953 por el asesinato de un policía y que fue 'absuelto' en 1998 a raíz del análisis lingüístico de su 'declaración' llevado a cabo por el lingüista forense M. Coulthard. Coulthard, a partir del análisis de la declaración y de distintos corpus de referencia demostró que la supuesta declaración de Bentley no podía ser *verbatim* (dictada palabra por palabra), como sostenía la policía y marca la ley, por lo que –de acuerdo con la postura de la familia en el caso de apelación– había sido manipulada de algún modo. Entre otros aspectos, destaca el uso del adverbio *then*. Según los corpus consultados por Coulthard, *then* ocurre una vez cada 500 palabras en el lenguaje general; una vez cada 930 en el corpus de declaraciones de testigos recopilado por el investigador; 1 vez cada 78 palabras en el corpus de declaraciones de policías (también confeccionado durante la investigación a cargo de Coulthard); y una vez cada 57 palabras en la supuesta declaración de Bentley. Pero el uso del adverbio pospuesto al pronombre *I* de sujeto ocurre solo una vez cada 16500 palabras en el lenguaje general; 1 de cada 5700 en declaraciones de testigos; 1 vez cada 100 palabras en declaraciones de policías; y una vez cada 190 palabras en la declaración de Bentley (Coulthard, 1993, 1994, 2006). Estos datos llevan a la conclusión de que efectivamente la supuesta declaración de Bentley no pudo haber sido *verbatim*, sino que fue un texto elaborado por la policía.

4. Las ratios de verisimilitud

El uso de los corpus lingüísticos de referencia también es fundamental para calcular las ratios de verisimilitud (*likelihood ratios*) de Bayes. Esta propuesta, de aplicación general en muchos ámbitos forenses no lingüísticos (como el análisis de ADN), se presenta como una solución "fiable y rigurosa" a las comparaciones forenses de textos y voces. "Esta propuesta se podría formular a partir de la siguiente pregunta: *¿cuál es la probabilidad, dada la evidencia de voz, de que las diversas muestras (de voz) hayan sido pronunciadas por la misma persona?*, y se concretaría en que los expertos forenses deberían proporcionar la probabilidad de la evidencia, dada la hipótesis [p(E|H)], fórmula en la que se basa la

Relación de Verosimilitud [...]" (Turell, 2010*b*: 16). En palabras de French *et al.* (2010: 143-144), "[t]he crucial insight of expressing a conclusion via a likelihood ratio (LR) is that it matters not only how likely the evidence is on the hypothesis that the suspect was responsible for leaving it, but also how likely it is given the alternative hypothesis that it was left by someone else". Esta es la opinión también expresada por Rose y Morrison (2009).

Para poder aplicar esta metodología cuantitativa es necesario disponer de datos de distribución poblacional de las distintas variables analizadas. Para ello, hay que definir una población de referencia y constituir corpus con muestras de voz o texto. Sin embargo, todos los autores reconocen la dificultad práctica y teórica de establecer la población de referencia:

> El gran problema no sólo se refiere a la dificultad de establecer cuál es la talla y características idóneas de la población-control, en la que hay que considerar la múltiple variedad de elementos sociolectales, dialectales e idiolectales del habla, sino también, a la diversidad de factores de variabilidad relacionados con otros aspectos de tipo patológico o emocional y otros muchos vinculados a los propios procesos de registro, transmisión, reproducción, conversión, compresión, etc de las emisiones habladas. Es decir, en cierta forma, cada locutor es en sí mismo una población. Además, todo proceso de registro, transmisión o codificación de su voz supondrá una mayor o menor modificación de su cualidad original y, en muchas ocasiones, la incorporación de importantes factores de degradación que dificultarán su evaluación por parte de los expertos forenses. (Delgado, 2005: 124)

> Further difficulties arise when we consider how to delimit the relevant population for comparison. It is commonly assumed that the population should be controlled for aspects of the speaker's regional and social background, since these factors may significantly affect the patterning of linguistic and phonetic features relevant to a case. We further note that population data must also be controlled for a wide range of other factors which are known to have significant effects on aspects of speech, voice and/or language. These include environmental effects of the recording situation (e.g. use of telephone, Lombard speech, transmission medium, recording hardware, distance from and orientation to the microphone) and short-term effects on the speaker resulting from e.g. smoking, intoxicants, or health problems. The range of potential factors to be controlled for is in fact very large indeed. (French *et al*., 2010)

5. Conclusiones

La implementación de modelos bayesianos basados en ratios de verisimilitud es una buena solución para la comparación forense de textos orales y escritos, aunque parcial, principalmente por la dificultad teórica y práctica de establecer poblaciones de referencia. A pesar del carácter matemático del modelo bayesiano, coexisten elementos subjetivos (además de los problemas ya comentados con la población de referencia): la expresión de los resultados mediante una escala de probabilidad verbal, la definición de las variables a analizar, la selección de los fragmentos pertinentes (¿el perito debe descartar fragmentos en los que la voz del sospechoso aparezca intencionadamente distorsionada?; ¿y los fragmentos de baja calidad?; ¿y los fragmentos en los que hay un cambio de lengua?;

etc.). Esta contradicción entre la objetividad de las ratios de verisimilitud y la subjetividad de la expresión de los resultados la observa Delgado (2005: 126):

> Desde un punto de vista teórico y general, la utilización de estimaciones de verosimilitud para materializar los resultados de análisis forenses parece una buena solución. Sin embargo, y con independencia de las particularidades ya referidas para la identificación de locutores, la propuesta de interpretación del teorema de Bayes, entendida en su globalidad, no acaba de ofrecer el grado de satisfacción que sería deseable. Si bien los papeles del científico, juez, fiscal, abogado, aparecen diferenciados, no se entiende muy bien qué sentido tiene la relación de sus roles y conclusiones en una igualdad matemática que conjuga supuestas valoraciones objetivas con apreciaciones de carácter subjetivo.

Sin embargo, siguiendo las tesis de Delgado (2001, 2005), Turell (2010*b*) y French *et al.* (2010), entre otros, la mejor solución se encuentra en la combinación de métodos de análisis (cualitativos y cuantitativos).

Referencias

Burridge, D. y Mulder, J. (1998). *English in Australia and New Zealand: An introduction to its structure, history and use*. Melbourne: Oxford University Press.

Byrne, C. y Foulkes, P. (2004). The 'mobile phone effect' on vowel formants. *Forensic Linguistics: The International Journal of Speech, Language and the Law 11*(1), 83-102.

Coulthard, M., (1993). Beginning the study of forensic texts: corpus, concordance, collocation. En M. P. Hoey (Ed.). *Data Description Discourse* (pp. 86-97). London: HarperCollins.

Coulthard, M. (1994). On the use of corpora in the analysis of forensic texts. *Forensic Linguistics: The International Journal of Speech Language and the Law,1*(1), 25-43.

Coulthard, M. (2004). Author identification, idiolect and linguistic uniqueness. *Applied Linguistics, 25*(4), 431-447.

Coulthard, M. (2006). …and then… Language Description and Author Attribution. Disponible en
http://www.aston.ac.uk/downloads/lss/english/Andthen_Coulthard.pdf

Delgado, C. (2001). *La identificación de locutores en el ámbito forense*. Tesis doctoral. Madrid: Facultad de Ciencias de la Información, Universidad Complutense.

Delgado, C. (2005). Comentarios sobre el contexto actual de la identificación forense de locutores. En M. T. Turell (Ed.). *Lingüística forense, lengua y derecho. Conceptos, métodos y aplicaciones* (pp. 113-129). Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra. DOCUMENTA UNIVERSITARIA.

French, J.P. (1998). Mr Akbar's nearest ear versus the Lombard reflex: a case study in forensic phonetics. *Forensic Linguistics. The International Journal of Speech, Language and the Law* 5 (1), 58–68.

FRENCH, J.P., NOLAN, F., FOULKES, P., HARRISON, PH. Y MCDOUGALL, K. (2010). The UK position statement on forensic speaker comparison: a rejoinder to Rose and Morrison. *The International Journal of Speech Language and the Law,17*(1), 143–152.

GIBBONS, J. Y M.T. TURELL (Eds.) (2008). *Dimensions of Forensic Linguistics*. Amsterdam/ Philadephia: John Benjamins.

KÜNZEL, H. J. (2001). Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies. *Speech Language and the Law. The International Journal of Forensic Linguistics*, *8*(1), 80–99.

KÜNZEL, H. J. (2002). Rejoinder to Francis Nolan's 'The 'telephone effect' on formants: a response. *Speech Language and the Law. The International Journal of Forensic Linguistics*, *9*(2), 83–6.

NOLAN, F. (1994). Auditory and acoustic analysis in speaker recognition. En J. Gibbons (Ed.). *Language and the law* (pp. 326-345). New York/London: Longman.

ROSE, P. (2002). *Forensic Speaker Identification*. London: Taylor & Francis.

ROSE, P. (2003). The technical comparison of forensic voice samples. En I. Freckelton y H. Selby (Eds.). *Expert Evidence* (cap. 99). Sydney: Thomson Lawbook Company.

ROSE, P., Y MORRISON, G. (2009). A response to the UK Position Statement on forensic speaker comparison. *The International Journal of Speech Language and the Law 16*(1), 139–163.

SVARTVIK, J. (1968). The Evans statements: a case for forensic linguistics. *Gothenburg Studies in English, 20*.

TURELL, M. T. (2010*a*). The use of textual, grammatical and sociolinguistic evidence in forensic text comparison. *The International Journal of Speech, Language and the Law*, *17*(2), 211-250.

TURELL, M. T. (2010*b*). *Los retos de la lingüística forense en el siglo XXI. In Memoriam Enrique Alcaraz Varó*. Alacant: Departamento de Filología Inglesa, Universitat d'Alacant.

Structured Parallel Coordinates: a visualization for analyzing structured language data

Chris Culy

Verena Lyding

Henrik Dittmann

European Academy of Bolzano/Bozen,
Italy

We present a visualization tool called Structured Parallel Coordinates (SPC), a specialization of Parallel Coordinates (cf., e.g., Inselberg, 2009), customized for the presentation and analysis of different types of structured language data, as found in corpora. Parallel Coordinates are a way of representing multidimensional data using a two-dimensional display. Interactive versions of Parallel Coordinates are flexible tools for data analysis, since selecting parts of the visualization allows for filtering the data (Inselberg, 2009). Language datasets often have dimensions which are interrelated or which have internal structure, a situation that is not accounted for by standard Parallel Coordinates. We describe the visual features and interactions provided by SPC to account for structured language data and give technical details of the tool. We present three sample applications of SPC that are closely linked to principle tasks in corpus analysis: (1) KWIC results as SPC, (2) ngrams and frequencies, and (3) ranking comparisons.

keywords: visualization, Parallel Coordinates, tools for analysis

Se presenta una herramienta de visualización llamada Structured Parallel Coordinates (SPC), una especialización de Parallel Coordinates (por ejemplo, Inselberg, 2009), adaptado para la presentación y análisis de diferentes tipos de datos de lenguaje estructurados. Parallel Coordinates son una forma de representar datos multidimensionales mediante una pantalla de dos dimensiones. Versiones interactivas de Parallel Coordinates son herramientas flexibles para analizar los datos, ya que seleccionando unas partes de la visualización se permite filtrar de los datos (Inselberg, 2009). Conjuntos de datos del lenguaje suelen tener dimensiones que están relacionadas entre sí o que tienen una estructura interna. Las características visuales y los interacciones preparadas por SPC son presentadas para explicar los datos y proporcionar los detalles técnicos. Se exponen tres ejemplos de SPC que están vinculados a las tareas principales en el análisis de corpus: (1) los resultados de KWIC como SPC, (2) ngrams y frecuencias, y (3) comparaciones de escala.

keywords: visualización, Parallel Coordinates, herramienta para el análisis

1. Introduction

Visualizations are a powerful means to support the processing, analysis and understanding of information by humans. The field of information visualization (InfoVis) is concerned with elaborating and evaluating possible ways to display different types of information and to create effective visualizations for it. Language data and any kind of linguistic information derived from it is often quite different from the types of information that InfoVis research has commonly focused on, like e.g. statistical and geospatial data. While InfoVis in general has matured, the specific concern with Linguistic Information Visualization (LInfoVis) is only recently getting more attention (see (Culy & Lyding, 2010a) and (Rohrdantz, Koch, Jochim, Heyer, Scheuermann, Ertl, *et al.*, 2010) for some examples), and applications targeted to language data are still scarce and not always linguistically informed (see (Wattenberg & Viégas, 2008) and (Hassan-Montero & Herrero-Solana, 2006) for some good general text-based examples).

In this paper we present *Structured Parallel Coordinates (SPC)*, a visualization tool for the presentation and analysis of different types of structured language data, as found in corpora. It is targeted to use by language analysts ranging from linguists to language teachers and learners. We describe the visual features and interactions provided by the tool, and explain how they respond to requirements of the prospective users and the characteristics of the data we are dealing with. We demonstrate, based on three elaborated sample applications, how *SPC* can be customized for different tasks.

2. Related Work

Structured Parallel Coordinates are a specialization of the *Parallel Coordinates* visualization (cf. d'Ocagne (1885), Inselberg (2009)). *Parallel Coordinates* are a way of representing multidimensional data using a two-dimensional display. Each dimension is represented along a vertical axis, and the values for a piece of data are connected by a line (see Figure 1). Interactive versions of *Parallel Coordinates* are flexible tools for data analysis, since selecting points and lines in the *Parallel Coordinates* display is the same as filtering the data (Inselberg, 2009). They are typically used with data dimensions that are conceptually independent, such as car size, year of manufacture, and mileage. *Parallel Coordinates* have been applied in many different contexts (e.g. (Inselberg, 2009) or (Steed, Fitzpatrick, Swan, & Jankun-Kelly, 2009)), with few, if any, detailed applications to language. *Parallel Tag Clouds (PTC)* (Collins, Viégas, & Wattenberg, 2009; Lee, Henry Riche, Karlson, & Carpendale, 2010) is similar to *SPC* visually, but it uses the size of words to indicate their frequency, and is not a true *Parallel Coordinates* visualization, since multiple dimensions cannot be selected. We have also implemented *PTC* as an application of *SPC*.

3. Structured Parallel Coordinates

3.1. The corpus analysis context

With *Structured Parallel Coordinates* we aim at providing a tool that extends the concept and functionality of the standard *Parallel Coordinates* visualization to support the

processing and analysis of language data derived from corpora. One particular problem in corpus linguistics is how to explore large result sets efficiently. A query might return thousands of examples from a corpus of millions of words and associated information (cf. the Corpus of Contemporary American English (COCA) (Davies, 2008), which has over 410 million words, or the recent freely available PAISÀ corpus of Italian (2011) with 500 million words). Our *Structured Parallel Coordinates* is a promising approach to visualizing corpus query results by providing the capability to explore and refine a set of query results without having to go back to the original data and redo a possibly complex query, in the same kind of spirit as, for example, Word Trees (Wattenberg & Viégas, 2008) or its linguistically specialized counterpart Double Tree (Culy & Lyding, 2010b). Other formats for corpus search results, such as KeyWord In Context (KWIC) lines and word lists, do not provide the flexibility and depth of information that we can provide by using *SPC*.

At the same time, in analyzing corpus data we deal with information that has dimensions that are not necessarily conceptually independent, but can be interrelated or have internal structure, unlike the typical uses of *Parallel Coordinates*. One fundamental type of structure is the sequential order of linguistic units like words, phrases, or paragraphs, plus statistical information associated with it. Another type of structure comes from meta-information associated with corpus texts, e.g. dates, where the data for each point in time can be treated as a dimension, and these dimensions are ordered (chronologically) with respect to each other. Rank orderings of (co-)occurrences of linguistic units provide an example of dimensions that have an internal structure: the ranks. *SPC* is designed to specifically account for the special nature of structured language data such as these, unlike general *Parallel Coordinates* visualizations which are not tailored to either language data or data that is structured across dimensions

3.2. Visual features and interactions in SPC

As with other *Parallel Coordinates* implementations, *SPCs* place data of different dimensions on vertical axes (one axis for each dimension) that are lined up horizontally. The axes may represent purely textual data (e.g. words), or purely numeric data (e.g. frequencies), and a single instance of *SPC* can contain axes of both types. Figure 1 shows ngrams of 'preposition' + 'verb' + 'any word class' plus their counts, extracted from a subset of the Italian corpus PAISÀ. On the first axis words are displayed, the second and third axes show word classes, and on axis four, counts are given on a numerical scale.[126] The different data dimensions may have an order among each other, as is the case for sequences of words in KWIC data, with one dimension/axis for each word position in a KWIC. They may also be without order, or ordered and unordered dimensions may occur together, as is in fact the case in Figure 1. A light red line placed vertically between ordered and unordered axes visually indicates the separation of these differently interrelated dimensions; here it visually separates the ngrams from their counts.

126 The remaining axes will be discussed in section 3.3.2. for Figure 3.

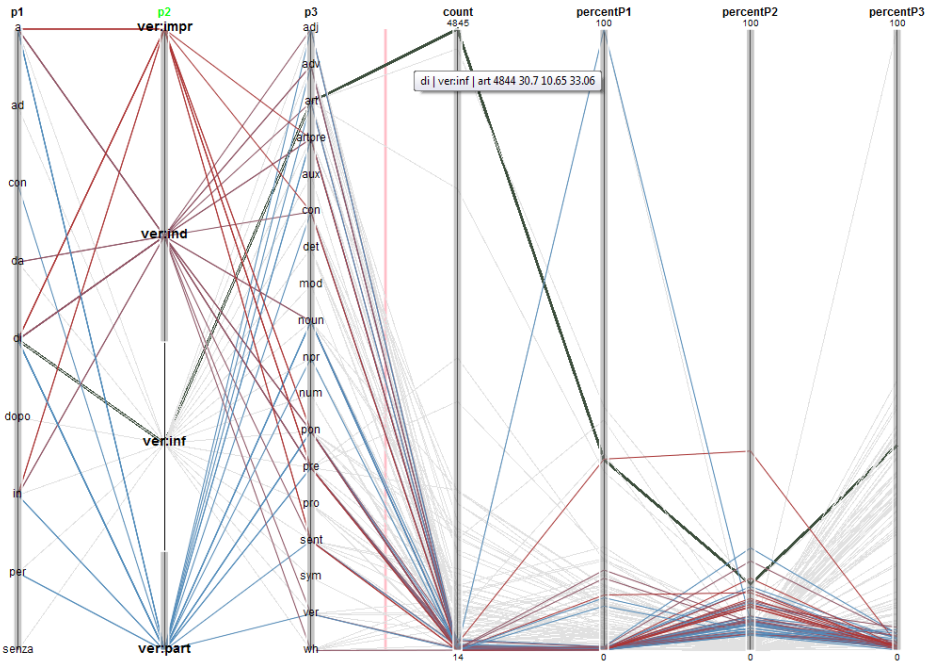


Figure 1. Ngrams with counts for sequences of: preposition + verb + any word

The colors of the lines shade from red to blue according to the order of the data on a determined axis (initially the left-most one). The specified axis, axis 2 in Figure 1, is indicated by a green label, and can be changed by clicking on the label of any other axis.

Data in SPC can be filtered by selecting data points (by dragging the mouse over it) on any axis. The connecting lines of data points which do not satisfy the resulting filters are rendered in gray.

By looking at the connecting lines, in Figure 1 we can easily see that infinitive verbs (*ver:inf*) have the widest combinatory variety of all verbs. To inspect this situation more closely data is filtered by all verbs that are not infinitives. Looking at the *count* axis, we can see that ngrams containing infinitive verbs do not just have the greatest variety, but make up all ngrams with a frequency higher than about 100 occurrences. In Figure 1, details on the most frequent ngram are given in a pop-up. Here this is "di" + "*ver:inf*" + "*art*" (*article*) with a total of 4844 occurrences.

3.3. Three applications of SPC

3.3.1. SPC for KWIC analysis

Working with KWIC results that show search units in context is perhaps *the* principle approach to corpus-based linguistic research. As corpus searches often yield large numbers

of results, to sort, filter and generally get an idea of the nature of the results is a necessary step of almost any KWIC based analysis. The strength of *SPC for KWIC analysis* is its compact representation format (for each position only one occurrence of every word type is shown), and the possibility to dynamically filter the data by selecting items.

Figure 2 shows an *SPC* for the KWIC results for the query for the lemma *vedere* ('to see') in a small corpus of Italian press releases (about 120.000 tokens), with two words of context to the left and right. The KWIC results are displayed with each axis representing a word position. Axes are ordered according to the sequential order of words in the KWIC. The words are displayed on each axis and presented in alphabetical order. KWIC sequences are represented by lines connecting the words. The results are filtered by position of the keyword, restricting the hits to future forms of the verb. The visualization shows that future forms make up the biggest part of the results. Having the results filtered by these forms, we can detect a particularity of the words preceding, directly or with a distance of 2, the future forms of *vedere*: they often represent events like here *incontro* ('meeting') and *conferenza* ('conference'). Hovering over the topmost line of the filtered data marks the line in bold and shows the sequence of data points (here "*L'incontro vedrà la partecipazione*" – 'The meeting will see the participation') in a pop-up box.

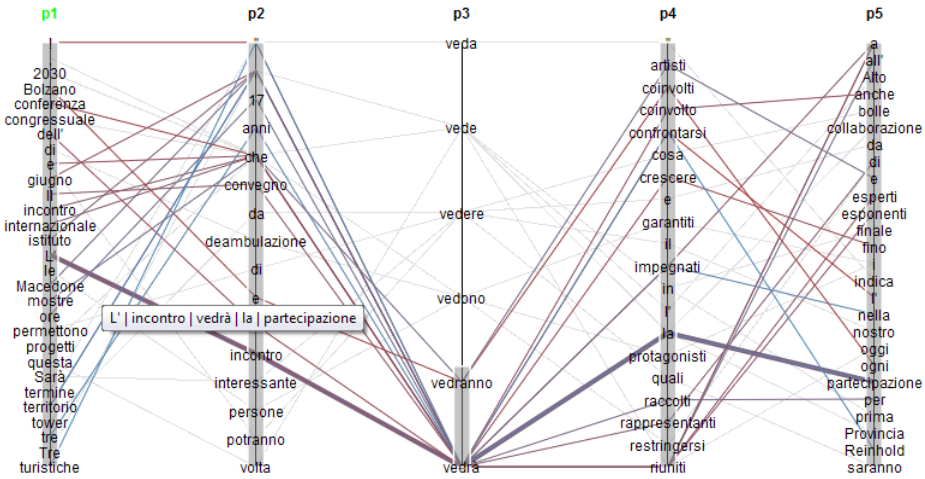


Figure 2. KWIC as SPC for the lemma vedere 'to see' in corpus of press releases

3.3.2. SPC for ngrams plus frequencies

Besides the question of which words (or other units of text) occur together, statistical information about their co-occurrence is of central relevance. *SPC for ngrams plus frequencies* is a visualization that combines these two types of information. Ordered data dimensions (the words of the ngrams) are presented together with unordered dimensions (statistical information related). In Figure 3 the first three dimensions show textual data—the results of a corpus query for pronouns followed by a form of the verb "to be" followed

by "happy" or "sad" —, presented on the axes, in their original reading order, while the other four dimensions contain absolute and relative frequency information related to the ngrams, where *count* indicates the total number of the ngrams and *percentP1* indicating the percentage of a word in position *p1* occurring in the particular ngram compared to all occurrences of this word. In Figure 3, for example, "she" occurs in the ngram "she's happy" just 2 times and for 4.44% of all occurrences of "she" in the results set. On the axes with numerical data the upper values are indicated on top of the axes, with it being 165 for the absolute count of ngrams and 100 for the axes indicating percentages. In Figure 3, on axis *p1* "he" and "she" are selected, on *p2* "'s", and on *p3* "happy". The resulting visualization shows that "she's happy" is a lower percentage of all the uses of "she" (in this result set) than "he's happy" is of the uses of "he", with about 20%. Interestingly, the situation is inverse for ngrams for the unabbreviated form of "is". In fact "she is happy" makes up for 37.77% of all occurrences of "she", while "he is happy" makes up for only 28.91% of the total occurrences of "he". The data is taken from a 100 million word Corpus of British English compiled from the web (ukWaC, (Ferraresi, Zanchetta, Baroni, & Bernardini, 2008).

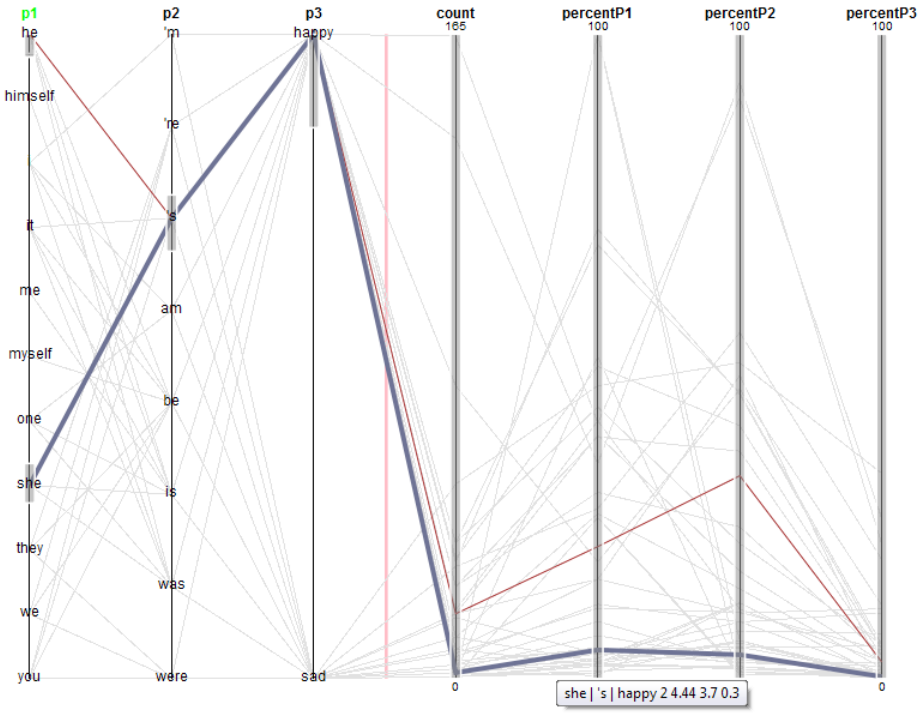


Figure 3. SPC for ngrams plus frequencies of preposition followed by lemma "to be" followed by "happy/sad"

3.3.3 SPC for ranking comparisons

The comparison of occurrences of linguistic phenomena is what *SPC for ranking comparisons* is designed for. The comparison could be of the same phenomenon across different corpora or subcorpora, or it could be of different aspects of a single phenomenon. Each data axis contains the linguistic units under inspection ordered by frequency. Figure 4 shows a visualization of the top 20 most frequent words starting with *[Ss]elbst* ('self'), counted by lemma, in 5 years of newspaper text, ranging from 1991 to 2006. The data is taken from a corpus of the German language South Tyrolean newspaper *Dolomiten*. In contrast to the KWIC and ngrams visualizations presented above, the axes are not ordered by the linear order of word sequences but with respect to the ordering of corpus metadata, the year of the originating text. Words are placed on the axes according to their rank ordering. "[NA]" indicates that a word that occurs in the top *n* words in another year is not among the top *n* in the year for that axis. In the case of multiple words with the same frequency, the words are given the same rank and the list of words is associated with the data point (rank). Since the list could be quite long (e.g. for hapax legomena), only one word is shown on the axis, and the others are shown when the user hovers the mouse over the word shown. E.g. in Figure 4 on the axis with data from 2001, *Selbstbestimmung* ('self-determination') shares its rank with another word (here *Selbstverständlichkeit* ('implicitness')), as is indicted by three dots following the term. This way of presenting words with the same frequency was decided upon in response to initial user feedback.

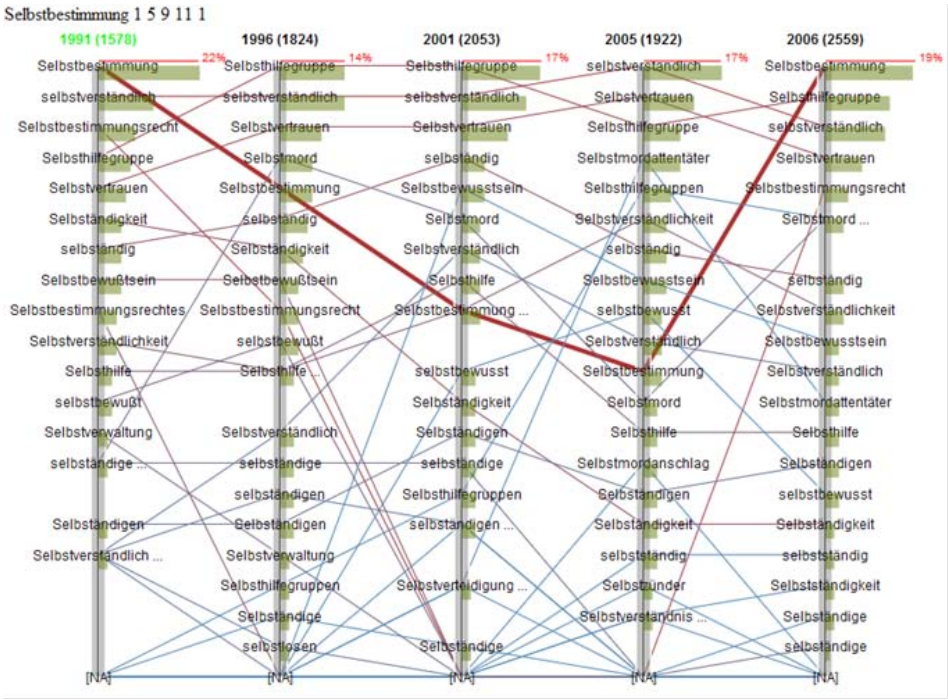


Figure 4. SPC for rank orderings of the top 20 words starting with "[sS]elbst" in 5 years of the Dolomiten

The ordering within the dimensions causes the positive or negative declination of the lines to have a meaning. Lines sloping upwards from left to right indicate an increase in rank ordering (though not necessarily of frequency, either absolute or relative to the series). Similarly, lines sloping downwards from left to right indicate a decrease in rank ordering. In Figure 4 the connecting line to the term "Selbstbestimmung" (*'self-determination'*) is highlighted. It shows that from 1991 to 2005 it was continuously decreasing in rank, and from 2005 to 2006 got back to its original (first) ranking position. For the easy comparison of number of occurrences relative to total number of occurrences within each year, we added small horizontal bars to each data point on each axis to indicate percentages of occurrences. The visualization allows for a quick perception of the relative frequencies of items within and across dimensions. For the analysis in Figure 4 we can see that from 2001 to 2005 "Selbstbestimmung" decreased in rank, while its proportion in the results set did not change that much. To the contrary, the term holds the first rank both in 1991 and 2006, but in 2006 is only makes up 19% of the hits as opposed to 22% for 1991.

3.4. Technical details

SPC is designed to provide a core set of general visualization and interaction functionalities that can easily be customized and implemented for specific applications. It is implemented in JavaScript using the toolkit Protovis (Bostock & Heer, 2009) and runs in a browser. Functions and properties both on visual and interaction aspects of the visualization, as well as the fundamental aspects of the data ordering, can be easily customized for different applications of *SPC* as we have seen.

4. Conclusion

To sum up, *SPC* is a new type of *Parallel Coordinates* that is designed specifically for language data. It is a general tool that can be used to analyze different kinds of data with the current applications. For example, some of our colleagues are using *SPC* to analyze learner texts. *SPC* can also be extended to provide additional kinds of functionality, as we showed above in implementing *PTC* as an *SPC* application. *SPC* is an innovative tool for corpus analysis, which illustrates opportunities that are created when visualization techniques are adapted to the special needs of language information. *SPC* and the applications are freely available under an Open Source license.

References

Bostock, M., & Heer, J. (2009). Protovis: A Graphical Toolkit for Visualization. *IEEE Transactions on Visualization and Computer Graphics, 15*(6), 1121-1128.

Collins, C., Viégas, F. B., & Wattenberg, M. (2009). Parallel Tag Clouds to Explore and Analyze Faceted Text Corpora. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST).* (pp. 91-98).

Culy, C., & Lyding, V. (2010a). Visualizations for exploratory corpus and text analysis. In *Proceedings of the 2nd International Conference on Corpus Linguistics (CILC-10)*. A Coruña, Spain. (pp. 257-268).

Culy, C., & Lyding, V. (2010b). Double Tree: An Advanced KWIC Visualization for Expert Users. In *Proceedings of the 14th International Conference on Information Visualization (IV)*. (pp. 98-103).

Davies, M. (2008). The Corpus of Contemporary American English (COCA): 410+ million words, 1990-present. http://www.americancorpus.org.

Ferraresi, A., Zanchetta, E., Baroni, M., & Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the Fourth Web as Corpus Workshop*. (pp. 47-54).

Hassan-Montero, Y., & Herrero-Solana, V. (2006). Improving tag-clouds as visual information retrieval interfaces. In *Proceedings of InSciT2006*, Mérida.

Inselberg, A. (2009). *Parallel Coordinates: VISUAL Multidimensional Geometry and its Applications*. New York: Springer.

Lee, B., Henry Riche, N., Karlson, A. K., & Carpendale, S. (2010). Spark Clouds: Visualizing Trends in Tag Clouds. *IEEE Tranactions on Visualization and Computer Graphics, 16*(6), 1182-1189.

d'Ocagne, M. (1885). *Coordonnées Parallèles et Axiales: Méthode de transformation géométrique et procédé nouveau de calcul graphique déduits de la considération des coordonnées parallèlles*. Paris: Gauthier-Villars.

Piattaforma per l'Apprendimento dell'Italiano Su corpora Annotati (PAISÀ), http://www.corpusitaliano.it, 2011-

Rohrdantz, C., Koch, S., Jochim, C., Heyer, G., Scheuermann, G., Ertl, T. *et al.* (2010). Visuelle Textanalyse. *Informatik-Spektrum*, *33*(6), 601-611.

Steed, C. A., Fitzpatrick, P. J., Swan, J. E., & Jankun-Kelly, T. J. (2009). Tropical Cyclone Trend Analysis Using Enhanced Parallel Coordinates and Statistical Analytics. *Cartography and Geographic Information Science, 36*(3), 251-265. doi:10.1559/152304009788988314.

Wattenberg, M., & Viégas, F. B. (2008). The word tree, an interactive visual concordance. *IEEE Transactions on Visualization and Computer Graphics*, *14*(6), 1221-1228.

Request markers in drama: data from the *Corpus of Irish English*

Fátima Faya Cerqueiro[127]
Universidad de Castilla-La Mancha

Abstract

In the Late Modern English period there was a change in the main courtesy marker used in polite requests, from *pray*, the former marker, to *please*, which still is the default marker in Present-day English. Drama texts are considered very useful for the study of pragmatic features due to their dialogic interaction, and therefore the analysis of a drama corpus may yield interesting results regarding the shift and the evolution of these forms. Courtesy markers *pray* and *please*, together with their verbal structures will be examined in the drama section of the *Corpus of Irish English*, which covers the whole Late Modern English period.

Key words: *please*, *pray*, requests, drama, *Corpus of Irish English*

A lo largo del período de inglés moderno tardío se produce un cambio en el principal marcador de cortesía en peticiones, por el cual *pray* deja de ser la forma más usada con dicha función para cederle su puesto a *please*, que es aún hoy la forma de cortesía por excelencia en peticiones. Las obras de teatro se consideran un recurso muy útil para el estudio de elementos pragmáticos debido al diálogo y la interacción presentes en este tipo de texto, por lo que el análisis de un corpus dramático puede aportar resultados interesantes sobre el cambio y la evolución de estos marcadores. De tal forma, *pray*, *please* y otras estructuras con sus formas verbales se analizarán en el *Corpus of Irish English*, ya que abarca todo el periodo de inglés moderno tardío, y su sección de textos dramáticos está muy bien representada.

Palabras clave: *please*, *pray*, peticiones, teatro, *Corpus of Irish English*

127 For generous financial support thanks are due to the Autonomous Government of Galicia (INCITE grant 08PX-IB204016PR) and the Spanish Ministry for Science and Innovation and the European Regional Development Fund (grant HUM2007/60706).

1. Introduction

In the Late Modern English period we observe a change in the use of main request markers, whereas *pray* was the most common courtesy marker in requests at the beginning of this period, it was eventually replaced by *please* and the former marker disappeared entirely in the twentieth century.

A preliminary study in *ARCHER* (*A Representative Corpus of Historical English Registers*)[128] showed that *pray* and *please* were found mainly in three types of texts, namely letters, fiction and drama. The analysis of those items in novels and letters have already brought interesting results about the evolution of these markers, and especially about the replacement of *pray* by *please* (cf. Faya Cerqueiro, 2007 and 2009). Nevertheless, requests markers have not been studied in drama texts yet. Therefore, an analysis of plays will help to complete the whole picture of the main request markers in the Late Modern English period, and will allow text-type comparisons. For this purpose I will make use of the *Corpus of Irish English* (cf. Section 3 below).

2. Drama and historical pragmatics

Drama is probably the most profitable fictional genre for the study of pragmatic issues, especially those regarded as typical of the spoken language. Even though it should be admitted that this genre contains an imitation of actual speech, it represents the spoken medium as close as possible and if it is "used with the necessary caution, plays may also yield insights into what counted as polite or impolite behaviour and how, for instance, greetings, insults or compliments were realised at that time" (Jucker, 1994: 535). Culpeper and Kytö (1999) classify drama as constructed dialogue with minimum of narratorial intervention, since apart from stage directions, plays contain dialogue almost exclusively. There are important contributions to historical pragmatics using drama as a corpus, proving the relevance of this text-type in pragmatic analyses, mostly focused on Early Modern English, among them we find Brown and Gilman (1989) and Kopytko (1993), who applied Brown and Levinson's politeness model to Shakespearean works, and Middle English, such as Mazzon (2009).

The selection of a particular text-type is a key issue when dealing with pragmatic aspects, since some features are more likely to occur in a certain group of genres, for example in drama, which is supposed to represent oral language. In this respect, Rissanen points out that it is safe to hypothesise that those variants which are more frequent in recorded speech than in other text-types were more frequent in the spoken language of the period (cf. Rissanen, 1986: 98).

As mentioned above, taking advantage of the multigenre approach that *ARCHER* allows, I selected the period 1850-1959 (649,170 words), in order to analyse the markers *please*, *pray* and *if you please*, paying attention to their distribution and pragmatic functions. Of the 9 different genres in *ARCHER* (drama, fiction, sermons, journal or diaries,

128 I am grateful to the VARIENG Research Unit (University of Helsinki) for providing me access to several corpora, among them *ARCHER* and *A Corpus of Irish English*.

legal, medicine, news, science, letters), only four of them contain examples of the courtesy markers under consideration. These are letters, drama, fiction and journal or diaries. Therefore, the analysis of the data in *ARCHER* confirms there are evident genre differences and further analysis of those three text types was necessary. As should be expected, only those genres in which a certain kind of interaction takes place contain these request markers. We find the highest overall figures in drama, followed by letters and fiction, as we can see in Figure 1 below:

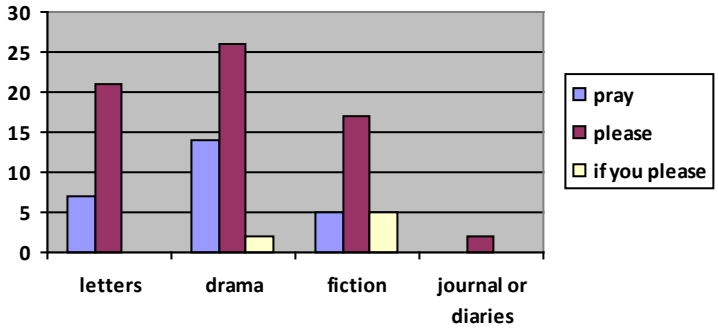


Figure 1. Data from ARCHER (1850-1959)

3. The Corpus of Irish English

The *Corpus of Irish English* collects Irish documents written in English from the early fourteenth century up to the twentieth century, allowing diachronic analyses.[129] The different genres represented in this corpus comprise poetry, glossaries, sketches and full-length plays, although drama is the best represented genre in the corpus. The material compiled from the sixteenth to the eighteenth centuries in the corpus includes not only "genuine representations of Irish English by native Irish writers" but also "texts by non-Irish writers where the non-native perception of the Irish English is found" (Hickey, 2003: 242). As regards number of words, the drama selection of this corpus contains an approximate number of 500,000 words, although the twentieth century provides almost half of them. The representation of works and number of words in each century is as follows:[130]

- Sixteenth century: 3 plays (by 3 different authors) / 1,744 words
- Seventeenth century: 8 plays (by 7 different authors) / 8,044 words
- Eighteenth century: 10 plays (by 8 different authors) / 121,462 words
- Nineteenth century: 10 plays (by 4 different authors) / 118,307 words
- Twentieth century: 13 plays (by 4 different authors) / 244,644 words

129 I will not consider here dialectal variation in the corpus.
130 For a complete list of plays in the corpus, see Hickey (2003: 243-245)

4. Data analysis

The replacement of *pray* by *please* has been generally placed around the nineteenth and the twentieth century (cf. Akimoto, 2000: 76), and this is also confirmed by the analysis of Irish drama. Figure 2 shows the evolution of frequencies of courtesy markers *please* and *pray* in the data analysed.[131] There are no instances of the markers in the sixteenth-century plays included in the corpus, whereas in the seventeenth century we can find different forms of *pray*, among which there is an extraordinary high frequency of *prithee*. *Pray* appears as a completely grammaticalised form in the eighteenth century. The only forms of *please* found in the seventeenth century are conventionalised phrases in which there is an experiencer object, such as *(and) please your majesty*/*grace* or *and't please you*, which are also present in the eighteenth century.

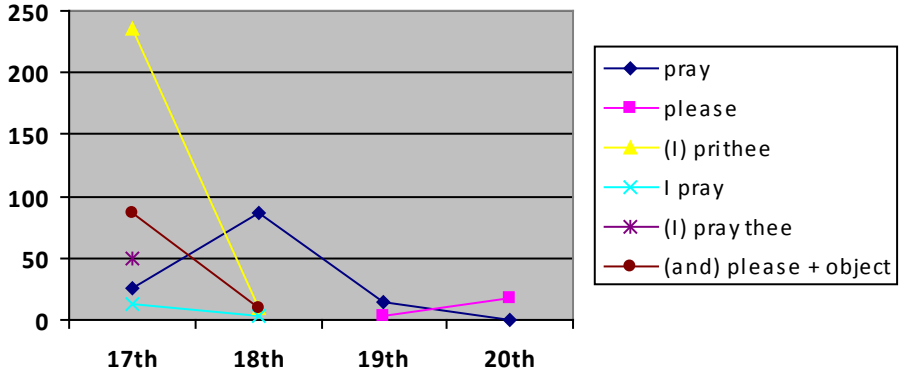


Figure 2. Pray, please and other parenthetical forms in Irish drama

Pray is found in two main pragmatic functions, either softening a request in the imperative, as in (1), or as an attention getter, commonly preceding a question, and often accompanied by other vocatives, as in (2):

1) JUST. Your humble Servant, honest Humphry - don't mind me – **pray** don't let me interrupt you. (Sheridan, Richard Brinsley. *St. Patrick's Day or The Scheming Lieutenant*)

2) WITWOUD. Is that the way? **Pray** Madam, do you pin up your Hair with all your Letters? I find I must keep Copies. (Congreve, William. 1700. *The Way of the World*)

Some authors have pointed at an origin of *please* in the conditional parenthetical *if you please* (cf. Traugott and Dasher, 2002: 255-257 and Brinton, 2006: 326), whereas previous studies on letters and novels have suggested a verbal origin of *please*: *be pleased to >please to > please* (cf. Tieken and Faya Cerqueiro, 2007). Thus, *please* would have emerged from sentences such as those in example (3), an interrogative in which the politeness expression *be pleased to* is further downtoned by the modal *will* and (4), the

131 Frequencies are calculated per 100,000 words. Since four plays by Lady Gregory do not contain any example of pray or please, they have been excluded from the normalised frequency count in the nineteenth century.

only imperative form in the corpus, which is nevertheless frequently found in eighteenth century letters:

3) SULLEN My head aches consumedly.

   MRS. SULLEN **Will** you **be pleased**, my dear, **to** drink tea with us this morning? It may do your head good. (Farquhar, George. 1707. *The Beaux' Stratagem*)

4) BONIFACE **Please** to bespeak something else. I have everything in the house. (Farquhar, George. 1707. *The Beaux' Stratagem*)

The analysis in novels and letters projected the idea of a possible evolution from these constructions. Thus, in letters from the second half of the eighteenth century, PROP constructions, such as imperative *please to* are the most frequent ones (cf. Faya Cerqueiro, 2007), whereas they are completely absent from this drama corpus in the whole eighteenth century. Surprisingly, *please to* constructions show a very low frequency in this corpus, they appear only in the eighteenth century: the imperative *please to* with a frequency of 0,82, whereas *please to* with presence of modal *will* reaches a frequency of 4,12. Both structures are absent in the remaining centuries. *Be pleased to* is not found in the imperative in the corpus, but the construction with *will* reaches a frequency of 0,82 in the eighteenth century, and 1,02 in the nineteenth century.

Figure 3 below shows the contrast between the evolutions of *please* and *if you please*. In the eighteenth century, when *please* as a courtesy marker was not yet available, *if you please* shows a higher frequency in this corpus than in any other period, and even than in the other text-types analysed. As we have seen in Figure 1 in the data from *ARCHER*, fiction and drama are the only genres which include this parenthetical conditional in the period 1850-1960, which is absent in *ARCHER* letters.

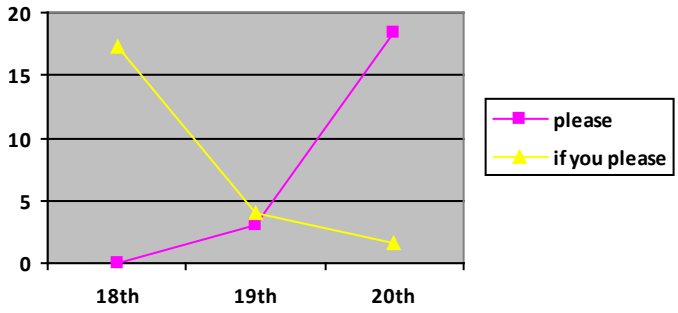


Figure 3. Please and if you please as parentheticals in Irish drama

Following Rissanen's observation mentioned above, features from other periods which are frequent in drama are expected to have been frequent also in spoken language. Thus, it is possible that *if you please* was more a feature of spoken language, whereas other expressions such as the imperative constructions, which are more frequent in letters could

have been more typical of writing. Instances (5) and (6) show the parenthetical *if you please* used as a courtesy marker.

5) AIMWELL Has the lady been any other way useful in her generation?

   BONIFACE Yes, sir, she has a daughter by Sir Charles, the finest woman in all our country, and the greatest fortune. She has a son too by her first husband, Squire Sullen, who married a fine lady from London t'other day. **If you please**, sir, we'll drink his health? (Farquhar, George. 1707. *The Beaux' Stratagem*)

6) HARDCASTLE Ha! ha! ha! The story is a good one. Well, honest Diggory, you may laugh at that - but still remember to be attentive. Suppose one of the company should call for a glass of wine, how will you behave? A glass of wine, Sir, **if you please** (to DIGGORY) - Eh, why don't you move? (Goldsmith, Oliver. 1773. *She stoops to conquer*)

If you please is not at all constrained as regards word-order in the sentence, occupying different positions. This fact makes this conditional parenthetical an unsuitable source for the courtesy marker *please*, which would have emerged in initial position in the sentence, showing a word-order behaviour more typical of imperatives (cf. Faya Cerqueiro, 2009: 32-33).

Even if the coexistence of *pray* and *please* in this corpus is underrepresented as compared to other Late Modern English corpora, we still find some interesting instances of the same character using both markers. This is the case of Gwendolen and Algernon, both characters in Wilde's *The Importance of Being Earnest* (1895) in the nineteenth century, as we can see in examples (7-10) below:

7) CECILY [Severely] Cake or bread and butter?

   GWENDOLEN [In a bored manner] Bread and butter, **please**. Cake is rarely seen at the best houses nowadays.

8) GWENDOLEN **Pray** don't talk to me about the weather, Mr. Worthing. Whenever people talk to me about the weather, I always feel quite certain that they mean something else. And that makes me so nervous.

9) ALGERNON Oh! there is no use speculating on that subject. Divorces are made in Heaven - [JACK puts out his hand to take a sandwich. ALGERNON at once interferes] **Please** don't touch the cucumber sandwiches. They are ordered specially for Aunt Augusta. [Takes one and eats it]

10) JACK Well, produce my cigarette case first.

    ALGERNON Here it is. [Hands cigarette case] Now produce your explanation, and **pray** make it improbable. [Sits on sofa]

In the twentieth century the only character using *pray*, Miss Gilchrist in Behan's *The Hostage* (1959), makes also use of *please*, as we can observe in examples (11) and (12):

11) MEG How dare you? When I was ill I lay prostituted on that carpet. Men of good taste have complicated me on it. Away, you scruff hound, and thump your craw with the other hippocrites.

MISS GILCHRIST **Pray** do not insult my religiosity.

MEG Away, you brass.

MISS GILCHRIST I stand fast by my lord, and will sing my hymn now:

12) MULLEADY I can't, MISS GILCHRIST I haven't paid my rent.

MISS GILCHRIST I will pray for you, Eustace. My shoes, **please**.

MULLEADY [fetching her shoes] Will you come back, Miss Gilchrist?

These instances exemplify the coexistence of both markers in the nineteenth and, still, in the twentieth century. The features of drama and the depiction of characters provide some nuances in the particular uses of the markers. Whereas characters using *please* in the nineteenth century are young, modern people, a young gentleman, Algernon, and a young lady, Gwendolen, Miss Gilchrist in the second half of the twentieth century represents a profound and old-fashioned religiosity.

5. Concluding remarks

The data in drama differ to a great extent with the data in novels or letters from previous studies, especially as regards indications of the origin of *please*, emphasising obvious genre differences. Contrary to what data in *ARCHER* could point at, this drama corpus contains a conservative stage of language in the nineteenth and twentieth centuries. We cannot appreciate notable evidence concerning the development of the courtesy markers *please* and *pray* since the difference in their frequencies from the nineteenth to the twentieth century data represents a huge contrast. Even though, the interaction of drama is nevertheless interesting from the pragmatic point of view and provides insights on the usages of both markers at different points in the history of English and the sort of characters that made use of them.

6. References:

Akimoto, M. (2000). The grammaticalization of the verb 'pray'. In O. Fischer, A. Rosenbach and D. Stein (Eds.), *Pathways of Change. Grammaticalization in English* (pp. 67-84). Amsterdam and Philadelphia: John Benjamins.

ARCHER = A Representative Corpus of historical English Registers 3.1. 1990-1993/2002/2007/2010. Compiled under the supervision of Douglas Biber and Edward Finegan at Northern Arizona University, University of Southern California, University of Freiburg, University of Heidelberg, University of Helsinki, Uppsala University, University of Michigan, University of Manchester, Lancaster University, University

of Bamberg, University of Zurich, University of Trier, University of Salford, and University of Santiago de Compostela.

Brinton, L. (2006). Pathways in the Development of Pragmatic Markers in English. In A. van Kemenade and B. Los (Eds.), *The Handbook of the History of English* (pp. 307-334) London: Blackwell.

Brown, R. & Gilman, A. (1989). Politeness theory and Shakespeare's four major tragedies. *Language in Society*, *18*, 159-212.

Culpeper, J. & Kytö, M.. (1999). Modifying pragmatic force hedges in Early Modern English dialogues. In A. H. Jucker, F. Lebsanft and G. Fritz (Eds.), *Historical Dialogue Analysis* (pp. 293-312). Amsterdam and Philadelphia: John Benjamins.

Faya Cerqueiro, F. (2007). The courtesy markers *pray* and *please* in the late 18th century: evidence from the Corpus of Late Eighteenth-Century Prose. In M. Losada Friend, P. Ron Vaz, S. Hernández Santano and J. Casanova (Eds.), *Proceedings of the 30th AEDEAN International Conference*. Huelva: Servicio de Publicaciones de la Universidad de Huelva. [CD-Rom]

Faya Cerqueiro, F. (2009). *Please* in nineteenth-century English: origin and position of a courtesy marker. In C. Prado-Alonso et al (Eds.), *New Trends and Methodologies in Applied English Language Research. Diachronic, Diatopic and Contrastive Studies* (pp. 25-36). Bern: Peter Lang.

Hickey, R. (2003). *Corpus Presenter: Software for Language Analysis with a Manual and* A Corpus of Irish English *as Sample Data*. Amsterdam: John Benjamins.

Jucker, A. H. (1994). The feasibility of historical pragmatics. *Journal of Pragmatics*, *22*(5), 533-536.

Kopytko, R. (1993). *Polite Discourse in Shakespeare's English*. Poznan: Adam Mickiewicz University Press.

Mazzon, G. 2009. *Interactive Dialogue Sequences in Middle English Drama*. Amsterdam and Philadelphia: John Benjamins.

Rissanen, Matti. 1986. Variation and the study of English historical syntax. In D. Sankoff (Ed.), *Diversity and diachrony* (pp. 97-110). Amsterdam and Philadelphia: John Benjamins.

Tieken-Boon van Ostade, I. & Faya Cerqueiro, F. (2007). Saying *please* in Late Modern English. In J. Pérez-Guerra et al. (Eds.), *'Of varying language and opposing creed': New insights into Late Modern English* (pp. 421-444). Bern: Peter Lang.

Traugott, E.C & R. Dasher. (2002). *Regularity in Semantic Change*. Cambridge: Cambridge University Press.

A Preliminary Study of Neutral Motion Verbs in *LOB* and *FLOB*

Iria Gael Romai
Universidad de Santiago de Compostela

The semantic domain of motion and space has been exhaustively studied in the last decades, being considered a cognitive universal. Research in the particular field of motion is mainly based on Talmy's (1991, 2000, 2007) typological classification of languages into Satellite-framed (which codify path through satellites) and Verb-framed, (which codify path within the verb). Thus, V-language users tend to encode fewer path segments than S-language users in both speech and written language. Moreover, in S-languages, path information is expressed in a more compact way than in V-languages. In this preliminary study three neutral English Run verbs (walk, run and jump) that express manner of motion have been taken into consideration by comparing two sub-periods of Present-day British English (the 1960s and the 1990s) as represented in the LOB and the FLOB corpora respectively.

Space, motion, lexicalization, Historical Linguistics.

El campo semántico del espacio y el movimiento es considerado como un universal cognitivo y, por ello, se ha estudiado intensivamente durante las últimas décadas. La investigación en este campo se basa principalmente en la clasificación tipológica de lenguas propuesta por Talmy (1991, 2000, 2007), que las divide en 'lenguas enmarcadas en satélite' (codifican el 'path' a través de satélites) y 'lenguas enmarcadas en verbo' (codifican el' path' dentro del verbo). Por tanto, los hablantes de lenguas-V tienden a codificar un menor número de segmentos 'path' que los hablantes de lenguas-S, tanto en lenguaje hablado como escrito. Además, en lenguas-S, la información sobre el 'path' está expresada de un modo más compacto. En este estudio preliminar se han tenido en cuenta tres verbos neutros del tipo 'Run' (walk, run y jump) que expresan modo de movimiento, comparando dos sub-periodos del inglés británico contemporáneo (la década de los 60 y la de los 90) a través de los corpora LOB y FLOB respectivamente.

Espacio, movimiento, lexicalización, Lingüística histórica.

1. Introduction

The present research is a pilot study of some of the verbs that express neutral manner of motion meaning in English: *walk*, *jump* and *run*, comparing two sub-periods of Present-Day English (the 1960's and the 1990's) in order to see whether there have been changes in usage along the time axis. This study is part of a larger project whose aim is to provide a contrastive analysis of the development of verbs of manner of motion in English and Spanish as represented in different corpora.

Manner of motion verbs is the largest class in the semantic domain of motion events. According to Levin (1993: 264-5) we can divide manner of motion verbs into two sub-classes: *Roll* verbs and *Run* verbs. *Roll* verbs typically express manners of motion of inanimate entities as well as motion Around an Axis. In turn, *Run* verbs describe the manners of movement of animate entities

2. Motion events

A clear definition of a motion event is that proposed by Talmy (2000: 25): a motion event "consists of one object (the *Figure*) moving or located with respect to another object (the reference object or *Ground*)". Moreover, according to the previous definition there are many participants involved in the expression of motion events:

- Figure.
- Path, i.e. direction of movement.
- Extent.
- Manner
- Ground, which can be divided into source (initial location), goal (final location), milestone (location passed along path).

The domain of motion is essential in all languages but they differ in the way of codifying motion events because they have developed different types of lexicalization patterns, which raise specific forms of narrative style and mental imagery in the motion domain. In this connection, Talmy put forward a typological classification of motion events by dividing languages into '*Satellite-framed*' languages and '*Verb-framed*' languages. The difference here lies in the lexicalization of path. If one language codifies or 'frames' a path within the verb, then it is a 'verb-framed' language, whereas if it codifies path through satellites, it is referred to as being 'satellite-framed'. Thus, motion events in V-languages are typically expressed by the combination of a path verb and a subordinate adverbial of manner, in contrast with S-languages, which express them by means of a manner of motion verb + path satellite.

(a) *Satellite-framed construction type:*

MOTION, MANNER ↓ VERBfinite ↓ *go, run out*	PATH ↓ SATELLITE ↓ *in*	SOURCE/GOAL ↓ N+(adposition,case) ↓ *of the house* *to the house*

(b) Verb-framed construction type:

MOTION,PATH ↓ VERB finite ↓ *salir* 'exit' *entrar* 'enter'	SOURCE/GOAL ↓ N+(adposition, case) ↓ *de la casa* 'of the house' *en la casa* 'in the house'	MANNER ↓ VERB nonfinite ↓ *corriendo* 'running' *corriendo* 'running'

Figure 1. Satellite and verb-framed constructions.

These patterns make speakers of V-languages and S-languages focus their attention on different components of motion events. S-language speakers consider manner as an inherent component and, as a result, the domain of manner is widely elaborated in these languages. In contrast, V-language speakers focus less on manner and more on changes of location and settings.

Nevertheless, several languages do not fulfil either of the patterns. This is the case of serial-verb languages such as Mandarin Chinese, Thai or West-African languages, since they exhibit series of morphologically unmarked and monosyllabic verbs. These languages are classified by Talmy as S-framed, by considering the manner verb as the main verb and path verbs as satellites due to their impossibility of being used as lexical verbs. Moreover, there are instances of path verbs functioning on their own, something characteristic of V-languages. Serial-verb languages, therefore, do not fulfil either of the patterns:

(1) Thai

(a)	*chán*	*won*	*jɔɔn*	*klàp*	*khâw*	*hɔɔŋ*
	I	circle	reverse	return	enter	room

'I returned circling back into the room.'

(Zlatev and Yangklang 2004: 163)

(b) *chán* *dəən* *won* *klà* *jɔɔn* *khâw* *paj*
I walk circle return reverse enter go
'I am walking in a circle, returning back inside.'
(Zlatev and Yangklang 2003: 166).

In examples (a) and (b) from Thai, manner and path are combined in different verbs. Thus, as Zlatev and Yangklang have proposed, a third typological category should be added: that of 'equipollently-framed languages', i.e. languages in which manner and path are expressed at equal terms. This category includes not only serial-verb languages but also bipartite verb languages and generic verb languages.

However, a tripartite typology does not seem to be the solution because, due to language-specific features, many of them cannot be ascribed to any of the three typological categories described above. Therefore, it is fair to conclude that it is necessary to place languages on a cline of path salience because, due to their specific morphosyntactic, lexical and cultural features, they may show both V- and S-language behaviour, even simultaneously.

2.1. Path

In the expression of motion events, path information must be somehow present. Being an obligatory element, it is not possible to look for a range of accessibility of path as a category, but we should consider path in relation to two different aspects:

- ❖ Number of path components into which a motion event can be divided (i.e. number of sub-trajectories that conform an overall trajectory).
- ❖ Distribution of these path segments into the different clauses present in the event (i.e. how compacted the path segments are).

Thus, according to Talmy's typology, V-language speakers tend to encode fewer path segments than S-language speakers in both speaking and writing. Moreover, in S-languages, path information is expressed in a more compacted way than in V-languages.

These previous two points are connected in some way and affect the use of manner of motion verbs. V-languages often make use of path verbs in order to express motion events because whenever a change of path direction occurs, they are subjected to the 'boundary crossing constraint'[132] and the use of manner of motion verbs is therefore affected by this. By contrast, S-languages show a single verb (usually a manner of motion verb) with different path satellites attached to it; as path information is conveyed through satellites, this pattern admits a more frequent use of manner of motion verbs. Besides, the different ways of codifying path information influences the attention to grounds (sources, goals, and milestones). As Slobin points out:

132 If there is some spatial crossing, it is not possible, at least in V-languages, to attach several ground elements to a single verb and therefore, these ground elements must be expressed through different verbs that convey path information themselves.

> Apparently language typology contributes to a typical level of event granularity. The determining factor seems to be the heavy use of a series of separate clauses in V-languages, as compared to the accumulation of path particles and prepositional phrases with a single verb in S-languages... [even in the case of] English-speaking children, at early ages, [it is possible to see a disposition] to describe complex paths, and in compact constructions. Across ages, the collection of complex locative elements in English exceeds the possibilities provided by path verbs in V-languages (2004: 239-40).

S-language speakers usually establish a one-per-one equivalence between ground elements and path segments, due to the possibility, inherent in these languages, to express several path components attached to a single verb. In S-languages, when several paths are introduced, more ground elements per clause can be expressed. V-languages, in turn, provide this information through static descriptions of setting.

3. RUN VERBS

3.1. Description and characteristics

This study is focused on *Run* verbs because they are generally used in sentences which provide movement information through the verb itself or through other parts of the sentence and can be considered one of the core elements in spatial semantics when expressing change of location.

Run verbs can be classified into two groups: general verbs (or 'neutral' in Slobin's terminology: *jump*, *walk*, *run*), and more expressive ones (*traipse*, *slink*). This latter type is much more elaborated in S-languages like English, in opposition to V-languages.

Therefore, this field is a salient one (both in memory and in verbal accounts), and has been enriched from Old English onwards. The speakers' need for 'finer' distinctions in this field has made them develop it extensively. One possible reason is the high codability of these verbs, which predisposes speakers to pay more attention to this domain. This codability lies in at least two factors: manner of motion being expressed by a single finite verb, and the high frequency of the lexical item itself (Slobin 2004: 237).

Thus, according to the literature, English manner of motion verbs are used in everyday language in different media, adding not only manner to the motion event but also vividness to the description.

Finally, it is important to remark that metaphorical examples were excluded for the present analysis, since they do not retain their original dynamic meanings. In many cases, these examples are connected with economy or politics, but also with other subjects such as time, art, mind and feelings, idioms, etc.

Table 1. Metaphorical cases of neutral manner of motion verbs.

Metaphorical cases of Neutral manner of motion verbs	*LOB*		*FLOB*
Walk	11		12
Run	156		199
Jump	11		9
TOTAL	**178**		**220**
		398 **TOTAL**	

3.2 Neutral Run verbs: Evidence from LOB and FLOB

If we look at the results from LOB and FLOB we can perceive a significant decrease in the use of neutral manner of motion verbs from the 60's to the 90's. Of the three, *Run* is the one that undergoes a more evident decrease, which coincides with an increase of its frequency in metaphorical uses.

Table 2. Frequency of manner of motion verbs in the LOB/FLOB corpora.

Neutral manner of motion verbs	N. of tokens *LOB*	N. of tokens *FLOB*
Walk	233	186
Run	131	95
Jump	50	31
TOTAL	414	312

Besides, it is important to mention that, as these verbs are neutral, manner is not really marked in them. Nevertheless, there exist other ways of emphasizing manner, which are attested in both corpora in combination with neutral manner of motion verbs, thus making the description more vivid. Manner adverbs (A:M), prepositional phases (PP:M), noun phrases (NP:M) and absolute participial clauses (A PP CL) are attested in both corpora.

Table 3. Emphasis on manner in LOB and FLOB.

Neutral manner of motion verbs	*LOB*				*FLOB*			
	A:M	**PP:M**	**NP:M**	**APP CL**	**A:M**	**PP:M**	**NP:M**	**APP CL**
Walk	32	25	3	25	22	22	4	26
Run	18	6	-	13	11	5	-	8
Jump	11	10	-	4	5	3	-	6
TOTAL	**61**	**41**	**3**	**42**	**38**	**30**	**4**	**40**
TOTAL AVERAGE	**147**				**112**			

This emphasis on manner decreases in terms of frequency from one period to the other, with the exception of noun phrases in which the results are very similar. Therefore, it seems that, in contrast with the 1960's, in the 1990's the manner component is less emphasized by speakers. They seem to prefer more simple sentences with respect to manner.

3.3 Path segments

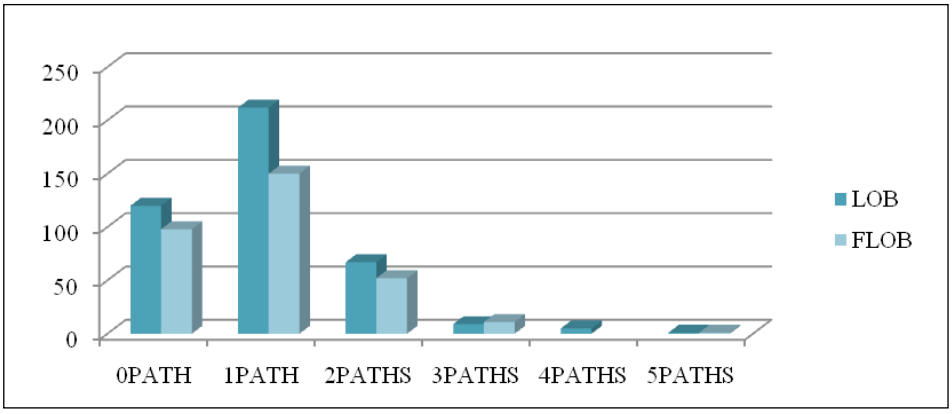


Figure 2. Frequency of path segments in the LOB/FOB corpora.

After analysing path segments we find 362 tokens of *Run* verbs + a path segment in both corpora, and 218 with no path segments at all. This contrasts with 119 instances of V + two path segments, 20 of V+ 3 path segments, 5 of V+ 4 path segments and 2 of V + 5 path segments. Thus, English speakers prefer to use less compacted and less complex constructions instead (placing English closer to V-languages). Besides, these three neutral manner of motion verbs show a decrease in the use of path segments from *LOB* to *FLOB*. According to this, we can consider that the codification of path segments has undergone change from the 60's to the 90's, bringing English closer to V-languages. Moreover, there are four occurrences of these verbs functioning as gerunds depending on other lexical motion verbs, which is a typical pattern shown by V-languages:

(2) Next day I persuaded the Anchorite to **come walking** with me in the same neighbourhood (*LOB*, K12: 113-4).

3.4 Ground elements

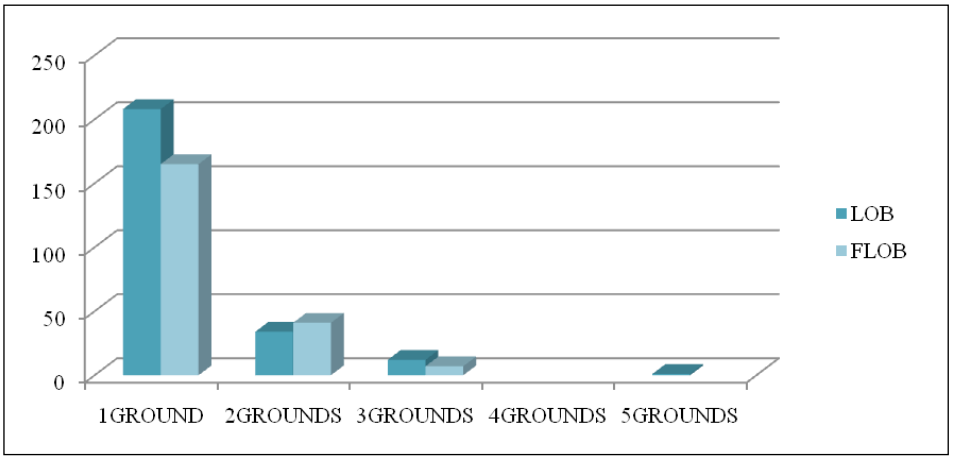


Figure 3. Instances of ground elements in the LOB/FLOB corpora.

With respect to ground elements, the introduction of one ground in the expression of a motion event seems to be the preferred option. Again, English speakers choose a less complex construction, going against the literature on motion events. Moreover, there seems to be no equivalence between paths and grounds of the examples under analysis.

Table 4. Frequency of the different grounds in the LOB/FLOB corpora.

	LOB			*FLOB*		
Ground elements	**SOURCE**	**GOAL**	**MILESTONE**	**SOURCE**	**GOAL**	**MILESTONE**
Walk	16	90	88	22	72	86
Run	4	43	54	4	24	40
Jump	10	8	4	7	9	4
TOTAL	**30**	**141**	**146**	**33**	**105**	**130**

Considering them separately, sources are the ground more underprivileged by speakers. We can perceive a lack of interest in mentioning initial locations, which are expressed by static descriptions instead. In most cases, what is introduced is a goal (highlighting the final destination) or a milestone (resource to express or describe the environment or space in which the figure is moving).

4. Conclusions and suggestions for further research

Taking all these aspects into consideration, it seems as if English were detaching itself from S-languages in the cline of manner and path salience, thus coming closer, at least in some respects, to V-languages, maybe due to language contact. However, the low number of tokens is not enough as to confirm a clear conclusion along these lines, although my intention is to deepen and develop this particular aspect.

In further stages of my research I will look at *Run* verbs in English and Spanish as represented in different corpora, seek for the historical and cultural reasons accounting for the differences between both languages analyze metaphorical cases of *Run* verbs and take a look at the influence of language on thought.

5. Acknowledgements

For generous financial support thanks are due to the Autonomous Government of Galicia (INCITE grant 08PXIB204016PR), the Spanish Ministry for Science and Innovation and the European Regional Development Fund (grant HUM2007/60706) and the Spanish Ministry for Education (FPU grant AP2008-00876). I would also like to express my gratitude to Teresa Fanego and Paloma Núñez-Pertejo for their suggestions while preparing this presentation.

6. References

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. Chicago: University Press.

Slobin, D. I. (2004). "The many ways to search for a frog: Linguistic typology and the expression of motion events". In S. Strömqvist & L. Verhoeven (eds.). *Relating events in narrative*: *Typological and contextual perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates, 219-257.

Strömqvist, S. and Verhoeven, L. (eds.). 2003. *Relating events in narrative*: *Typological and contextual perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Talmy, L. (1991). "Path to realization: A typology of event conflation". *Proceedings of the Berkeley Linguistics Society, 17*, 480-519.

Talmy, L. (2000). *Toward a cognitive semantics*: *Vol. II*: *Typology and process in concept structuring.* Cambridge, MA: The MIT Press.

Talmy, Leonard. 2007. Lexical typologies. In: Language typology and syntactic description, volume 3: Grammatical categories and the lexicon, Second edn., ed. by Timothy Shopen. Cambridge: Cambridge University Press.

Zlatev, J., & Yangklang, P. (2004). "A third way to travel: The place of Thai and serial verb languages in motion event typology". In S. Strömqvist & L. Verhoeven (eds.). *Relating events in narrative*: *Typological and contextual perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates, 159-190.

Disciplinary differences in the use of sub-technical nouns: a corpus-based study

María José Luzón Marco

Universidad de Zaragoza

Abstract

Research on academic vocabulary items has suggested that these words have specific behaviours related not only to the genre but also to the discipline (e.g. Hyland and Tse, 2007). In this research I use corpus-based methodology to analyse how a type of sub-technical vocabulary ("research nouns" and "discourse nouns") is used in two different disciplines (Applied Linguistics and Environmental Engineering). The purpose is to determine whether there are differences in the use of these nouns in both disciplines in terms of frequency, the lexico-grammatical patterns in which they occur, and the discourse functions associated with these patterns. The results provide corpus evidence of disciplinary variation in the frequency and collocational behaviour of sub-technical nouns. They also reveal that some of these nouns contribute to multi-word units that are part of the specific phraseology of the research paper in these disciplines.

Keywords: *academic vocabulary, sub-technical vocabulary, disciplinary variation, corpus analysis, phraseology*

Resumen

En esta investigación se analiza el uso de un tipo de vocabulario muy frecuente en el discurso académico (los sustantivos que hacen referencia al discurso y a la investigación) en dos disciplinas diferentes (Lingüística Aplicada e Ingeniería del Medio Ambiente). El objetivo es determinar si existen diferencias en el uso de dichos sustantivos en ambas disciplinas en cuanto a la frecuencia, a los patrones léxico-gramaticales en los que aparecen y a las funciones discursivas asociadas con dichos patrones. Los resultados muestran variación disciplinar en la frecuencia y el comportamiento colocacional de estos sustantivos sub-técnicos. Los resultados también revelan que algunos de estos sustantivos forman parte de unidades multipalabra que pertenecen a la fraseología específica del artículo de investigación en estas disciplinas.

Palabras clave: *vocabulario académico, vocabulario sub-técnico, variación disciplinar, análisis de corpus, fraseología*

1. Introduction

The need to help learners of EAP both to effectively read and produce academic texts has led to research on academic vocabulary, in an attempt to identify and teach the more frequent words specific to academic discourse (Coxhead, 2000). While the General Services List (GSL) (West, 1953) includes the basic vocabulary of English, the Academic Word List (AWL) (Coxhead, 2000) consists of 570 word families that are not in the GSL. Although the importance of the AWL for receptive purpose has been widely acknowledged, some researchers have questioned the idea that all the words in the list should be the focus of teaching for productive purposes in EAP. The efficiency of the list in EAP is questionable in the light of research that has revealed interdisciplinary variation in the use of academic vocabulary (Charles, 2007; Hyland and Tse, 2007, 2009). Hyland and Tse's (2009) analysis of the academic vocabulary in different disciplines showed that individual lexical items behave in different ways across disciplines and that these words are commonly part of 'lexical bundles' which also reflect disciplinary preferences. If the AWL is considered as a core vocabulary for students in any discipline, these students may be exposed to more vocabulary than they need and lack exposure to more specific vocabulary (Hyland and Tse, 2007; Paquot, 2007). In addition, the AWL has also been criticised for its exclusion of collocations. Drawing on research showing the existence of an EAP-specific phraseology, some researchers have claimed for the need to "introduce new words together with information on how to use them, especially their collocational and colligational environment" (Paquot, 2007). An interesting proposal in this sense is the compilation of a list of academic formulaic sequences that can be used for the teaching of academic speech and writing (Simpson and Ellis, 2010).

In this paper I study a type of non-technical vocabulary highly frequent in academic texts (which I will refer to as "research nouns" and "discourse nouns") in two different disciplines (Applied Linguistics and Environmental Engineering). Research nouns include: (i) nouns that refer both to the cognitive process involved in research (e.g. *hypothesis*) and to elements associated with the different phases of the research process itself (e.g. *result, data, analysis*), and (ii) general language nouns that are used by the writer to report on their experience of the real world and to talk about the object of research (e.g. *presence, feature*). Discourse nouns are nouns that refer to discourse elements (e.g. *example, comparison, study*). The purpose of this research is to determine whether there are interdisciplinary differences in the use of these nouns in terms of frequency, the lexico-grammatical patterns in which they occur, and the discourse functions associated with these patterns.

2. The study: corpus collection and data analysis

The corpus used in this study consists of 108 research articles, belonging to two different disciplines: Applied Linguistics (AL) and Environmental Engineering (EE). Each of the two subcorpora consisted of 54 articles published in three prestigious international journals in each field. The reference sections, tables and abstracts were deleted. The total number of words in the Applied Linguistics corpus was 405,830 (7,511 words per text

on average) versus 271,433 words in the Environmental Engineering subcorpus (5,026 words per text on average). Although both AL and EE are applied disciplines, AL is a soft discipline and EE is a hard discipline (Becher and Towler 2001) and thus differ in how research is conducted and knowledge is constructed. This difference will probably be reflected on a different behaviour of lexical items in both disciplines.

The methodological approach adopted in this research was both quantitative and qualitative. In the first stage of the quantitative data analysis, a frequency count of both corpora was conducted using Wordsmith Tools (Scott, 2004), to determine whether these nouns were used with the same frequency in both corpora. Both lists were manually examined to identify research/discourse nouns and two lists consisting of the 50 most frequent such noun lemmas in each discipline were compiled. As some of these words (such as *process,* and *result*) can be used as nouns or verbs, all concordances of these forms were carefully examined in order to eliminate verb occurrences, and final counts for each noun were re-calculated. From the lists of the most frequent noun lemmas I selected a few whose frequency was strikingly different in both corpora to make a detailed analysis of their collocational and colligational behaviour. Due to space limitations, only four of these lemmas are discussed here.

3. Results

Table 1 shows a combination of the two lists with the 50 most frequent sub-technical nouns in the AL corpus and the EE corpus. The resulting table includes 73 nouns, since there is not an exact match in the 50 top nouns in both corpora. The table lists the nouns with the highest frequency in **any** of the corpora in descending order. Therefore, nouns with a high frequency in one of the corpora and a low frequency in the other corpus occur high in the list. Since I was not dealing with comparable corpora in terms of extension, results were normalised per 10,000 words.

The lemmas whose relative frequency is higher in AL occur in bold type. The lemmas whose relative frequency is higher in EE occur in normal type.

Table 1. Noun frequencies per 10,000 words

Lemma	AL	EE	Lemma	AL	EE
STUDY	**23.08**	**11.4**	**TARGET**	**3.8**	**0.5**
DATA	18.15	18.7	**CATEGORIES**	**3.7**	**0.6**
RESULTS	8.94	18.6	FACTORS	3.6	3.7
ANALYSIS	**14.4**	**9.6**	**DIFFERENCE**	**3.6**	**2.7**
EXAMPLE	**12**	**3.6**	**ISSUE**	**3.5**	**3.1**
CONDITIONS	2.5	11.12	**APPROACH**	**3.5**	**3.1**
PROCESS	3.8	10.8	**FINDINGS**	**3.5**	**0.8**
STUDIES	**10.76**	**7.7**	**WAYS**	**3.5**	**0.3**
PRESENCE	0.8	10.6	PROCEDURE	2	3.5
EFFECT	3.9	10.5	**STRATEGY**	**3.3**	**0.5**
EXPERIMENTS	0.3	9.7	**ASPECTS**	**3.3**	**0.5**
LEVEL	**9.3**	**5.9**	EXPERIMENT	2.2	3.3
QUESTION	**9.2**	**0.7**	**BELIEFS**	**3.1**	**0**
SYSTEM	2.06	9.2	COMPARISON	2	3.1
FEATURES	**8.66**	**0.6**	**ANSWER**	**3**	**0.2**
INFORMATION	**8.57**	**5.5**	**REASON**	**2.8**	**1**
CASE	**5.7**	**8.5**	**PERSPECTIVE**	**2.8**	**0.2**
METHOD	2.1	7.2	**ISSUES**	**2.7**	**1.3**
STRATEGIES	**6.8**	**0.4**	LITERATURE	2.6	2.7
RATIO	0.9	6.7	**ANALYSES**	**2.6**	**1.3**
VALUES	1	6.7	**PURPOSE**	**2.6**	**0.9**
DIFFERENCES	**6.6**	**2.3**	**PURPOSES**	**2.5**	**0.7**
QUESTIONS	**6.1**	**0.5**	**FUNCTIONS**	**2.5**	**0.7**
CASES	**3.1**	**5.7**	**CRITERIA**	**2.5**	**0.4**
RESULT	3.1	5.5	CONDITION	2.5	1.1
SITUATIONS	**5,5**	**0,5**	**REASONS**	**2.4**	**0.5**
LEVELS	**5.3**	**4.1**	**THEORY**	**2.4**	**0.5**
RATES	2.2	5.2	**VARIABLES**	**2.3**	**0.7**
EFFECTS	3	4.9	CHARACTERISTICS	1.4	1.9
PROPERTIES	4.1	4.6	TECHNIQUE	0.3	1.8
EVIDENCE	**4.3**	**1.9**	MEASUREMENT	0.2	1.3
EXAMPLES	**4.3**	**0.4**	PROCEDURES	1.1	1.3
PROBLEM	**4.2**	**1.8**	TECHNIQUES	0.4	1.3
RESPONSES	**4.2**	**1**	CHARACTERISTIC	0.8	1.1
PROBLEMS	**4.1**	**1**	REQUIREMENTS	0.6	1.1
PROCESSES	1.5	3.8	QUANTITY	0.1	1
FUNCTION	2.1	3.8	**TOTAL**	**363.2**	**253.4**

Table 1 shows that the concentration of these nouns is higher in the AL corpus (the frequency of the top 50 noun lemmas per 10,000 words is 363.2 in the AL corpus and 253.5 in the EE corpus). If we examine the kind of nouns that are more frequent in each of the corpus, we can see that discourse noun lemmas prevail in the AL corpus (e.g. STUDY, EXAMPLE, QUESTION), while research noun lemmas are more frequent in the EE corpus (e.g. RESULTS, CONDITIONS, PROCESS, PRESENCE, EFFECT).

Differences in the use of nouns (and bundles including these nouns) across disciplines have been accounted for by differences in how knowledge is constructed in disciplines (Charles, 2007; Hyland, 2008). In natural science disciplines "knowledge is advanced by using experimental methods to provide evidence which will support or reject the hypothesis under investigation" (Charles, 2007: 209), hence the need for research nouns, which "impart a greater real-world, laboratory-focused sense to writing" (Hyland, 2008: 14). By contrast, soft disciplines build knowledge by constructing arguments and interpretation (Becher and Trowler, 2001). As Hyland (2008: 16) puts it: "while claims are often based on observations of real-world phenomena, knowledge is typically constructed as plausible reasoning rather than as nature speaking directly through experimental findings". This explains the importance of discourse nouns.

4. Analysis of some sub-technical noun lemmas

4.1. Example

As Table 1 shows, the relative frequency of the lemma EXAMPLE in both corpora is 12 (AL) and 3.6 (EE). The difference in the frequency of the lemma in both corpora is reflected in the difference in the number of clusters yielded by the concordance software. While in the EE corpus there were no 4-word clusters and only 3 3-word clusters (*for example the* (21 occurrences), *as an example* (6), *for example in* (5)), in the AL corpus there are 5 4-word clusters (*in the following example* (9), *an example of a* (7), *is an example of* (6), *for example in the* (6), *is an example from* (5)) and 23 3-word clusters, the most frequent ones being: *for example the* (43), *an example of* (26), *for example in* (22), *see for example* (16). Regarding 2-word clusters, the frequency between both corpora is strikingly different. For instance, there are 308 occurrences of *for example* in the AL corpus and 79 in the EE corpus. Although the frequency of *for example* is lower in the EE corpus than in the AL corpus, there are very few cases where EXAMPLE is not part of this exemplification marker in the EE corpus.

By contrast, EXAMPLE is part of a high number of clusters in the AL corpus. It collocates with the lemmas FOLLOWING (17), TYPICAL (9) (e.g. "as shown in the following example"), SHOWS (12), SHOWN (10), SEEN (7), ILLUSTRATES (5), ILLUSTRATED (5) (e.g. "as shown in example 1", "As example 3 shows", "This is illustrated in example 6"). In most cases these phrases are used by the writer to introduce a fragment of discourse that provides evidence of the point made by the writer. The collocations with adjectives such as *typical* or verbs such as *show, illustrate* help to present the example as persuasive evidence.

The high number of the lemma EXAMPLE in the AL corpus and the discourse functions of the patterns where it occurs are related to the heavy presence of exemplification in the humanities articles (Hyland, 2007) and its important role in the construction of knowledge in this discipline, where exemplification helps to discuss things in the real world: examples in soft knowledge fields help to contextualise and to "persuade the reader that the phenomenon actually exists".

Some of the clusters to which EXAMPLE contribute are psycholinguistically salient clusters (e.g. *an example of, see for example, in the following example, in this example, in example*+number*, a typical example*), and their high frequency suggests that it would be useful to include them in a formula list when teaching the discourse of Applied Linguistics, together with the function of these formulae.

4.2. Evidence

The relative frequency of *evidence* in both corpora is 4.2 (AL) and 1.9 (EE). In both corpora the noun collocates with the verbs *support, indicate, suggest* (*support* being the most frequent verb collocating left), and in the AL corpus it also collocates with *show*. A frequent pattern is *evidence to + support/indicate*.

(1) No reliable scientific evidence was ever reported to indicate an attributable risk of asbestos-related

The most frequent left verb collocate in both corpora is *provide* (with different lemmas), followed by *present*. Although the verbs with which EVIDENCE collocates in both corpora are similar, there is a clear difference as regards its semantic prosody[133]. In EE EVIDENCE has a negative prosody, which seems to be missing in AL. Of the 54 occurrences of EVIDENCE, it occurs in a negative context in 24 cases (44% of the cases) (16 of these occurrences are modified by *no*):

(2) a. The TPR profile **no longer shows evidence of** a bimetallic phase

b. Researchers have found **no evidence of** CuSO4 in aged Cu/CaO [21]

c. **No evidence of** sintering was found for any of the catalysts.

In the AL corpus there are only 18 out of 173 occurrences of EVIDENCE in a negative context ("little evidence", "the evidence does not support"). Evidence tends to occur in positive patterns, e.g. *we found some evidence, there is evidence that, this study/ paper provides evidence.* The high frequency of EVIDENCE in a negative context in EE could be related to the fact that hard sciences aim at the testing of hypotheses and predictions and to the wish to show that conclusions are based on empirical evidence: the existence or presence of something can only be claimed if there is evidence for it.

133 "When the usage of a word gives an impression of an attitudinal or pragmatic meaning, this is called a *semantic prosody*" (Sinclair 1999). Some words, like *cause* take on a negative value from its habitual co-occurrence with negative words.

The analysis of this noun shows that even if the lemma collocates with the same set of verbs in both corpora, it is used very differently and should therefore be taught differently in both disciplines, focusing on the relevant collocational patterns and discourse functions in each discipline.

4.3. Results

The relative frequency of RESULTS in both corpora is 8.9 (AL) and 18.6 (EE). In the EE corpus RESULTS occurs very frequently in the pattern "inanimate subject (research noun)+ active verb", where verbs such as *indicate, show, suggest* or *demonstrate* are used to link experimental or observational evidence to conclusions. Table 2 shows the most frequent collocates of RESULTS in the pattern "*results* (subject)+ active verb" in both corpora.

Table 2. Verb lemma collocates of RESULTS

	suggest	show	indicate	showed	demonstrate	support	Total
ES	26	23	21	13	**10**	0	93
AL	18	15	8	12	0	4	59

The number of occurrences of RESULTS with these verbs is much higher in the EE corpus than in the AL corpus. Another striking result is that, while there are 10 co-occurrences of RESULTS and DEMONSTRATE in the EE corpus, these lemmas do not collocate in the AL corpus.

Two findings from the context in which DEMONSTRATE occur are relevant here. First, in 4 out of the 10 occurrences of *results demonstrate*, the verb collocates with CLEARLY (e.g. "The results do clearly demonstrate"). In 5 out of the 10 cases, the collocation *results demonstrate* occurs in the conclusions, usually in the first sentence of this section. These findings suggest that *results demonstrate* signal a stronger link between evidence and conclusion than *results suggest/ indicate*, but it is still difficult to see the difference with *results show*. The higher use of *results demonstrate* in the EE corpus could again be related to the fact that hard sciences aim at testing hypotheses.

In the EE corpus RESULTS also occurs frequently in phrases with a location function in the research paper, i.e. used to indicate where results are to be found. In the EE corpus there are 22 co-occurrences of RESULTS with SHOWN and 13 with PRESENTED ("The results are shown in Table 1"). It also co-occurs with SHOWS (8) (e.g. "Table/Figure X shows the results of"). These patterns did not occur in the AL corpus.

Other collocations present in the EE corpus and absent in the AL corpus are *results obtained* (33 occurrences) and *results observed* (5).

(3) The effect of incident laser energy on the degradation of dye was investigated and the results obtained are presented in Table 1.

The past participles in the collocations *results obtained* and *results observed* make reference to the physical activity of doing research and thus help researchers to represent knowledge as deriving from the experimental activity.

4.4. Analysis

The relative frequency of *ANALYSIS* in both corpora is 14.4 (AL) and 9.6 (EE). Although research nouns tend to be more frequent in hard sciences than in soft sciences, this is not the case with the lemma ANALYSIS in the AL corpus. This high frequency of ANALYSIS in the AL can be accounted for by the fact that doing research in this discipline involves analysing discourse in several ways. When comparing how the lemma ANALYSIS is used in both corpora, the main difference is in the modifiers with which it collocates. The top most frequent collocates of ANALYSIS in both disciplines are as follows:

> AL : Genre (39), data (28), quantitative (16), discourse (15), **statistical (12),** conversation (11), qualitative (10)
>
> ES : **Statistical (14),** chemical (11), GC (9), asbestos (5)

As can be seen the only common collocate is *statistical*. Both in the AL corpus and the EE ANALYSIS occurs frequently in multi-word units that are in fact technical terms in these disciplines: *genre analysis, discourse analysis, conversation analysis, chemical analysis, GC analysis, asbestos analysis*. There are other multi-word units, like *data analysis, quantitative analysis or qualitative analysis*, where both terms are general and do not make reference to concepts that are specific of AL, but they are certainly collocations typical of this discipline, and absent in others, which give a clear insight of the kind of research that applied linguists are expected to do.

5. Conclusions

The high frequency of research and discourse nouns in the corpora analysed in this study suggests that they are an important device to construct knowledge and develop an argument. However, this study has shown that particular research and discourse nouns are more relevant in some disciplines than in others.

This study provides corpus evidence of disciplinary variation in the frequency of some of these nouns, in their lexico-grammatical patterns, in the frequency of the bundles where they occur and in the rhetorical functions of the sentences where these bundles occur.

These results have clear implications for EAP teachers. They provide support for previous research that shows the need for discipline-specific academic vocabulary lists (Hyland and Tse, 2007; Martínez et al., 2007). There are many research and discourse nouns highly frequent in one of the disciplines, which are, however, rare or much less frequent in the other discipline and therefore should not be given the same relevance when teaching both disciplines. In addition, the analysis of individual lexical items has shown that these nouns

occur in discipline-specific collocations or in clusters with specific rhetorical functions. This involve that the teaching of these items should be based on a previous analysis of their collocational and colligational patterns in a discipline corpus.

6. References

Becher, T. and Trowler, P. (2001). *Academic Tribes and Territories* (2nd ed.). Buckingham: Open University Press/SRHE

Charles, M. (2007). Argument or Evidence? Disciplinary Variation in the Use of the Noun "that" Pattern in Stance Construction. *English for Specific Purposes*, *26*(2), 203-218.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34* (2), 213–238.

Hyland, K. (1999). Academic attribution: Citation and the construction of disciplinary knowledge. *Applied Linguistics, 20 (3),* 341-367.

Hyland, K. (2007). Applying a gloss: exemplifying and reformulating in academic discourse. *Applied Linguistics, 28*(2), 266-285.

Hyland, K. (2008). 'As can be seen: Lexical bundles and disciplinary variation'. *English for Specific Purposes, 27*(1), 4-21.

Hyland, K. and Tse, P. (2009). Academic lexis and disciplinary practice: corpus evidence for specificity. *International Journal of English Studies*, *9* (2), 111-130.

Hyland, K. and Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly, 41*(2), 235–253.

Martinez, I. A., Beck, S. C., & Panza, C. B. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. *English for Specific Purposes*, *28*(3), 183-198.

Paquot, M. (2007) Towards a productively-oriented academic word list. In J. Walinski, K. Kredens and S. Gozdz-Roszkowski (Eds.), *Corpora and ICT in language studies* (pp. 127–140). Frankfurt am Main: Peter Lang.

Simpson-Vlach, R., & Ellis, N.C. (2010). An academic formulas list: new methods in phraseology research. *Applied Linguistics 31*(4), 1-26.

Scott, M. (2004). *WordSmith Tools*. Oxford: OUP.

Sinclair, J. (1999). *Concordance tasks.* Online at http://www.twc.it/happen.html.

West, M. (1953). *A general service list of English words.* London: Longman.

Voice-overs in British TV ads: Characterising a written-to-be-spoken corpus[134]

Barry Pennock
Universitat de València

Abstract

In this paper I will analyse the transcription of the voice-overs which form part of the MATVA corpus of UK television commercials in order i) to identify the salient characteristics of this type of discourse compared to both a spoken and a written reference corpus and ii) to look more closely from a pragmatic point of view at the use of the second person possessive determiner your*, which is a very frequent token in my corpus. My first hypothesis is that voice-overs in TV ads have characteristics that are different from all other registers. My second hypothesis is that the vocabulary of voice-overs is closer to spoken discourse than written discourse. My third hypothesis is that voice-over discourse in TV ads will embody pragmatic strategies originated by the fact that the strategic goal of TV ads is to promote or sell products and services and that this will be reflected in the vocabulary of this register. In order to discover if my hypotheses are correct, I will, on the one hand, use the Wordsmith wordlist, keyword, and concordance tools and, on the other, carry out a pragmatic analysis based on Brown and Levinson (1987).*

Keywords

advertising, television, linguistic politeness, multimodal, pragmatics, spoken corpus, voice-over

Resumen

En este trabajo analizaré la transcripción del discurso de la voz en off que forma parte del MATVA corpus de anuncios televisivos del Reino Unido para a) identificar sus características comparándolo tanto con un corpus de referencia hablado como con un corpus de referencia escrito y b) analizar el uso del posesivo de segunda persona your desde el punto de vista pragmático. Mi primera hipótesis es que las voces en off de los anuncios televisivos tendrán características distintas a las de los demás registros. Mi segunda hipótesis es que las características propias de las voces en off están más relacionadas con el discurso oral que con el discurso escrito. Mi tercera hipótesis es que las voces en off de los anuncios televisivos contendrán estrategias pragmáticas originadas por el hecho de que la meta estratégica de los anuncios televisivos es promocionar o vender productos y servicios y que este hecho se verá reflejado en el vocabulario de este registro. Para descubrir si mis hipótesis son correctas, por una parte, utilizaré las herramientas del programa Wordsmith para crear una lista de palabras organizadas de más a menos frecuentes, una lista de palabras claves y una concordancia y, por otra, emplearé el análisis pragmático basado en Brown y Levinson (1989).

Palabras Clave

televisión, anuncios, corpus hablado, multimodal, pragmática, cortesía lingüística, voz en off

134 This research has been made posible thanks to the project "Efectos Pragmático-cognitivos de los elementos paralingüísticos y extralingüísticos sobre la audiencia en los anuncios de televisión en lengua inglesa" (UV-AE-10-24541) financed by the *Universitat de València* through the *Projectes d'Investigació Precompetitius* scheme.

1. Introduction

In this paper I will analyse the transcription of a series of voice-overs, which are part of a corpus of UK television commercials, in order i) to identify the salient characteristics of this kind of discourse compared to both a spoken and a written reference corpus and ii) to look more closely at the use of the second person possessive determiner *your*. My first hypothesis is that voice-overs in TV ads have characteristics that will be different from all other registers. My second hypothesis is that the vocabulary of voice-overs is closer to spoken discourse than written discourse. My third hypothesis is that voice-over discourse in TV ads will embody pragmatic gambits originated by the fact that the strategic goal of TV ads is to promote or sell products and services and that this will be reflected in the vocabulary of this register (Pennock-Speck & del Saz-Rubio, 2006). In order to discover if my hypotheses are correct, I will, on the one hand, analyse the corpus with the *Wordsmith* wordlist, keyword, and concordance tools (Scott, 2010) and, on the other, I will carry out a pragmatic analysis of the corpus based on Brown & Levinson (1987).

2. Description Of Corpus

My corpus has been taken from the larger MATVA Corpus (Multimodal analysis of television ads) which consists of 4 days of daytime television recorded from ITV 1, the oldest and most popular commercial television channel, on the 19th and 20th of March and the 24th and 25th of June from approximately seven am to around six pm. What I will be analysing here is the finished June section of the corpus which consists of the actual ads themselves in mpg format, the linguistic information extracted from the ads, and the descriptive, non-linguistic information on the ads.

The linguistically analysable material in the corpus is made up of the transcription of on-screen text, voice-over text, dialogue and testimonial text, and song lyrics. There is also additional ancillary information such as that found in Table 1:

Table 1: Ancillary information

Ancillary Information	**Example**
unique ad id	NestleCerealsVIPsBalcony01
broadcast date	24/06/2009
broadcast time	08.00-12.30
duration	10 seconds
previous programme	GMTV
following programme	GMTV Weather
product name	Nestlé Cereals
product type	Food
sub-product type	-
voice-over description	Gender: Female
	Age: Over 40
	Accent: RP
	Voice quality: happy, optimistic
ad description	Cartoons. A group of cereal boxes appear on a balcony and a crowd of cereal boxes cheer at them in the street. Then we hear the sound of a storm and the group at the balcony go back inside while the crowd cheers and claps.

The MATVA corpus is designed so that it can be analysed from multiple perspectives – including the analysis of linguistic, paralinguistic and extralinguistic elements. However, in the research I present here, due to space constraints, I will be focusing exclusively on voice-over discourse. I have a strict definition of what constitutes a voice-over, that is, I only count disembodied voices as such. If a voice is heard off-screen and then the owner of the voice appears, I do not classify it as a voice-over but as an intervention by a character or a person giving a testimonial.

3. Corpus Analysis of Voice-Overs

My June corpus contains a total of 637 TV ads. Many of these, as one would expect, are duplicates; for example, there are over 30 *Sky Box* ads. In a study of impact, taking into account duplicated ads would be useful. However, as what I am interested in here are the special characteristics, if any, of voice-over discourse, duplications have to be eliminated as otherwise certain words would be more frequent than others merely because the ad in which they appear is repeated several times during the day. Thus, after eliminating duplicated ads, I was left with a total of 276 ads –62 did not include voice-overs while 214 did. However, some of these 214 ads were different from a visual point of view but contained identical voice-overs. I therefore eliminated all but one of each of these. For example, there are several distinct ads promoting the airline company *Flybe*. One features flights to Scotland, another, flights to skiing resorts, yet another, flights to English cities, and so on but the voice-over is the same for all of them: "Change your outlook with your local airline. *Flybe.* Sponsors of West Country Weather". Eliminating all but one of these ads, and others with identical voice-overs, left a total of 180 different commercials in all.

To avoid as much repetition as possible I also decided to eliminate identical phrases within ads. For example, the *Argos-end-of-catalogue-sale* commercials, which are both visually different and feature different products, begin and end with identical phrases. So, I removed all but one instance of the phrases which appeared in all of them, namely: *The incredible Argos end of catalogue sale is now on. Don't miss out on hundreds of great deals.* The same process was carried out in with other ads.

Once the voice-over texts were selected and edited, I used the Wordsmith programme to analyse the corpus. The statistics offered by *Wordsmith* show that the corpus contains 6,310 tokens. Of these 6,115 are used in the word list[135]. The total number of distinct tokens is 1,910. I generated a wordlist showing the most frequent words (Table 2) but without implementing the *Wordsmith* stoplist function so tokens like *the, and, to, a, of, for, with,* etc. appear. My list is similar to the list made up of the first forty tokens taken from the CANCODE corpus of spoken English published in McCarthy (1999: 236). I have included in parentheses the rank number of the tokens from McCarthy's list next to the tokens in my list (see Table 2). Fourteen of the words in both lists coincide. Nevertheless, there are some conspicuous absences among the most frequent words in my list when compared to the CANCODE list. For instance, *I* is found only once and is ranked 1,220 in my wordlist whereas it is ranked second in McCarthy's. Other very frequent words in

135 Numbers were not included in the wordlist.

McCarthy's list such as *know (265); they (168);well (224)* are not very frequent at all in mine and others do not appear at all: *yeah.* Some differences may be due to the fact that, as Stubbs (2002: 42) points out, "raw frequency lists often have odd gaps". The "odd gaps" that may appear in my corpus are probably due to its small size. However, the main reason is probably because voice-overs are effectively non-interactive monologues unlike the discourse found in the CANCODE corpus.

In order to shed more light on the characteristics of voice-over discourse I compiled two keyword lists using written and spoken corpora included in the ICAME collection. The written reference corpus consisted of the Brown corpus (1961) and the Wellington Corpus of Written English (1986 to 1990) while the spoken reference corpus comprised the Wellington Corpus of Spoken New Zealand English (1988 to 1994) and the Corpus of London Teenage Language (1993).

Table 2: MATVA June Corpus of Voice-overs

N	Word	N	Word	N	Word	N	Word
1	THE (**1**)	12	ON (**23**)	23	UP (**-**)	34	ON (**23**)
2	AND (**3**)	13	NOW (**-**)	23	OUR (-)	35	NOW (**-**)
3	TO (**5**)	14	IT (-)	25	THIS (**38**)	36	BUT (**18**)
4	YOU (**4**)	15	FROM (-)	26	ALL (**36**)	37	OUT (-)
5	A (**6**)	16	AT (-)	27	OR (-)	38	ARE (-)
6	OF (**10**)	17	CAN (-)	28	CALL (-)	39	GO (-)
7	FOR (**29**)	18	FREE (-)	29	GET (-)	40	HELP (-)
8	WITH (-)	19	JUST (-)	30	IT'S (**16**)	41	PRICE (-)
9	YOUR (-)	20	NEW (-)	31	SO (**20**)	42	HALF (-)
10	IN (**-**)	21	ONE (-)	32	WE (**22**)	43	SUMMER (-)
11	IS (**19**)	22	UP (**-**)	33	AN (-)	44	LIKE (**30**)

Getting the written and spoken reference corpora ready to be used as reference corpora was a necessary but laborious task. In some cases it was necessary to eliminate the *less-than/greater-than* brackets(< >) around primary text and in other cases it was necessary to add them on either side of the many annotations so this type of text would be ignored by *Wordsmith*. If this had not been done, *Wordsmith* would have ignored the primary text that we are interested in and analysed the annotations.

Let us first turn our attention to the written keylist in Table 3. It is no surprise that brand names like *L'Oreal* and *Dettol* and charities like the *NSPCC* from my corpus would stand out compared to both the written and spoken reference corpora. Other words are positively key because they either did not exist when the reference corpora were being compiled or were extremely rare: *DVD*, *online*. Among the lexical items, words like *free, call, now, plus, price, pounds, skin, sale,* and *mascara* are noticeable. The frequency of *summer* is due to the fact that the ads were broadcast in June. The items *skin* and *mascara,* on the other hand, reveal the frequency of cosmetics ads in our corpus. In all, the positively key lexical items are the kind one would expect in a TV ad corpus. With regard to function words, both *your* (rank 1) and *you* (rank 2)*,* are positively key while the pronouns *I, he, his* and *the* are negatively key.

Table 3: Written/MATVA keylist

N	Key word	Keyness	N	Key word	Keyness
1	YOUR	263.94	12	SUMMER	91.80
2	YOU	232.62	13	POUNDS	83.21
3	FREE	188.74	14	DETTOL	81.01
4	L'OREAL	150.46	15	JUST	75.76
5	UK	131.84	16	SKIN	73.91
6	CALL	122.53	17	TV	71.67
7	NOW	118.09	18	DVD	69.44
8	COM	116.16	19	ONLINE	69.44
9	NSPCC	104.16	20	SALE	66.86
10	PLUS	98.58	21	MASCARA	66.61
11	PRICE	94.45	22	IT'S	65.56

Turning now to the spoken keyword list, among the 158 words that are featured, the pronoun *you* does not appear at all thus indicating that it does not stand out in our corpus compared to the spoken reference word list. This suggests that the use of *you* is similar in both my corpus and the spoken reference corpus. However, the item *your,* ranked 8th, is positively key in the spoken corpus whereas the pronouns: *I* and *he* are negatively key. Interestingly, numbers are ranked first in the spoken keylist probably because in advertising, unlike the spoken reference corpus, the mentioning of prices is essential in the promotion and selling process. Among the individual lexical items which are positively key we find: *free, price, UK, summer, call, skin, plus, visit, sale, help(s), eye, TV, Paris* and *mascara.* All of these, except *UK,* are also found in the written keylist. Once more, most of these items seem to be typical or even stereotypical items in promotional discourse.

Table 3: Spoken/MATVA keylist

N	Key word	Keyness	N	Key word	Keyness
1	#	895.89	12	VISIT	84.67
2	FREE	245.89	13	SALE	84.40
3	L'OREAL	141.98	14	COM	80.87
4	PRICE	120.85	15	DETTOL	76.44
5	UK	113.10	16	HELP	71.06
6	SUMMER	108.09	17	EYE	68.92
7	NSPCC	98.29	18	TV	68.27
8	YOUR	97.65	19	DVD	65.52
9	CALL	94.38	20	PARIS	65.25
10	SKIN	90.28	21	HELPS	65.06
11	PLUS	87.16	22	MASCARA	64.25

4. Pragmatic Analysis

As we have seen, the function word that stands out in both the written and spoken keylists is *your*. It is much more common in our corpus than in the written reference corpus and

the fact that it appears in the spoken reference corpus keylist when *you* does not, makes it even more conspicuous. Due to space constraints we will limit ourselves to looking exclusively at just this token in more detail. I feel that *you* is frequent for a reason and that reason is its pragmatic significance. In effect, my results show that *your* is normally an instantiation of one of Brown & Levinson's (1987: 103) strategies to mitigate a face threatening act (FTA). One of these is positive politeness strategy 1:

> "Notice, attend to H[136]" in which "S should take notice of aspects of H's conditions (noticeable changes, remarkable possessions, anything which looks as though H would want S to notice and approve of it)."

What better way of drawing attention to a person's looks, needs, wants, or possessions than the second person possessive determiner? Of course, *your* does not occur in isolation. Practically in every one of the 78 instances of *your* its redressive effects are augmented by phrases that reinforce the Strategy-1 interpretation although we also find other politeness strategies such as positive politeness Strategy 9: "Assert or presuppose S's knowledge of and concern for H's wants (Brown and Levinson: 125). The following examples of this strategy are from the *Computeach* ad in my corpus: *Are you worried about your future?* and *Do you feel your back is against the wall?***.** Strategy 8: "Joke" (Brown & Levinson: 124) is found in a *Gillette Fusion* ad: *Sometimes you need a little push to let go of your Mach3 razor* which is part of a visual joke in which famous sportsmen use golf balls, tennis balls and soccer balls to remove an old razor from a man's hand. I only found two examples in which *your* was not reinforced by further attention to the hearer's face: *Ask your pharmacist about Alli* and *Type your postcode into blood.co.uk.* I also only found two examples of *you* as a part of negative politeness redress.

Negative politeness strategies are quite rare in TV ads and the only clear examples appear in ads for the NSPCC charity. The first is negative politeness strategy 4: "Minimize the imposition" (Brown and Levinson: 176): *They're just children and they need your help.* I classify the next as negative politeness strategy 5: "Give deference" (Brown & Levinson: 178): *Please open your eyes and your heart.*

What makes it clear that the ad creators are using politeness strategies to redress FTAs is that unlike other genres like peer criticism (see Pennock-Speck & Clavel-Arroitia forthcoming), there is no doubt that addressing the viewer directly embodies an intentional politeness strategy. After all, in most, if not all of the examples in our corpus in which *your* is employed, other words or phrases could have been used. I have brought up the issue of intentionality at this point as it is found to be problematic in other genres (Culpeper, 2009; Garcés-Conejos Blitvich 2010). The creators of ads deliberately mitigate FTAs as the whole purpose of TV advertising is to impress the viewer favourably. The main strategic goal of ad creators is the promotion and sale of products and services while the use of certain types of linguistic messages is simply a tactical goal to achieve said strategic goal (Pennock-Speck & del Saz-Rubio, 2006). In a multimodal genre like TV ads, words, of course, are just one way of achieving the advertiser's goals, voice-quality, images, narrative and music also embody tactical goals as we saw in the *Gillette Fusion* ad (see Pennock-Speck, 2006, del Saz-Rubio & Pennock-Speck 2009).

136 Of course, S would be the creator of the ad and H, the viewer.

5. Conclusions

My first hypothesis, i.e., that voice-overs in TV ads are different from other registers, seems to be correct given the positive keyness of tokens such as *free, new, now, help, price, half* which most speakers would recognize as belonging to a promotional register. My second hypothesis was that my scripted corpus would be nearer to spoken discourse. The frequency of *you*, appearing in the top five in the MATVA corpus, the CANCODE corpus and the spoken demographic and task-oriented section of the BNC (see Fuster-Márquez & Pennock Speck, 2009: 59), as well as in my own reference corpus seems to show that the scripted monologues that make up voice-overs are indeed closer to spoken discourse when compared to my written reference corpus and to the written section of the BNC. However, the very infrequent appearance of the pronouns *I, he, she* and *they* make the MATVA voice-over corpus very different not only from oral but also from written discourse. The relative frequency of *you* is due, without doubt, to the onus on *speaking* directly to the audience. The scarcity of the third person personal pronouns stems from the fact that voice-over discourse is directed almost exclusively, and in a direct way, at the audience and their needs and wants.

This brings us to the most salient item in the MATVA corpus, *your,* which goes a long way to proving my third hypothesis, that voice-over discourse embodies pragmatic strategies of persuasion. Firstly, because the great number of instances of this possessive is connected to the strategic goal of TV ad discourse which is to make people aware of, or purchase products and services which, in turn, entails engaging in direct communication with the audience. Secondly, because it is a way of conveying the positive politeness strategy of claiming common ground that is often instantiated through strategy 1. I have also shown that *your* is usually part of, or reinforces, other positive politeness strategies which are typical of this kind of genre. However, the scarcity of *your* in negative politeness strategies is due to the indirect nature of the prevalent soft-sell strategies in TV ads –it is extremely rare to push a product in an overt way.

To sum up, corpus analysis is an invaluable tool in order when it comes to carrying out pragmatic analysis on a relatively large number of voice-overs. It would have been impossible to prove or disprove my hypotheses or to discover the politeness-strategic role of *your* any other way.

References

Brown, P. & Levinson, S. C. (1987). *Politeness: Some Universals in Language Usage.* Cambridge: Cambridge University Press**.**

del Saz-Rubio, M. M. & Pennock-Speck, B. (2009). Constructing female identities through feminine hygiene TV commercials. *Journal of Pragmatics* 41: 2535–2556.

Culpeper, J. (2009). Impoliteness: Using and Understanding the Language of Offence. Retrieved from ESRC project website: http://www.lancs.ac.uk/fass/projects/impoliteness/.

FUSTER, MÁRQUEZ, M. & PENNOCK-SPECK, B. (2008). The spoken core of British English: A diachronic analysis based on the BNC. *Miscelánea. A Journal of English and American Studies,* 37: 53-74.

GARCÉS-CONEJOS BLITVICH, P. (2010). The status quo and quo-vadis of impoliteness research. *Intercultural Pragmatics* 7, 4: 535–559.

MCCARTHY, M. (1999). What constitutes a basic vocabulary. *Studies in English Language and Linguistics,* 1: 233-249.

PENNOCK-SPECK, B. (2006). Styling the voice, selling the product. In C. Mourón-Figueroa, & T. I. Moralejo Gárate, (Eds.), *Contrastive Linguistics: Proceedings of the 4th International Contrastive Linguistics Conference. Santiago de Compostela, September, 2005* (pp. 973-980). Santiago: Servicio de Publicacións da Universidade de Santiago de Compostela.

PENNOCK-SPECK, B. & CLAVEL-ARROITIA, B. (In press). Mitigating the force of criticism in student peer reviews. In M. D. García Pastor (Ed.) *English as a Foreign Language: Proposals for the Language Classroom.* Catarroja: Perifèric Edicions.

PENNOCK-SPECK, B. & DEL SAZ RUBIO, M. M. (2006). A genre approach to goals and their implementation applied to a TV programme for the Virginia Farming Community. *Ibérica (*Revista de la Asociación Europea de Lenguas para Fines Específicos) 11: 7-28.

PENNOCK-SPECK, B. & DEL SAZ-RUBIO, M. M. (2009). Voice-overs in Standardized English and Spanish Television Commercials. *Atlantis.*

SCOTT, MIKE (2010). *WordSmith Tools* (version 5). Liverpool: Lexical Analysis Software.

STUBBS, M. (2002). *Words and Phrases: Corpus Studies of Lexical Semantics*. Blackwell Publishing: Oxford.

Un corpus de dietarios de viajes: los límites entre el dialecto y el idiolecto[137]

María Pilar Perea
Universitat de Barcelona

La edición de todos los dietarios de viajes que el dialectólogo mallorquín Antoni M. Alcover (Manacor 1862 - Palma 1932) publicó entre 1901 y 1924 ha dado lugar a un corpus de carácter biográfico y documental que sobrepasa el millón de palabras. Las fuentes provienen de los relatos extraídos de los 14 tomos de la primera época del Bolletí del Diccionari de la Llengua Catalana *(1901-1926) y de las narraciones que aparecieron en publicaciones periódicas como el* Diario de Mallorca *(1901-1902) y* La Aurora *(1912 y 1913).*

En este estudio se presentan las dificultades que presenta un corpus de estas características y se analizan los elementos fonéticos, morfológicos y léxicos más destacados, propios de la variedad mallorquina, los cuales experimentan variación a lo largo del período cronológico que abraza la publicación de los dietarios. Además de los rasgos dialectales, que el narrador adapta a los potenciales lectores de sus textos, con la ayuda del corpus se intentan definir los rasgos idiolectales que caracterizaron la escritura del autor.

Palabras clave: corpus, dietarios de viaje, catalán, dialecto, idiolecto

A documentary and biographical corpus of more than one million words has been created from the travel diaries that dialectologist Antoni M. Alcover (Manacor 1862 - Palma 1932) published between 1901 and 1924. The materials are drawn mostly from the 14 volumes of the first period of the Bolletí del Diccionari de la Llengua Catalana *(1901-1926) and from Majorcan journals such as the* Diario de Mallorca *(1901-1902) and* La Aurora *(1912 and 1913).*

This study will present the difficulties within the treatment of this dialectal corpus and analyze its main phonetic, morphological and lexical characteristics, which changed across the period of publication of the diaries. The corpus offers not only information about Alcover's dialect, which he adapted to readers' depending on their place of origin, but also his own idiolectal features.

Keywords: corpus, travel diaries, Catalan, dialect, idiolect

137 Este trabajo se adscribe en el proyecto de investigación "Portal de léxicos y gramáticas dialectales del catalán del siglo XIX" (FFI2010-18940 (subprograma FILO)), financiado por el Ministerio de Ciencia e Innovación.

1. Introducción

La edición de los dietarios de viajes[138] que el dialectólogo mallorquín Antoni M. Alcover (Manacor 1862 - Palma 1932) publicó entre 1901 y 1924 ha dado lugar a un corpus de carácter biográfico y documental que sobrepasa el millón de palabras. El concepto "dietario" reúne no sólo los ocho relatos que llevan ese título, sino también las crónicas de las impresiones experimentadas en diversas excusiones, que fueron encabezadas con los nombres de "impresiones de viaje", "excursiones" o "escapadas". El objetivo de los viajes era estudiar las formas vivas de la lengua para obtener materiales que le permitieran redactar su *Diccionari català-valencià-balear* (DCVB). Los dietarios incluyen descripciones de los lugares que visitó, las cuales son muy frecuentes en las narraciones de los viajes realizados en el extranjero. Adicionalmente, algunos textos reúnen comentarios sobre las formas dialectales usadas en las localidades donde llevó a cabo encuestas, tanto desde el punto de vista fonético como morfológico y sintáctico. Este hecho incrementa la complejidad del etiquetado de los materiales.[139]

En este estudio se muestran las características que presenta un corpus dialectal de este tipo así como las dificultades en su procesamiento y se analizan los elementos fonéticos, morfológicos y léxicos más destacables, propios de la variedad mallorquina, que experimentaron variación formal a lo largo del período cronológico en que se publicaron los dietarios. Además de definir las características dialectales, el corpus muestra los rasgos idiolectales que caracterizan la escritura del autor.

Junto con la introducción y las conclusiones, el trabajo incluye cinco apartados: *a*) la descripción de los datos; *b*) la cuantificación de los materiales y la determinación de sus principales características lingüísticas; *c*) la manifestación de los rasgos dialectales más sobresalientes; *d*) la sugerencia de los rasgos que pueden considerarse idiolectales; y *e*) la constatación de la imprecisa delimitación entre dialecto e idiolecto.

2. Descripción de los materiales textuales

El corpus de viajes está formado por 16 documentos que fueron redactados entre 1901 y 1924. Los textos tienen una extensión variable: doce se han extraído de los diferentes tomos del *Bolletí del Diccionari de la Llengua Catalana* (BDLC),[140] revista editada por el propio Alcover, y los cuatro restantes corresponden a publicaciones efectuadas en periódicos locales. Exceptuando el dietario que se publicó en castellano en la *Gaceta de Mallorca*, en 1907, y que reproduce en parte lo que apareció en catalán en el tomo V del BDLC, los otros tres dietarios, con el título de impresiones de viaje, se editaron en el *Diario de Mallorca* en 1901 y en diversos números del semanario de Manacor, *La Aurora*, que reúnen los viajes que realizó al extranjero en 1912 y 1913.

138 En 2002 se publicaron en CD-ROM los dietarios (Perea, 2002) y entre 2003 y 2004 se editaron los 19 tomos del *Bolletí del Diccionari de la Llengua Catalana* (Perea, 2003 y 2004). Actualmente estos materiales forman parte del *Portal Antoni M. Alcover* (alcover.iec.cat). Otros dietarios anteriores se incorporarán próximamente al corpus.

139 Ueda y Perea (2010) presentaron un método de lematización del tomo V del *Bolletí del Diccionari de la Llengua Catalana* (1908), que contiene el "Dietari de la meua exida a Alemania y altres nacions durant l'any del Senyor 1907".

140 Cf. Los tomos del BDLC están agrupados, en general, de manera bianual. Cf. en Ueda & Perea (2010) una descripción más completa de estos materiales.

3. Cuantificación y características lingüísticas de los documentos

En este trabajo se analizan los quince documentos que Alcover redactó en catalán. El corpus cuenta hasta ahora con 946.108 palabras,[141] de las cuales 45.720 son diferentes.

La Tabla 1 muestra los elementos cuantitativos más relevantes de cada documento.

Tabla 1. Cuantificación de los elementos más relevantes de los documentos

Documents	Period	tokens (running words) in text	tokens used for word list	types (distinct words)	sentences
DM_1901	1901	101.604	101.403	10.832	4.161
BDLC1	1901-1902	10.661	10.525	1.827	606
BDLC2	1904-1905	9.383	9.300	2.338	327
BDLC3	1906-1907	56.917	56.641	7.168	3.166
BDLC05	1907	179.767	177.570	13.604	9.291
BDLC4	1908-1909	6.754	6.687	1.884	318
BDLC6	1910-1911	5.367	5.318	1.623	313
Aurora1912	1912	102.867	102.017	9.532	4.323
BDLC7	1912-1913	29.328	29.167	4.887	1.612
Aurora1913	1913	184.687	184.072	13.997	7.090
BDLC8	1914-1915	12.007	11.909	2.969	548
BDLC10	1918-1919	43.830	43.520	6.290	2.373
BDLC11	1920	83.200	82.639	11.584	4.142
BDLC12	1921-1922	98.792	97.673	11.028	5.260
BDLC13	1923-1924	20.944	20.665	3.465	810
Overall	1901-1924	946.108	939.106	45.720	44.340

Existen diversos aspectos del corpus que dificultan los procesos de lematización y de etiquetaje morfológico: 1) la grafía prefabriana —es decir, no estandarizada— tiene como consecuencia que una misma palabra presente mucha variación ortográfica,[142] hecho que, en contrapartida, muestra aspectos fonéticos y morfológicos de carácter dialectal, que habrían quedado ocultos si se hubiera utilizado una grafía convencional; 2) los numerosos topónimos y antropónimos, especialmente extranjeros, o palabras no catalanas, que el autor usa cuando pretende reproducir los idiomas con los cuales entra en contacto, tienen un papel cuantitativamente relevante, e incluyen la problemática del tratamiento de los nombres propios; 3) las cuantiosas y valiosas informaciones dialectales, especialmente cuando están transcritas fonéticamente, provocan que el programa *WordSmith* no reconozca esos símbolos e incluya, en el recuento, palabras truncadas, al substituirse la transcripción fonética por signos de interrogación.

141 El corpus se ha tratado con el programa *WordSmith 5.0.0.334*; en el futuro se lematizará y etiquetará completamente.

142 Por ejemplo, la presencia, ausencia o diversa orientación de los acentos (*deça* / *deçà*); las vacilaciones en las grafías vocálicas (*a* / *e* átonas: *condamnada*, *condemnada*; *ambaxada*, *embaxada*; *ambaixada*, *embaixada*) o consonánticas (*ç* / *ss*: *carabaça*; *carabassa*; *caloraça*, *calorassa*; *x*/ *ix*: *axeca* / *aixeca*; [ks]: *lecsicográfica*, *lecxicografica*, *lexicografica*, *lexicográfica*, *lexicogràfica*), entre muchas otras.

Desde un punto de vista lingüístico, una primera aproximación a los textos facilita, por un lado, la determinación de las palabras más frecuentes[143] y, por otro, la extracción de las que aparecen conjuntamente en los 15 documentos analizados.

En primer lugar, ordenadas por una frecuencia decreciente, las 50 primeras palabras que aparecen en los documentos (cf. el Apéndice 1) se refieren principalmente, y como es usual, a partículas gramaticales: *a*) preposiciones (*de* es la palabra más frecuente; *a*, *per* 'por' (pero hay que considerar sus variantes formales, como la aglutinación *pera* (prep. *per* y *a*) o la representación *pe'se* (prep. + art. "salat")); *b*) conjunciones (*i* y *que* son la segunda y tercera más frecuentes; *o*, *com*); *c*) pronombres átonos (*hi*, *se*, *s*, *me*, *m*, *mos*, *li*); *d*) adverbios (*no*, *molt* 'mucho', *dins* 'dentro', *ben* 'bien'); *e*) contracciones (*del*, *dels* o *des*); *f*) sólo dos formas verbales (el auxiliar *he* y *fan* 'hacen'); *g*) adverbios y/o adjetivos (*més* 'más', *tot* 'todo', *tots*); y *h*) dos adjetivos (*sant* 'santo', *gran* 'grande'), que se localizan a partir de la posición 49.

Existen diversos casos de elementos ambiguos: *es* puede referirse al verbo copulativo, al pronombre reflexivo o al artículo derivado de *ipse*, llamado "salat"; *la* o *les* pueden ser tanto artículos como pronombres átonos femeninos; lo mismo sucede, en cuanto al masculino, con *el*, *lo* o *els*; *l* puede representar la elisión del artículo o del pronombre de tercera persona masculinos o femeninos; *son* puede referirse al verbo copulativo o al posesivo átono; *ha* es el auxiliar de *haver*, pero es un componente de la forma impersonal *hi ha* 'hay'; *sa* en general es el artículo 'salat', pero también puede ser el posesivo átono o el adjetivo; etc.

En segundo lugar, 42 de las 50 palabras más frecuentes (cf. el Apéndice 1) aparecen conjuntamente en los quince textos. Es sencillo obtener las 199 palabras que aparecen en todos los textos examinados (cf. el Apéndice 2). En este caso, obviando las partículas gramaticales, destacaremos algunos de los elementos morfológicos y léxicos coincidentes.

En cuanto a la morfología pronominal, son comunes los demostrativos: *aquest*, *aquell*, *aquells*, *aquella*, *aquelles*, *aquests*; y los posesivos: *nostra*, *nostres*. Las formas verbales *som*, *han*, *fer*, *té*, *era*, *veure*, *va*, *dit*, *dir*, *fan*, *esser*, *fet*, *tenen*, *haver*, *feta*, *fent*, *eren*, *feu*, *acaba*, *estudiar*, *sia*, *tenim*, *fou*, *fets*, *venir*, *queden* también aparecen en los quince textos. Además de algunas formas auxiliares (*han* y *som*), los infinitivos y participios forman parte, en general, de perífrasis. Las terceras personas del singular y del plural, junto con los deícticos, revelan el carácter descriptivo de los textos. La primera persona del plural (*tenim*) se hace eco de la voz del narrador, que queda en segundo plano ante las descripciones de carácter impersonal.

Respecto al léxico, aparecen los substantivos *dia*, *gent*, *llengua*, *part*, *anys*, *ciutat*, *coses*, *diccionari*, *obra*, *temps*, *manera*, *tren*, *mon*, *hora*, *costat*, *vida*, *escola*, *missa*, *hores*, *nom*, *senyor*, *poble*, *estudi*, *vespre*, *mestre*, *estudis*, *formes*, *boca*, *bolletí*, *motiu*, *tom*, que revelan los objetivos e intereses del dialectólogo con relación a sus encuestas; y los adjetivos: *gran*, *major*, *bona*, *nou*, *santa*, *catalana*, *diferents*, *ferm*, *bons*, *passat*, *dur*, *actual*, *popular*. Hay también elementos ambiguos (*deu*, *seu*, *partida*, *estat*, *veu*, *prou*, *mitja*, *vol*, *dona*, *seus*, *cor*, *segons*, *dita*, *viu*). Puesto que el corpus no está lematizado,

143 Cf. en Woods (2001) las posibilidades que ofrece este tipo de cuantificación.

se observan contrastes gráficos en elementos gramaticales muy usados. Se trata de la preposición *amb*, junto con sus variantes, *ab* y *am*, o la conjunción copulativa *i*, que presenta la variante (*y* y *e*, cuando, como en castellano, precede a una palabra empezada por *i*).[144]

Los dos elementos que se han cuantificado (cf. la tabla 2) tienen correlaciones temporales en cuanto a su frecuencia de aparición creciente o decreciente, como se observa en la tabla 3.

Tabla 2. Frecuencia de aparición de la preposición amb/ab/ab 'con' y de la conjunción i/y 'y'

Word	Freq	%	Texts		Word	Freq	%	Texts
AMB	4.389	0,46	13		I	27.320	2,88	15
AB	3.343	0,35	15		Y	3.343	0,35	15
AM	810	0,09	8					

Tabla 3. Correlación entre las frecuencias de la preposición amb/ab/ab 'con' y de la conjunción i/y 'y' con la cronología

File	Period	Words	AMB Hits	AM Hits	AB Hits		I Hits	Y (hits)
DM_1901	1901	96.288	1.167	1	13		67	5.079
BDLC1	1901-1902	10.100	148	-	1		57	567
BDLC2	1904-1905	8.784	66	-	20		22	395
BDLC3	1906-1907	53.105	472	-	9		49	2.766
BDLC05	1907	167.376	774	758	82		101	8.855
BDLC4	1908-1909	6.352	25	24	4		240	96
BDLC6	1910-1911	5.086	-	-	57		224	16
Aurora1912	1912	95.995	17	1	1.048		4.420	10
BDLC7	1912-1913	27.596	159		61		1.086	1
Aurora1913	1913	173.574	1.473	3	1		1.455	4
BDLC8	1914-1915	11.294	77		6		431	42
BDLC10	1918-1919	41.881	6	16	322		2.056	8
BDLC11	1920	79.578	1	6	516		3.619	38
BDLC12	1921-1922	94.461	4	1	961		5.377	52
BDLC13	1923-1924	19.950	-	-	241			5

4. Elementos dialectales presentes en los documentos

Si un examen pormenorizado del corpus informa sobre aspectos estilísticos característicos del autor o sobre sus intereses cuando redacta los dietarios, el contraste entre diferentes

144 Este último caso no se contabiliza puesto que *e* no es una grafía complementaria a *i*.

elementos gramaticales permite observar su elección de una determinada forma a lo largo de un cierto periodo temporal.

De los muchos elementos que pueden analizarse, este estudio se centra, puesto que tiene connotaciones dialectales, en los contrastes siguientes: 4.1) fonética: *ll* vs. [j] (*agenollada~agenoyada*); *g* vs. *tg* (*llegida~lletgida*); 4.2) morfología: i): nominal: artículos; demostrativos; posesivos; ii) verbal: realizaciones de la extensión incoativa; iii) pronominal; 4.3) derivación: sufijos; 4.4) léxico.

4.1. Contrastes en la fonética

En múltiples ocasiones, las grafías que aparecen en los textos están claramente vinculadas a la pronunciación. Se examinan a continuación dos elementos relacionados con el consonantismo:

i) *ll* vs. [j]. En los textos suelen contrastar las realizaciones palatalizadas con las yodizadas, propias del proceso histórico:[145] *agenoiar* [19]~*agenoyar* [11] y *agenollar* [27]; *cabei* [7]~*cabey* [11] y *cabell* [32]; *enfilai* [19]~*enfilay* [30] y *enfilall* [94], *taia* [18]~*taya* [18] y *talla*[146] [190].

ii) *g* vs. *tg*: Son mucho más frecuentes las grafías africadas, que muestran la tendencia balear a la africación de la prepalatal, que las realizaciones con la fricativa: *lletgir* [233] vs. *llegir* [9]; *passetjar* [66] vs. *passejar* [1]; *ratjola* [29] vs. *rajola* [1].

4.2. Contrastes en la morfología

4.2.1. Nominal

La formas del artículo salat *es, sa, s'*, *ses*, inusuales en el registro escrito, aparecen exclusivamente en los dos dietarios publicados en *La Aurora*, donde también alternan con las correspondientes formas literarias.

Además de los demostrativos de primer y tercer grado (*aquest* / *aquell*), en los textos puede verse un contraste entre la representación del segundo grado, en relación con la forma masculina plural: siempre *aqueis* [251] o *aqueys* [310], nunca *aqueixos*.[147]

Las realizaciones posesivas del estándar *meva* [97], *meves* [27], *teva* [-], *teves* [-], *seva* [137], *seves* [2] contrastan numéricamente con *meua* [342], *meues* [51], *teua* [3], *teues* [1], *seua* [946], *seues* [248].

145 El mallorquín y, en general, todo el balear, se vio afectado históricamente por un proceso que convertía en yod la [ʎ] resultante de los grupos consonánticos secundarios latinos: C'L, G'L y LY.

146 Se mantiene, sin embargo, sistemáticamente *talla*, cuando tiene valor de substantivo.

147 En realidad, se trata de un contraste fonético, el mismo que aparece entre *mateys*, despalatalizado, y *mateixos*.

4.2.2. Verbal

La extensión incoativa de la tercera conjugación -*esc*, propia del mallorquín, alterna con la estándar -*eix*. Se ilustra este hecho a través de las formas flexionadas de *seguir* (*seguesc* [62] vs. *segueix* [57]); *aclarir* (*aclaresc* [62] vs. *aclarex* [5] (BDLC5)); *agrair* (*agraesch~agraesc* [22] vs. *agraeix* [2]); *beneir* (*beneesch* [8] vs. *beneeix* [1]), con un cierto dominio numérico de la primera extensión incoativa. Hay formas donde sólo existe la variante mallorquina (*aduesch* [2]; *esclafesch* [1]; *meresch~meresc* [10], etc.).

4.2.3. Pronominal

El pronombre de primera persona del plural *mos* [325] (*veuremos*, *passetgemmos*, *abocarmos*) en posición enclítica alterna con la forma estándar *nos* [27] (*aydarnos*, *lletgirnos*); el mismo contraste existe en la posición proclítica *mos* [1725] (*mos embarcàrem*, *mos ne tornam*) vs. *nos* [99] (*nos trobam*, *nos entendrem*), la mayoría localizados en el *Diario de Mallorca*. Existe, en esta misma posición, una forma minoritaria diptongada del pronombre neutro *ho* (*e-hu* o *ehu*) [85] (*e-hu va fer*, *e-hu veu*), que se opone a *ho* [1345] (*ho esplica tot*, *ho demostran*). Lo mismo sucede con la diptongación del locativo *e-hi* [414] (*e-hi ha unes figures*, *e-hi torna comparèixer*) vs. *hi* [9469] (*hi ha una claraboya*, *hi entra*). Las formas diptongadas sólo aparecen en *Aurora*12,13.

4.3. Los sufijos

La posibilidad de ordenar inversamente las palabras facilita la determinación de la frecuencia de determinados sufijos. Concretamente, y como ilustración, se muestra el uso de unos sufijos muy poco rendibles en catalán que Alcover aplica a determinadas palabras para dotarlas de un particular valor aspectivo, intensivo o despectivo: -*enc(h)*/-*enca* [330] (*mancomunidadench*, *llatinench*, *catalanench*; *carabacenca*, *cadafalquenca*, *puigcadafalquenca*, *mancomunidenca*, *diccionarienca*, *barrufetenca*, *regionalistenca*, *lliguenca*, *fabrenca*); afectivo: -*inga* [3] (*maletinga*, *llengwetinga*, *llengwinga*); -*oi* [27] (*petitoy~petitoy*, *menudoi, caminoi~caminoy, caminois~caminoys*); -*oia* [14] (*micoia~micoya*, *menudoia*, *petitoia~petitoya*); -*ando* [8] (*hora-baxando*, *forasterando*, *grosserando*, *malfenerando*) Cabe también subrayar el contraste entre el sufijo balear -*idat* [1.094] (*amabilidat*, *particularidat*, *preciosidat*), mayoritario frente al catalán -*itat* [469] (*uniformitat*, *magnanimitat*, *preciositat*).

4.4. El léxico

En los dietarios aparecen muchas formas propias del mallorquín, documentadas, en su mayoría, en el DCVB. Se trata, entre otros muchas, de *potranca* [4] 'hacer fracasar'; *cateyfa* [4] (variante de *caterva*) 'multitud'; *nigul*/*nigulada* [14] 'nube'/'conjunto de nubes'; *treginada* [3] y su variante *teginada* [1] 'techo'; *encitronar* [24] (y la variante *encintronada*) 'acicalar'. Cabe incluir también *homo/s* [211/148] (vs. *home/s* [151/100]), *aigo/aygo* [38/71] (vs. *aigua* [7] / *aygua* [47]); *atlots* [9] o *al·lot/s* [97]; *horabaixa~horabaxa* [8], *guiterra* [3] o *guya* [1].

5. Formas idiolectales presentes en los documentos

Podrían calificarse de formas idolectales ciertas representaciones populares de algunos cultismos, como el contraste entre *ll* y *l* inicial en palabras como *llingüística* [234] y *lingüística* [17] (y derivados); *llitúrgia* [18] y *litúrgia* [3]; *llògic/a* [17] (o *illògic* [2]) y *lògich* [1]; las metátesis: *drestes* [1] (BDLC10) (por *destres*); *juriol* [48] (per *juliol* [187]); el uso de determinados vulgarismos (*devertida* [2] vs. *divertida* [1]; *gènit* [1] vs. *geni* [10], *jovintut* [11]); y especialmente las frases hechas mallorquinas: *enviar a fregir ous de lloca* [7] 'enviar a freír espárragos'; *aficar bolenga* [1] 'obstinarse en una idea'; *posar miques* [23] 'comer', expresión no documentada en el DCVB; *passar per malla* [1] 'escaparse'; *fent la pretxa* [29] 'conversar'; *a la bordellesca* [3] 'de mala manera'; *esclips i esclops* [1] 'peleas'.

Destacan también expresiones muy usadas en el discurso alcoveriano (*s'és mester* [77] 'es necesario'; *tot lo sant dia* [86]; *tot d'una* [317] 'enseguida'; *de tot y per tot* [24]), la exhortación *hala!* [102] *(¡y hala a conjugar verbs!* (BDLC3), onomatopeyas (*clach* [1] (*Si no sabia l'inglès, clach! quedaria parat i confús* (BDLC3)), eufemismos (*diantre(s)* [30], *dianxa* [1], *dianxe(s)* [5], *diastre* [1], *dixanta* [1] en lugar de *diablo*, en expresiones como *i ¡cap a Sineu com cent mil dianxes!* (BDLC 11), aunque el autor también utiliza *dimoni* [46] (*Es que a voltes costa un dimoni i mig guipar i aglapir totes les formes que hi ha dins qualsevol comarca* (BDLC8)); refranes (*qui no se cansa, alcança* [1] (BDLC7), o dichos moralizantes (*la ganància de Na Peixfrit, que comprava el peix, el fregia, i llavò el venia an el mateix preu que l'havia comprat, posanthi l'oli i la feyna de franch*[148] (BDLC7)).

6. Los límites entre el dialecto y el idiolecto

Siguiendo a Nolan (1994: 331), el idiolecto se considera la elección que cada individuo hace de las formas lingüísticas de carácter fonético, fonológico, morfológico, sintáctico, pragmático o discursivo que su lengua le ofrece; y, podríamos añadir a su vez, que su dialecto le ofrece. Dittmar (1996: 111) completa esta definición constatando que la elección se efectúa a partir de los hábitos adquiridos y de los rasgos estilísticos propios de su personalidad. En el caso de Alcover, los elementos que se han considerado idiolectales no se separan de su variedad dialectal, aunque, obviamente, pueden ser compartidos por la lengua de referencia. Sin embargo, cabe destacar dos rasgos que configuran su idiolecto y que lo alejan del dialecto, entendido como el uso lingüístico compartido por un conjunto de individuos: la individualidad y la univocidad.

Si hasta aquí el contraste establecido entre idiolecto y dialecto parece muy claro, no lo es tanto cuando se pretende determinar el idiolecto de Alcover a través de su producción escrita. El estudio de los elementos lingüísticos que configuran los dietarios de viajes, así como sus restantes escritos, posibilita la caracterización de su prosa, la cual, obviamente, es diametralmente distinta a la de cualquier otro autor contemporáneo y de su mismo entorno geográfico. Hemos señalado, por ejemplo, ciertas estructuras sintácticas muy usadas o el recurso a refranes, onomatopeyas y a un léxico característico. Sin embargo,

148 Es el equivalente a 'Es como el sastre del Campillo, que cosía de balde y ponía el hilo'.

determinadas soluciones aparecen en publicaciones concretas. Perea (2002) estudió la ortografía de Alcover, cuya evolución se ha corroborado en este análisis, que muestra que a partir de 1913 el autor se adaptó temporalmente a las directrices postuladas por las recientemente publicadas normas ortográficas. Desde esta perspectiva, cabe preguntarse hasta que punto controlaba su producción escrita y la adaptaba voluntariamente a la publicación a que iba destinada o a los lectores de los diversos géneros que cultivaba. También cabría prestar atención a cuestiones de variación semántica[149] y a la acción del corrector o del editor sobre sus textos, mediante el examen, cuando existen, de los manuscritos originales.

7. Conclusiones

Este trabajo ha consistido en una primera aproximación al análisis cuantitativo de un corpus de viajes desde la perspectiva de los dialectalismos y de los idiolectalismos propios de su autor y ha mostrado las posibilidades que ofrecen unos materiales de estas características. Aunque se ha obtenido una información significativa, quedan todavía diversas cuestiones por tratar; entre muchas otras, la presencia y el número de castellanismos (*enternir*, *acontexement*, etc.) o el uso contrastivo de ciertos sufijos, como el que pone en evidencia la problemática surgida entre los verbalizadores *-isar* e *-itzar* (cf. Alcover, 1918), etc. Un estudio más completo determinará la posible evolución cronológica que ciertas formas dialectales han podido experimentar, la lematización del corpus ayudará a cuantificar el grado de proximidad o alejamiento respecto al estándar y su etiquetado facilitará el análisis de estructuras sintácticas.

En el caso del dialectólogo mallorquín, y con los resultados que pueden obtenerse mediante el tratamiento cuantitativo de sus obras, que cubren un período cronológico muy extenso, cabe también preguntarse si su idiolecto ha evolucionado o si se adapta a un género (en este caso la literatura de viajes) o a un registro determinados, así como qué elementos son más proclives al cambio o a la estabilidad. Complementariamente, el resultado de las investigaciones, y quizá aplicando técnicas más propias de la lingüística forense, permitiría la atribución a este autor de artículos en la prensa que aparecieron sin firmar, aunque que existen indicios de que le pertenecían.

El futuro y progresivo análisis de la totalidad de la obra escrita alcoveriana —artículos, cuentos (*rondaies*), epistolario, etc.— contribuirá a definir aún más las características idiosincrásicas de su idiolecto, y hasta que punto se adaptó a los requisitos del género literario.

149 Cf. Louwerse (2004).

Apéndices

Apéndice 1

Lista de las 51 palabras más frecuentes

N	Word	Freq.	%	Texts
1	DE	50.603	5,35	15
2	I	27.320	2,89	15
3	QUE	26.228	2,77	15
4	A	25.424	2,69	15
5	LA	20.424	2,16	15
6	Y	17.934	1,90	15
7	D	15.908	1,68	15
8	ES	13.640	1,44	15
9	L	12.702	1,34	15
10	EL	12.183	1,29	15
11	PER	10.820	1,14	15
12	UN	10.462	1,11	15
13	HI	10.254	1,08	15
14	UNA	9.910	1,05	15
15	LES	9.307	0,98	15
16	NO	8.316	0,88	15
17	EN	7.953	0,84	15
18	HA	7.827	0,83	15
19	S	7.038	0,74	15
20	#[1]	7.002	0,74	15
21	DEL	6.866	0,73	15
22	ELS	6.560	0,69	15
23	N	6.392	0,68	15
24	LO	5.326	0,56	15
25	SA	4.976	0,53	15
26	COM	4.719	0,50	15
27	TOT	4.715	0,50	15
28	AMB	4.387	0,46	13
29	SE	4.206	0,44	15
30	M	4.127	0,44	15
31	CAP	4.018	0,42	15
32	ME	3.440	0,36	13
33	MÉS	3.440	0,36	15
34	AB	3.343	0,35	15
35	AN	3.073	0,32	10
36	SON	3.053	0,32	15
37	MOLT	3.033	0,32	15
38	DINS	2.909	0,31	15
39	HE	2.904	0,31	14
40	BEN	2.489	0,26	15
41	SES	2.476	0,26	12
42	DES	2.346	0,25	13
43	DELS	2.343	0,25	15
44	MOS	2.311	0,24	14
45	QU	2.210	0,23	13
46	O	2.204	0,23	15
47	LI	2.193	0,23	15
48	TOTS	2.159	0,23	15
49	SANT	2.139	0,23	15
50	GRAN	2.111	0,22	15
51	FA	2.100	0,22	15

Apéndice 2

Lista de las 199 palabras que aparecen en los 15 textos estudiados, ordenadas por frecuencia decreciente.

	N	Word	Freq.	%	Texts
1	38	DE	50.603	5,35	15
2	96	I	27.320	2,89	15
3	145	QUE	26.228	2,77	15
4	2	A	25.424	2,69	15
5	102	LA	20.424	2,16	15
6	199	Y	17.934	1,90	15
7	37	D	15.908	1,68	15
8	66	ES	13.640	1,44	15
9	101	L	12.702	1,34	15
10	56	EL	12.183	1,29	15
11	136	PER	10.820	1,14	15
12	187	UN	10.462	1,11	15
13	92	HI	10.254	1,08	15
14	188	UNA	9.910	1,05	15
15	103	LES	9.307	0,98	15
16	125	NO	8.316	0,88	15
17	61	EN	7.953	0,84	15
18	89	HA	7.827	0,83	15
19	149	S	7.038	0,74	15
20	1	# [145]	7.002	0,74	15
21	39	DEL	6.866	0,73	15
22	60	ELS	6.560	0,69	15
23	124	N	6.392	0,68	15
24	106	LO	5.326	0,56	15
25	150	SA	4.976	0,53	15
26	33	COM	4.719	0,50	15
27	181	TOT	4.715	0,50	15
28	153	SE	4.206	0,44	15
29	109	M	4.127	0,44	15
30	28	CAP	4.018	0,42	15
31	114	MÉS	3.440	0,36	15
32	3	AB	3.343	0,35	15
33	167	SON	3.053	0,32	15
34	119	MOLT	3.033	0,32	15
35	47	DINS	2.909	0,31	15
36	21	BEN	2.489	0,26	15
37	40	DELS	2.343	0,25	15
38	130	O	2.204	0,23	15
39	104	LI	2.193	0,23	15
40	184	TOTS	2.159	0,23	15
41	151	SANT	2.139	0,23	15

150 Este símbolo representa cifras u otros símbolos, que el programa WordSmith por defecto no contabiliza separadamente.

42	87	GRAN	2.111	0,22	15
43	74	FA	2.100	0,22	15
44	161	SI	2.068	0,22	15
45	97	JA	2.056	0,22	15
46	44	DIA	2.023	0,21	15
47	42	DEU	1.895	0,20	15
48	186	TRES	1.812	0,19	15
49	173	TAN	1.772	0,19	15
50	9	ALTRES	1.739	0,18	15
51	166	SOM	1.694	0,18	15
52	83	FINS	1.681	0,18	15
53	90	HAN	1.642	0,17	15
54	147	QUI	1.602	0,17	15
55	183	TOTES	1.495	0,16	15
56	77	FER	1.450	0,15	15
57	176	TÉ	1.433	0,15	15
58	93	HO	1.345	0,14	15
59	52	DOS	1.325	0,14	15
60	18	AQUÍ	1.304	0,14	15
61	137	PERQUE	1.296	0,14	15
62	64	ERA	1.284	0,14	15
63	159	SEU	1.276	0,13	15
64	182	TOTA	1.270	0,13	15
65	55	E	1.255	0,13	15
66	172	TAMBÉ	1.230	0,13	15
67	195	VEURE	1.222	0,13	15
68	86	GENT	1.201	0,13	15
69	190	VA	1.191	0,13	15
70	132	P	1.189	0,13	15
71	8	ALTRE	1.154	0,12	15
72	27	CADA	1.148	0,12	15
73	57	ELL	1.142	0,12	15
74	49	DIT	1.137	0,12	15
75	75	FAN	1.108	0,12	15
76	144	QUATRE	1.052	0,11	15
77	170	TAL	1.043	0,11	15
78	16	AQUEST	987	0,10	15
79	24	BON	981	0,10	15
80	105	LLENGUA	980	0,10	15
81	29	CASI	971	0,10	15
82	69	ESSER	968	0,10	15
83	12	AQUELL	953	0,10	15
84	165	SOBRE	950	0,10	15
85	189	UNS	950	0,10	15
86	43	DEVERS	948	0,10	15
87	48	DIR	922	0,10	15
88	88	GRANS	884	0,09	15
89	110	MAJOR	875	0,09	15

90	25	BONA	872	0,09	15
91	133	PART	872	0,09	15
92	11	ANYS	862	0,09	15
93	32	CIUTAT	855	0,09	15
94	79	FET	851	0,09	15
95	174	TANT	822	0,09	15
96	59	ELLS	809	0,09	15
97	15	AQUELLS	808	0,09	15
98	35	COSES	802	0,08	15
99	178	TENEN	799	0,08	15
100	129	NOU	788	0,08	15
101	6	AL	776	0,08	15
102	13	AQUELLA	771	0,08	15
103	19	ARA	747	0,08	15
104	63	ENTRE	739	0,08	15
105	134	PARTIDA	729	0,08	15
106	111	MALLORCA	716	0,08	15
107	169	SR	702	0,07	15
108	100	JUST	694	0,07	15
109	45	DICCIONARI	688	0,07	15
110	127	NOSTRA	679	0,07	15
111	131	OBRA	673	0,07	15
112	113	MARIA	648	0,07	15
113	62	ENCARA	645	0,07	15
114	177	TEMPS	622	0,07	15
115	70	ESTAT	620	0,07	15
116	112	MANERA	606	0,06	15
117	155	SEMPRE	606	0,06	15
118	152	SANTA	601	0,06	15
119	185	TREN	596	0,06	15
120	68	ESPANYA	589	0,06	15
121	122	MON	589	0,06	15
122	94	HORA	582	0,06	15
123	107	LOS	577	0,06	15
124	148	RES	577	0,06	15
125	164	SINO	568	0,06	15
126	41	DESPRÉS	559	0,06	15
127	142	PROU	555	0,06	15
128	30	CATALANA	550	0,06	15
129	118	MITJA	538	0,06	15
130	194	VEU	532	0,06	15
131	158	SET	527	0,06	15
132	31	CATALUNYA	521	0,06	15
133	36	COSTAT	500	0,05	15
134	108	LS	488	0,05	15
135	98	JOAN	487	0,05	15
136	46	DIFERENTS	474	0,05	15
137	141	PRIMERA	471	0,05	15

138	196	VIDA	458	0,05	15
139	91	HAVER	455	0,05	15
140	20	BARCELONA	446	0,05	15
141	67	ESCOLA	433	0,05	15
142	80	FETA	433	0,05	15
143	121	MOLTS	431	0,05	15
144	198	VOL	424	0,04	15
145	117	MISSA	420	0,04	15
146	99	JOSEP	418	0,04	15
147	95	HORES	411	0,04	15
148	14	AQUELLES	399	0,04	15
149	126	NOM	398	0,04	15
150	51	DONA	384	0,04	15
151	143	QUAL	383	0,04	15
152	157	SENYOR	380	0,04	15
153	120	MOLTES	379	0,04	15
154	140	PRIMER	378	0,04	15
155	160	SEUS	373	0,04	15
156	76	FENT	368	0,04	15
157	162	SÍ	367	0,04	15
158	53	DOTZE	366	0,04	15
159	138	POBLE	357	0,04	15
160	17	AQUESTS	351	0,04	15
161	78	FERM	351	0,04	15
162	65	EREN	350	0,04	15
163	10	ANTONI	317	0,03	15
164	34	COR	302	0,03	15
165	71	ESTUDI	291	0,03	15
166	175	TANTES	285	0,03	15
167	116	MIL	284	0,03	15
168	193	VESPRE	284	0,03	15
169	26	BONS	282	0,03	15
170	115	MESTRE	276	0,03	15
171	191	VALÈNCIA	273	0,03	15
172	135	PASSAT	264	0,03	15
173	168	SOS	262	0,03	15
174	7	ALGUNS	257	0,03	15
175	171	TALS	244	0,03	15
176	73	ESTUDIS	241	0,03	15
177	84	FORMES	237	0,03	15
178	82	FEU	236	0,02	15
179	4	ACABA	225	0,02	15
180	154	SEGONS	222	0,02	15
181	156	SENS	221	0,02	15
182	72	ESTUDIAR	215	0,02	15
183	163	SIA	214	0,02	15
184	179	TENIM	197	0,02	15
185	58	ELLA	192	0,02	15

186	54	DUR	176	0,02	15
187	22	BOCA	175	0,02	15
188	50	DITA	170	0,02	15
189	128	NOSTRES	165	0,02	15
190	81	FETS	152	0,02	15
191	85	FOU	149	0,02	15
192	192	VENIR	135	0,01	15
193	197	VIU	123	0,01	15
194	5	ACTUAL	100	0,01	15
195	23	BOLLETÍ	99	0,01	15
196	123	MOTIU	75		15
197	139	POPULAR	56		15
198	146	QUEDEN	53		15
199	180	TOM	31		15

Referencias

Alcover, A. M. (1918). "La z llatina i el sufixe dins el català". *Bolletí del Diccionari de la Llengua Catalana*, *X*, 1918-1919, 54-64.

Dittmar, N. (1996). Explorations in 'Idiolects'. En Robin Sackmann and Monika Budde (eds), *Theoretical Linguistics and Grammatical Description: Papers in honour of Hans-Heinrich Lieb***.** Amsterdam: Benjamins, 109-128.

Louwerse, M. M. (2004). Semantic Variation in Idiolect and Sociolect: Corpus Linguistic Evidence from Literary Texts. *Computers and the Humanitie*s, *38*, 207-221.

Nolan, F. (1994). Auditory and acoustic analysis in speaker recognition. En J. Gibbons (ed.). *Language and the Law*. London: Longman, 326-345.

Perea, M. P. (2002). *Dietaris*. Palma de Mallorca: Conselleria d'Educació i Cultura (CD-ROM).

Perea, M. P. (2003). *Bolletí del Diccionari de la Llengua Catalana*, Palma de Mallorca: Conselleria d'Educació i Cultura (CD-ROM).

Perea, M. P. (2004). *Bolletí del Diccionari de la Llengua Catalana* (nova edició ampliada amb índexs), Palma de Mallorca: Conselleria d'Educació i Cultura (CD-ROM).

Ueda, H. y Perea, M. P. (2010) "Método general de lematización con una gramática mínima y un diccionario óptimo. Aplicación a un corpus dialectal escrito". En Moscowich, I., B. Crespo, I. Laredo, P. Lojo (ed.). *Language windowing through corpus. Visualización del lenguaje a través de corpus*, 2. A Coruña: Universidade A Coruña, 919-932.

Woods, M. J. (2001). Spanish Word Frequency: A Historical Surprise, *Computers and the Humanities* 35**,** 231-236.

Http://alcover.iec.cat

The world has got some hint of her country speech: On the Enregisterment of the 'Northern Dialect'

Javier Ruano-García
Universidad de Salamanca

You are not known to the world, so that you may pass for them securely; only the youngest that came from the North, the world has got some hint of her country speech, which if thou canst imitate, we shall cozen the world, live in pleasure, and die in the Bed of Honour

John Lacy, *Sir Hercules Buffoon* (1684: II, i)

This paper analyses literary representations of northern English through the lens of enregisterment. It examines the repertoire of features that have commonly been identified as northern and have thus contributed to the enregisterment of the dialect. For this purpose, I shall undertake a corpus-based analysis of early modern plays included in the Salamanca Corpus. My aim is twofold. Firstly, to identify the most recurrent northern traits of the representations. Secondly, to show that dialect writing, though much neglected for linguistic analysis, gives insights into language variation and attitudes. In fact, these texts are inextricable from the historico-linguistic context in which they were produced, and from the attitude(s) towards the 'other' English which was reproduced.

Keywords: enregisterment, northern dialect, early modern drama, the Salamanca Corpus.

Este trabajo ofrece un análisis de representaciones literarias del dialecto del norte a la luz del enregisterment. Se examinan los rasgos lingüísticos que se han asociado habitualmente con el Norte y han contribuido al enregisterment del dialecto. Para ello, llevaré a cabo un estudio de corpus de textos dramáticos del período moderno temprano en los que se constatan rasgos norteños. Mi propósito es, por un lado, identificar los aspectos lingüísticos más recurrentes de las representaciones. Por otro lado, mostrar que este tipo de documentos dialectales reflejan cuestiones íntimamente relacionadas con la variación y las actitudes lingüísticas. De hecho, las representaciones de dialecto son inseparables del contexto histórico-lingüístico en el que se crearon, así como de las actitudes hacia la variedad que representan.

Palabras clave: Enregisterment, dialecto del norte, teatro del período moderno temprano, Corpus de Salamanca.

1. INTRODUCTION

Recent research in sociolinguistics and dialectology has introduced the concept *enregisterment* to refer to the process whereby certain linguistic features become associated with a particular place and specific sociocultural values (Agha, 2003; Beal, 2009a; Johnstone *et al*., 2006; among others). Agha (2003: 231) defines it as "the processes through which a linguistic repertoire becomes differentiable within a language as a socially recognized register of forms". Studies exemplifying it have shown that enregisterment occurs through a series of discursive practices. For example, Beal (2010: 94-95) holds that "speakers/writers may take part in the process of enregisterment via such practices as dialect writing, the compilation of dialect dictionaries and, more recently, websites dealing with issues of dialect and local identity" (see further Beal, 2009b).

This paper analyses literary representations of northern English through the lens of enregisterment. It examines the repertoire of features that have commonly been identified as northern and have thus contributed to the enregisterment of the dialect. For this purpose, I shall undertake a corpus-based analysis of early modern plays included in the Salamanca Corpus (henceforth SC). My aim is twofold. Firstly, to identify the most recurrent northern traits of the representations. Secondly, to show that dialect writing, though much neglected for linguistic analysis, gives insights into language variation and attitudes. In fact, these texts are inextricable from the historico-linguistic context in which they were produced, and from the attitude(s) towards the 'other' English that was reproduced. The paper is organised as follows. Section 2 gives an overview of enregisterment; section 3 pays attention to the context in which the dramatic representations of the North evolved in Early Modern English (EModE); section 4 makes a quantitative corpus-based analysis of the northern forms found in the corpus; section 5 gives some concluding remarks.

2. THE IDEA OF A DIALECT: ENREGISTERMENT, LANGUAGE VARIATION AND IDENTITY

Traditional dialectology and sociolinguistics have mainly been concerned with mapping varieties onto geographical or social space. However, attempts to capture socio-geographical variation by means of neat linguistic boundaries are rather elusive for obvious reasons. Drawing on linguistic anthropology, semiotics, discourse analysis and historical studies, research over the past few years has complicated the picture, drawing our attention to some key issues in the study of language variation. For example, attention has been paid to the concept of *enregisterment* to explore the ideological links between place, language and sociocultural values, which has provided successful answers to how we come to associate linguistic features to specific varieties, or how some of these come to perform identity.

In his seminal article "The social life of cultural value" (2003), Asif Agha introduced the concept *enregisterment* to explore the emergence and spread of Received Pronunciation (RP). Once a regional variety, RP was used by a close group of speakers in a likewise bound geographical enclave in the South-East of England, not being necessarily linked to

correctness, nor seen as a national model of pronunciation. However, from the eighteenth century onwards, the prescriptivist comments provided by pronouncing dictionaries, and the metalinguistic activity reflected by different types of discourse aided to the circulation of a repertoire of features which were gradually seen as a supra-local standard. As a result, these forms "have been represented collectively in the public imagination as a stable variety and maintained across time and region via metapragmatic practices that reiterate the value of this variety and its link to social status and correctness" (Johnstone *et al*., 2006: 80).

In line with RP, distinct sets of nonstandard forms have been seen as stable varieties, or dialects, linked to specific groups of people inhabiting specific places. Suffice it to say that this association involves linkages of different kinds in that a particular linguistic form is connected with socio-geographical identities, and at the same time is connected with an ideological scheme by which it is evaluated against another variant that makes it become noticeable (Johnstone, 2009: 160). Furthermore, these linkages vary depending on the speaker, as they may index different values for different persons. Some of the links may be, nevertheless, shared thanks to different practices and discourses that typify speech and contribute to their circulation. For example, personal narrative, dialect dictionaries, websites, commodities such as T-shirts and mugs, comedy sketches or literature exemplify the public representation of sets of features as stable varieties. By disseminating habits of speech, either of perception, recognition or production, these activities create public awareness of the values indexed by the features represented, as well as collective ideas about varieties, about dialects.

This process is not a straightforward one in that the relationship between language and social categories evolves and becomes established over time. One important point that should be made is that of the orders of indexicality, proposed by Silverstein (1976) and summarised by Beal (2010: 94), which relates to the levels at which linguistic forms are loaden with social meaning. The taxonomy comprises three orders that are as follows. The first order reflects the correlation of a linguistic form with a social category, observable by an outsider such a linguist. At the second one there is awareness of the link between the linguistic form and the social category, the former coming to be used variably according to style, context, etc. At the third order, the forms linked with a social category are the objects of overt comment. It is worth noting that the three orders may be connected with Labov's (1972) concepts of *indicators*, *markers*, and *stereotypes*, respectively (see Johstone *et al*., 2006: 82-83). Nonetheless, linguistic features subject to overt comment, or stereotypes to use Labov's term, are not necessarily temporary, as it is by overt comment in public discourse that they become enregistered and maintained across time.

Enregisterment refers then to "the processes through which a linguistic repertoire becomes differentiable within a language as a socially recognized register of forms" (Agha, 2003: 231-232). Put simply, enregisterment are the processes by which linguistic features acquire sociocultural meaning, this connexion becoming noticeable thanks to a variety of practices and discourses that put it on display. Once they have been recognised and fixed, speakers may reflexively respond to and use those features to express and / or perform identity (Beal, 2010: 94).

Literary representations of dialect show the correlation between language and social categories, and the ideas derived from it. In the ensuing sections, I pay attention to the dramatic renditions of the North in the Early Modern English period (EModE), and the set of forms which have contributed to the enregisterment of the northern dialect, in its broadest linguistic sense.

3. REPRESENTATIONS OF THE NORTH IN EMODE, WITH SPECIAL REFERENCE TO DRAMA

3.1. *Historico-linguistic and cultural context*

An important number of works with representations of different varieties of English were produced in the EModE period. The growing concern for a unified linguistic model existing at the time was clearly connected with a growing sense of dialect awareness in that Englishmen began to identify a specific variety with a form of linguistic and social prestige (Blank, 1996: 14). The gradual diffusion of this variety stimulated the publication of manuals of orthoepy, grammars, and dictionaries in which overt comments about 'bad' spellings and words were not infrequent. The idea of 'Englishness' was strengthened by linguistic estrangement. As such, the earliest definitions of *dialect* articulated the distance, both geographical and linguistic, between the regional periphery, explicitly located in the West and in the North, and the London centre.

Contemporary literature recreated these feelings. Some writers projected voicing contrasts within the literary works, especially dramatic, emphasising a juxtaposition of speech forms that invited to an opposition of characterological types. In keeping with the early definitions of *dialect*, EModE representations of other Englishes put for the most part northern and south-western varieties on display. They were depicted as subordinate dialects in opposition to the London model, explicitly creating an ideological scheme by which they and the values connected with their speakers were contrasted and ready for consumption. It is worth noting that the literary treatment of these varieties was not alike. A careful examination of the EModE literary renditions suggests that the South-West was loaden with categorical disdain, being personified by country bumpkins subject to ridicule. On the contrary, the representations of northern English present an old, uncorrupted, remote and plain dialect embodied in the persona of a simple, frank, honest northerner that is sometimes portrayed as likewise attractive and appealing. We could mention numerous examples in the EModE literary discourse that exemplify this, going from Deloney's rough Halifax-born Hodgekins in *Thomas of Reading* (1600), to John Lacy's honourable attractive Innocentia in *Sir Hercules Buffoon* (1684), and simple Lolpoop in Thomas Shadwell's *The Squire of Alsatia* (1688) (see further Ruano-García, 2010: 48-72).

There are reasons of a historical, cultural and linguistic kind which lie behind the literary articulation of the North in EModE and which go back to very early times. The ideas of remoteness and sharp speech documented in Trevisa's well-known *Polychronicon* (1387), or the linguistic divide running along the Danelaw are echoed in EModE literature in a way that enables the public performance of these popularised ideas. In addition, they

manifest the binary opposition between the North and South and, more importantly, the attitudes and ideas about the North, northerners and their language that authors implicitly reproduced in their writings. It is worth pointing out that these represent the outsiders' perspective, as most of them were produced by non-northern speakers, and often represented in London for a non-northern audience.

3.2. *Linguistic mechanisms*

As is true of other time periods, the textual fabrication of the North comprises a series of linguistic adjustments intended both to reproduce the language of the northerner, and his / her sociocultural status. These range from spelling manipulations to suggest regional sounds, to the use of lexis and morphosyntactic traits characteristic of the North. Unlike verse and prose, drama provides already from the sixteenth century interesting representations in which northernisms are relatively abundant. The dialogue structure of dramatic discourse naturally opens the possibility that linguistic contrasts can be made on stage so that spectators may note them and at the same time they may contribute to creating a particular literary effect. Semiphonetic spellings and lexicalisms characteristic of the North are more often relied upon to typify both the language and the dialect characters.

As indicated above, the great majority of EModE representations were produced by non-northern speakers. This leads us to consider that the northernisms included in EModE drama were salient enough to be recognised from the outside as prominent and be associated with a particular type of persona, or a set of non-linguistic characteristics such as social class or region. Additionally, their usage in literary products mainly consumed by London spectators suggests that they might have been familiar not only to northerners.

4. ON THE ENREGISTERMENT OF THE 'NORTHERN DIALECT': A BRIEF SURVEY

4.1. *Primary data*

My selection of data has been made according to different criteria. Firstly, I have considered cases of drama only. According to Culpeper & Kytö (2010: 41) they are speech-purposed documents and are thus likely to approximate speech patterns closely. Secondly, I have concentrated on specimens of literary dialects only (see Shorrocks, 1996: 386). These give a clear picture of linguistic juxtapositions, and contrasts of characterological types, as northern and 'standard' English are set in opposition. Thirdly, I have been mostly concerned with texts written by non-northern authors. Their writings implicitly reflect attitudes towards the North, the language and northerners from the other side of the 'borderline'. Finally, I have paid attention to spelling and phonological features, as an examination of the lexical patterns has already been made (Ruano-García 2010: 293-314). In sum, my primary data consist of 34 dramatic texts which amount to over 654,000 words.

4.2. *Analysis of data*

The dialect passages included in the plays selected indicates that the representation of the North largely relies upon a series of linguistic traits which are used recurrently. Table 1 provides an outline of the sounds which are most represented, which I have classified into lexical sets.

Table 1. Top dialect forms found in the corpus (≥ 10 tokens)

Lexical set	Types	Tokens	Incidence per 10,000 w
ME /a:/ (<OE /A:/)	46	350	5.35
ME /u:/	65	248	3.79
ME /o:/	20	242	3.69
ME /a(:)+l/	20	202	3.08
ME /i:/	49	184	2.81
ME /a+nasal/	18	157	2.4
ME /k/	9	93	1.42
ME /s/	8	87	1.33
ME /ai/	14	46	0.7
ME /hw/	8	23	0.35

The words showing the northern development of OE /A:/ into ME /a:/ are by far the most common instances of northernisms in the corpus, followed by those descending from ME /u:/, terms indicating the northern fronting of ME /o:/, and words distinguished by the 'l-vocalisation' process. Additionally, EModE literary fabrications of northern English are likewise built upon forms descending from ME /i:/, unpalatalised variants of ME /k/, as well as words showing the northern development of ME /al/. Although less in number and, therefore, less frequent in the corpus sample, these forms constitute, along with those of a higher incidence, a relatively stable set of features, which concurs with recent accounts of regional varieties that refer to most of them as characteristic of traditional northern English dialects. Wakelin (1991: 88), for example, lists some features which "broadly mark off northern from southern England". He refers to a bundle of isoglosses which include those relating to words with ME /u:/ (*cow*, *down*, *house*), ME /o:/ (*fool*, *goose*, *spoon*) or ME /a:/ (*bone*, *road*, *stone*), these being broadly realised in the North as [u:], [l@], [l@], respectively, against /aU/, /u:/, /@u/ in the South. Similarly, Trudgill (1990: 19-37) names a series of pronunciation features that broadly distinguish the North traditional dialect area from the South, amongst which he includes forms with ME /i:/ (*right*, *light*, *night*), ME /u:/ (*house*, *out*, *cow*), or ME /a:/ (*home*, *bone*, *stone*) which are realised as /i:/, /u:/, /l@/, respectively, against /al/, /aU/, /@u/.

Side by side with the modern evidence highlighting the salience of these traits, it is worth noting that some of the dialect features found in the corpus are also part of metalinguistic activities of the period, which emphasises their prominence during EModE. For example, Alexander Gil's celebrated description of contemporary dialects in *Logonomia Anglica* (1619: 15-16) refers to words with ME /a:/, ME /i:/ or ME /o:/ (see Britton 2007). In like manner, Simon Daines referred in *Orthoepia Anglicana* (1640) to the northern development of ME /i:/ (see Dobson, 1968: 329). Far from providing a detailed picture of the variability of realisations that these sounds might have had across the North, non-literary testimonies reinforce the linguistic ideas that EModE plays circulated about the dialect by means of the repertoire of forms documented.

Whilst these linguistic forms provide the basis for the articulation of the North, it is worth stressing that they were sometimes represented variably in spelling. As a rule, semiphonetic sequences are employed in EModE drama to reproduce the northern sounds, and these are used quite consistently but for a few cases. Taking the words with ME /a:/, Table 2 shows that at least five sequences are documented in the corpus (see García-Bermejo Giner, 2008: 60).

Table 2. Representation of northern forms with ME /a:/: a sample

Spelling in SC	PdE spelling	RP sound	Tokens	Some examples
<ea>	<o>, <oe>, <or>	/V/, /u:/, /@u/, /O:/	82	*beane*, *beath*, *eane*, *gea*, *heam*, *meare*, *nea*, *twea*
<e(+C+e)>	<o>, <or>	/V/, /u:/, /@u/, /O:/	82	*ene*, *beth*, *clethes*, *hely*, *lere*, *mere*, *ne*, *se, whe*
<a>	<o>, <or>	/V/, /u:/, /@u/	74	*ans*, *awn*, *bath*, *hally*, *twa*, *wa*, *wham*
<a+(C)+e>	<o>, <oe>, <or>	/@u/	51	*alane*, *bane*, *drave*, *hame*, *nase*, *sae*, *thase*, *waes, whae*
<ee+r+e>	<o>, <or>	/V/, /O:/	5	*meere*, *neene*

In like manner, words descending from ME /u:/ are also represented variably, with three different sequences having been recorded.

Table 3. Representation of northern forms with ME /u:/: a sample

Spelling in SC	PdE spelling	RP sound	Tokens	Some examples
<au> / <aw>	<ou>, <ow>	/aU/	167	*accawnt*, *caunsel*, *drawning*, *grawnt*, *mauth*, *stawt*
<oou> / <oow>	<ou>, <ow>	/aU/	46	*coows*, *doown*, *hoow*, *oour*, *penthoowse*, *oour*
<oo>	<ou>, <ow>	/aU/	35	*croon*, *mooth*, *oot*, *poond*, *thoosand*, *toon*

As with ME /a:/, the spelling variability attested suggests the existence of different realisations of ME /u:/. In fact, <au>/<aw> points to the earlier diphthongisation of /u:/ in some areas of the North, at the time that <oou>/<oow> and <oo> indicate that /u:/ did not diphthongise in some other areas. The latter realisation is reported by Beal (2005: 124-125) as characteristic of the far North and Scotland, whilst diphthongal reflexes appear to have been more widespread further down South, as in the county of Lancashire (see further Beal, 2005: 124-125).

5. CONCLUSION

It seems that EModE drama shows enregisterment of some northern English features, providing a discursive articulation of some images about the North, the dialect, and the northerner. The historical context in which these literary representations emerged and evolved not only accounts for the social, cultural and linguistic constrasts on the stage, as it also gives insights into the playwrights' ideas and attitudes towards the language, its cultural and social references or values. Although these ideas, or perceptions, may have been variable for each single writer, as they largely depend on what s/he might have known about the North or how s/he might have come in contact with northern English, these literary artefacts provide ample ground to share some of them. Thanks to corpus methodology, it has been shown that a specific repertoire of forms is fairly consistently used in the texts selected, and tightly connected with the type of character to whom it is ascribed. In this vein, the persona of the northerner, embodying a set of non-linguistic features such as frankness, simplicity and plain speech, is linked to a particular form of expression which marks him off from other characters, and which includes different realisations of ME /a:/ or ME /u:/. Sixteenth- and seventeenth-century dramatic renditions of northern English make therefore links between language, social categories and place by collective representations in which the dialect is seen and disseminated as a stable variety, as well as a linguistic and sociocultural construct.

Enregisterment seems thus a valid framework for a different, perhaps non-traditional, reading of the use of dialect in literary discourse. Although literary texts do not lend themselves to absolute linguistic realism, their highlighting specific traits suggests both that they were socially and linguistically recognised as a differentiable register of forms, and that this kind of artefacts show people's understanding and experiencing associations between place and language in and out of the speech community. The implications of this kind of discourse for the study of language variation are clear. As Johnstone *et al*. (2006: 99) put it: "sociolinguists interested in understanding patterns of variation and change in the speech community need to pay attention not just to people's talk but to the metapragmatic activities in which they create and circulate ideas about how they talk."

REFERENCES

AGHA, A. (2003). The Social Life of Cultural Value. *Language and Communication*, *23*, 231-73.

BEAL, J. (2005). English Dialects in the North of England: Morphology and Syntax. In E. Schneider *et al*. (Eds.), *A Handbook of Varieties of English. Vol. 2 Morphology and Syntax* (pp. 114-141). Berlin, MY: Mouton.

BEAL, J. (2009a). Enregisterment, Commodification, and Historical Context: "Geordie" versus "Sheffieldish". *American Speech*, *84* (2), 138-156.

BEAL, J. (2009b). The idea of a dialect: Dialect literature and the enregisterment of urban dialect in 19th-century england. Paper presented at *Studying the Representation of Dialect in Literature.* Sheffield, 25-26, September.

BEAL, J. (2010). *An Introduction to Regional Englishes*. Edinburgh: Edinburgh UP.

BLANK, P. (1996). *Broken English: Dialects and the Politics of Language in Renaissance Writings*. London: Routledge.

BRITTON, D. (2007). Alexander Gill's account of northern speech. In Stephan Dollinger, *et al*. (Eds.), *Tracing English through Time: Explorations in Language Variation* (pp. 17-32). Vienna: Braunmüller

CULPEPER, J. & M. KYTÖ (2010). *Early Modern English Dialogues: Spoken Interaction as Writing*. Cambridge: Cambridge UP.

DOBSON, E. (1968). *English Pronunciation 1500-1700*. Oxford: Oxford UP.

GARCÍA-BERMEJO, M. F. (2008). Early Sixteenth-Century Evidence for [I@] < OE /A:/ in the North? In María F. García-Bermejo Giner, Pilar Sánchez García, *et al*. (Eds.), *Multidisciplinary Studies in Language and Literature: English, American and Canadian* (pp. 59-63). Salamanca: Ediciones U Salamanca.

JOHNSTONE, B. (2009). Pittsburghese Shirts: Commodification and the Enregisterment of an Urban Dialect. *American Speech*, *84* (2), 157-75.

JOHNSTONE, B. *ET AL*. (2006). Mobility, Indexicality, and the Enregisterment of "Pittsburghese". *Journal of English Linguistics*, *34* (2), 77-104.

RUANO-GARCÍA, J. (2010). *Early Modern Northern English Lexis: A Literary Corpus-Based Study*. Bern, etc.: Peter Lang.

SHORROCKS, G. (1996). Non-Standard Dialect and Popular Culture. In J. Klemola *et al*. (Eds.), *Speech Past and Present: Studies in English Dialectology in memory of Ossi Ihalainen* (pp. 385-411). Frankfurt am Main: Peter Lang.

THE SALAMANCA CORPUS: A DIGITAL ARCHIVE OF ENGLISH DIALECT TEXTS. María F. García-Bermejo Giner et al (Eds.) <http://salamancacorpus.usal.es/SC/index.html>.

TRUDGILL, P. (1990). *The Dialects of England*. Oxford: Blackwell.

WAKELIN, M. F. (1991). *English Dialects: An Introduction*. London: Athlone Press.

A data-driven approach to alternations based on protein-protein interactions

Gerold Schneider and Fabio Rinaldi
Institute of Computational Linguistics, University of Zurich

Syntactic alternations like the dative shift are well researched. But most decisions which speakers take are more complex than binary choices. Multifactorial lexicogrammatical approaches and a large inventory of syntactic patterns are needed to supplement current approaches. We use the term semantic alternation for the many ways in which a relation between entities, conveying broadly the same meaning, can be expressed. We use a well-resourced domain, biomedical research texts, for a corpus-driven approach. As entities we use proteins, and as relations we use interactions between them, using Text Mining training data. We discuss three approaches: first, manually designed syntactic patterns, second a corpus-based semi-automatic approach and third a machine-learning language model. The machine-learning approach learns the probability that a syntactic configuration expresses a relevant interaction from an annotated corpus. The inventory of configurations define the envelope of variation and its multitude of forms.

Keywords: syntactic alternations, lexicogrammar, corpus-driven, semantic alternation, text mining, machine learning

Alternaciones sintácticas como la alternancia de dativo se han investigado extensivamente. Sin embargo la mayoría de las decisiones que toman los hablantes van más alla de simples opciones binarias. Métodos multifactorial léxico-gramaticales y un amplio inventario de patrones sintácticos son necesarios para complementar los métodos actuales. Utilizamos el término alternancia semántica para indicar las distintas maneras de expresar una relación entre entidades con el mismo significado. Para nuestro estudio utilizamos como corpus artículos científicos del campo biomédico. Las entidades que consideramos son proteínas, genes, enfermedades y medicinas, y estudiamos las relaciones entres ellas. En nuestro artículo presentamos tres métodos: en primer lugar, patrones sintácticos desarollados manualmente, en segundo lugar un enfoque semi-automático basado en corpus y tercero un enfoque que utiliza técnicas de Aprendizaje Automático. El sistema de Aprendizaje Automático extrae de un corpus anotado la probabilidad que una configuración sintáctica específica exprese una interacción relevante. El inventario de las configuraciones permite definir las variaciones sintácticas en todas sus formas.

Palabras clave: Alternaciones sintácticas, léxico-gramática, lingüística de corpus, alternancia semántica, text mining, aprendizaje automático

Introduction

The present paper suggests the use of a corpus-driven approach to alternations.[151] Instead of viewing alternations as a binary decision between two choices we suggest a view of alternations as a multifactorial phenomenon of many choices, relating the many different ways of expressing similar concepts to each other. We use a *corpus-driven* approach. Instead of focussing on a single phenomenon and its *envelope of variation* (Labov, 1969), a corpus-driven approach forces the researcher to interpret many features, which possibly interact with each other. The detection of the envelope of variation is typically not corpus-driven and often not clear. For example, Arppe, Gilquin, Glynn, Hilpert and Zeschel (2011) state:

> Our focus on alternations is the result of theoretical heritage from generative syntax and a matter of methodological convenience. Most linguistic decisions that speakers make are more complex than binary choices … alternations are as simplistic and reductionistic as the theories of language that originally studied them (Arppe *et al.,* 2011).

Corpus-driven approaches are for example used to discover collocations (Evert, 2008), or diachronic word class shifts (Mair, Hundt, Leech & Smith, 2002). For the discovery of collocations, word forms or lemmas are used as uncontested features, for word-class shifts agreed-on part-of-speech tags can be used. In the case of alternations, there is a considerably less stable base than in collocations or part-of-speech tags, as Arppe *et al.* (2011) warn us. In particular, there are manifold restrictions, strong lexicogrammatical interactions, the sheer number of alternation is contested. In other words: in (probably) the majority of cases where an alternation could syntactically be used, it will lead to a different semantic or an inacceptable or at least not native-like utterance (Pawley & Syder, 1983). In computational terms, there is a precision problem: many application of an alternation rule lead to incorrect results. Also, the vast majority of utterances where two speakers express the same concept differs in more respects than in the choice of a single alternation.

As *semantic equality* is the touchstone of alternation (only those applications of an alternation rule that keep the semantic content largely unchanged are part of the envelope of variation), we would like to keep it as a base. Classical approaches to alternations start from the precision perspective: apply alternations and overgenerate; lose on recall anyway. We would like to suggest starting from a recall perspective: aim at collecting and recognizing all utterances that express the same concept and find out which complex set of alternation choices were involved.

Although we base our suggestions on large amounts of data and carefully annotated corpora, our investigation is exploratory in nature. In section 2, we give a brief introduction to the concept of corpus-drivenness. In section 3 we illustrate and motivate our view of alternations as a multifactorial phenomenon. In section 4, we present our method for collecting and detecting different ways of expressing the same concept, based on carefully annotated corpora from carefully restricted concepts, using an Information Retrieval approach.

151 This research is partially funded by the Swiss National Science Foundation (grant 100014-118396/1). Additional support is provided by Novartis Pharma AG, NITAS, Text Mining Services, CH-4002, Basel, Switzerland.

The corpus-driven approach

The distinction between *corpus-driven* and *corpus-based* has been described by Tognini-Bonelli (2001). In corpus-based approaches, existing hypothesis are tested, while in corpus-driven or *data-driven* approaches, hypotheses arise from the corpus data. Corpus-driven approaches have a advantages and disadvantages. An advantage is that, in areas of gradience and subtle differences, it can bring patterns to the surface that went unnoticed by linguists (e.g. Hunston & Francis, 2000). Variationist linguistics is often very subtle and gradient.

A disadvantage of corpus-driven approaches is that they rely on the quality of the corpus: "… since the information provided by the corpus is placed centrally and accounted for exhaustively, then there is a risk of error if the corpus turns out to be unrepresentative" (Tognini-Bonelli, 2001:88). For corpus-driven approaches, large amounts of data are necessary, and relying on frequencies implies a tacit hypothesis, namely that significant frequency differences in the investigated data are indicative.

Alternations as an open set

In a classical approach, alternations are two syntactic configurations that are used to convey the same meaning. In English, well-known examples are the dative shift which links sentences (1) and (2), or the Genitive alternation, which links sentences (3) and (4).

Peter gave Mary a book.

Peter gave a book to Mary.

Mary's car is fast.

The car of Mary is fast.

Restrictions on alternations

There are many restrictions on the application of alternations. These restrictions are typically referred to as the *envelope of variation* (Labov, 1969) or the *choice context* (Rosenbach, 2003). These restrictions rule out contexts in which a speaker does not really have a choice between the two variants. For example, while

Peter gave a book to those students who had achieved a grade A mark.

is acceptable, hardly anybody would produce (6):

?Peter gave those students who had achieved a grade A mark a book.

(6) violates the linguistic tendency to put long constituents after short ones, the principle of end weight. Similarly, while

Mary's picture of the house is great.

is acceptable, (8) is highly unusual, because nested Saxon genitives are extremely rare, and because the of-PP in (7) does not necessarily express a possessor relation.

?Mary's house's picture is great.

There are at least a dozen such restrictions for each alternation (e.g. Jucker, 1993). Only a minority of all candidate configuration tokens are really available for the alternation. While syntactic restrictions can be listed, there is an almost infinite set of semantic restrictions. The verbal semantics of *give*, for example entail that sentences (9) and (10) are only equivalent if the printout of a speech is intended.

Mary gave a speech to the students.

Mary gave the students a speech.

The deep-syntactic role typically depends on verb semantics. In the nominalization alternation, for example, *destruction of the city* implies *city* as object, while in *implication of the discovery* the word *discovery* is a subject. Such behaviour can be found in most alternations. For example, *God's* creation and *the creation of God* are probably not in the envelope of variation.

Interactions between Lexis and Grammar

Interactions between lexis and grammar have been investigated for the dative shift, e.g. Bresnan and Nikitina (2009) or Lehmann and Schneider (2011). Pronouns as indirect object favour the double NP construction, and there are many idioms and collocational preferences. For example, in *give birth to baby* the alternative *give a baby birth* is basically not used. The configurations favouring the double NP-construction most strongly in Lehmann and Schneider (2011) are given in table 1.

Table 1. Dative shift lemma triplets ordered by preference for the double-NP construction

lemma triplet	dshift	to	for	% dshift	iObj
ask you question	4876	3	8	99.8	you
tell you truth	1203	4	1	99.6	you
tell you story	958	3	3	99.4	you
ask him question	1089	6	1	99.4	him
show you picture	1698	13	1	99.2	you
give you number	470	3	1	99.2	you
bring you update	456	5	0	98.9	you
give them information	519	6	0	98.9	them
bring them home	502	6	0	98.8	them
ask them question	404	3	2	98.8	them

Non-core ditransitive verbs have a different behaviour from prototypes. The most prototypical verb, *give*, has a preference for the double-NP construction, while marginal ditransitives such as *provide*, are rarely used with the double-NP construction.

It is unclear if a list of ditransitive verbs can be compiled in the first place. There are indications that they form an open class. Lehmann and Schneider (2011), for example, deliver the following examples.

One husband, accompanying his wife to a fitting there, responded to her lament that she had nowhere to wear the ballgown he had selected for her by promising to **throw** the ***dress*** a lavish ***party***. (TLN956252198)

Cry ***Orwell*** a ***river***, Mr. Timberlake, while CNN's Jeanne Moos reminds us what 20 years of VMAs have really been all about. (CNN:20030827SE.02)

By moving to pastures new, successful managers can **negotiate** ***themselves*** a new ***package*** of options from scratch. (BNC:AH2:375)

Alternations are not a binary switch

The verb *provide* which we have mentioned also illustrates that the alternation can take many forms: the double-NP construction is not the alternative to an NP + of-PP construction, but to an NP + with-PP construction, or an NP + to-PP construction. The double-NP construction is markedly rare, the BNC contains a few dozen of double-NP constructions, about 4000 with-PP constructions, and about 2000 to-PPs. Examples are given in (14) to (16), the automatic syntactic dependency analysis of sentence (14) is also given in figure 1.

You provided him his death, others have provided him a grave. (BNC-Wri K8S)

The forwards played extremely well as a unit, driving in unison and **providing** their backs **with good ball**. (BNC-Wri K5A)

Salespeople may also be called upon **to provide** after-sales service **to customers**. (BNC-Wri K94)

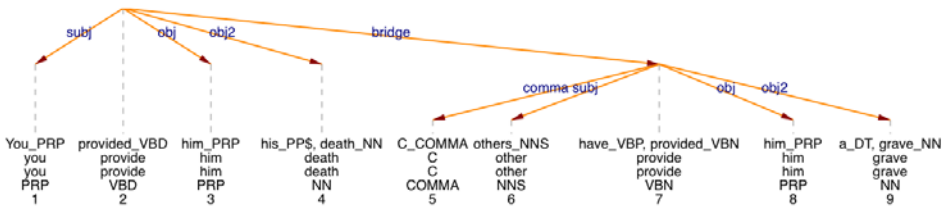


Figure 1. Syntactic analysis of sentence (14).

Many verbs also have an NP + for-PP alternative, which typically expresses benefactor. Levin (1993) provides a detailed verb-lexicon based analysis of alternations. Based on the observation that *load X onto Y* (e.g. sentence (17)) and *load Y with X* (e.g. sentence (18)) express the same meaning she created alternation classes.

In June 1989 the East Londoner had taken a load to Barcelona, where depot staff **loaded his trailer with a mixed consignment** to be taken back to London. (BNC-Wri AHM)

The men, from Pickfords Removals, **were loading a machine onto a trailer** when part of it collapsed, trapping the men beneath them. (BNC-Wri K1G).

Levin compiled 51 coarse classes, containing a total of 193 fin-grained classes. Levin classes cover 3100 verbs.

Ditransitive constructions, in fact most constructions, need to be disambiguated in the context, which means that a dictionary-based approach like Levin's will massively overgenerate, and token-wise disambiguation is necessary.

But I **call** you **a whore**! (BNC-Wri FEE)

Shall I get them **to call** you **a cab**? (BNC-Wri G0B)

PropBank (Palmer, Gildea, & Kingsbury, 2005) and FrameNet (Fillmore, Johnson, & Petruck, 2003) are projects that assign *thematic roles* to verbs and their arguments in context (see Baker and Ruppenhofer (2002) for a comparison between Levin and FrameNet), and that have been used for token-wise disambiguation, for example in the CoNLL-2005 shared task (Carreras & Màrquez 2005). The baseline performance was about 40% F-Score. A baseline takes simple class-based decisions, for example assigning the thematic role *agent* to all subjects and *patient* to all objects. The best system described in Carreras and Màrquez (2005), Punyakanok, Koomen, Roth and Yih (2005), reaches an F-Score of 79.44%. These two percentages illustrate the difficulty of the task – more difficult than for example syntactic parsing. Automatic syntactic parses were provided to the participants.

Lexis is a major disambiguation tool in such tasks, as far as the sparseness problem allows. For example, also humans largely disambiguate (19) and (20) based on the lexis of the second NP, (17) and (18) share the lexis of one object NP. Data sparseness is a problem as most lexical items are very rare (Zipf's law). Due to data sparseness, decisions of classifiers are usually taken at a level between the class-based baseline and a fully lexical decision.

Semantic Alternations

As semantic equality is the touchstone of alternation (only those applications of an alternation rule that keep the semantic content largely unchanged are part of the envelope of variation), we would like to keep it as a base. Classical approaches to alternations start from the *precision*-centered perspective: apply and overgenerate, filter with constraints; lose on recall anyway. We would like to suggest starting from a *recall*-centered perspective: aim at collecting and recognizing all utterances that express the same concept and find out which complex set of alternation choices were involved. Information Retrieval, in particular *Text Mining*, is an applied science that aims to find all textual forms which express a sought-for concept. The detection of *events* (also often termed *relations*) is

particularly relevant for the domain of alternations. Events are typically verb-based, the participants of an event are arguments of the verb, and all configurations that are used to connect them to the verb should be detected.

Methods

We use the following Text Mining scenario for our method: detection of *protein-protein* or *gene-disease-drug interactions* from biomedical texts. Biomedical Text Mining is a domain that has highly developed linguistic resources, for example protein databases, corpora that are annotated for events (IntAct, etc, REFs) and frequent shared tasks where state-of-the-art approaches are competing. In order to recognize a verb's syntactic arguments, we use a syntactic dependency parser (Schneider, 2008). We have used different approaches, which are briefly summarised in the following.

Manual Alternation Patterns

Initially we used a model freely combining classical alternations such as passive, dative shift, genitive, and nominalization. Although it was fairly successful, it overgenerated considerably (Rinaldi, Schneider, Kaljurand, Hess & Romacker, 2006).

Manual class-based disambiguation

Every sentence that contains at least two proteins can express a protein-protein relation. The manual annotation of the corpus, as well as the application phase, needs to discern between those syntactic connections that express a relevant interaction and those that do not. We use the following method: we collect all protein-pairs connected by a dependency chain (‚path') from a term-annotated corpus (we use the GENIA corpus (Kim, Ohta, Tateisi & Tsujii, 2003)). We refer to the syntactic chain up from both proteins to where they meet as *path,* which is then used as a training *feature*. In figure 2, we see the path connecting the gene *nAChR* to the disease *schizophrenia*.

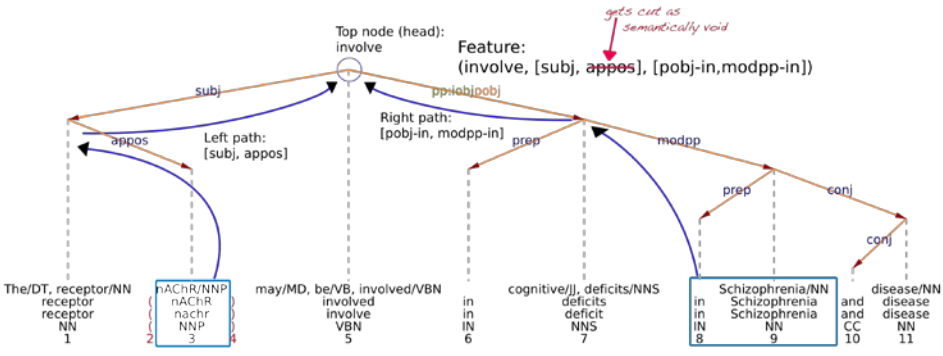


Figure 2. Syntactic path connecting the gene nAChR to the disease schizophrenia

Syntactic relations that are semantically void, like apposition and conjunction, are cut. The manual annotation decisions on which paths are relevant can be used directly for the application phase, augmented by a backoff chain to fight sparse data. The approach is described in detail in Schneider, Kaljurand and Rinaldi (2009). We have participated in the BioCreative II.5 competitive evaluation of biomedical text mining systems (Leitner, Mardis, Krallinger, Cesareni, Hirschmann & Valencia 2010). We achieved the best run for the detection of protein-protein interactions (according to the AUC iP/R metric (Manning, Raghavan & Schütze, 2008). Our system was overall considered as one of the best three.

Probabilistic Model

While the approach described in section 4.2 still involves a manual step, fully automatic learning is possible. If a sufficiently large training corpus is provided, the probability of being relevant for each syntactic path connecting any two entities (E1 and E2) can be calculated, using *Bayesian statistics*, as the following division: how often a given path expresses an interaction between the two entities (*count* in table 2), divided by how often the path appears between the two entities in the whole corpus (*potential* in table 2). We call a path relevant if it expresses an event, i.e. an interaction between two entities.

$$p(\text{relevant} \mid \text{path}_{E1,E2}) = \frac{\mathcal{f}(\text{relevant}(\text{path}_{E1,E2}))}{\mathcal{f}(\text{path}_{E1,E2})}$$

For each candidate entity pair, e.g. two proteins appearing in the same sentence, we suggest paths which have a probability above a certain threshold as being relevant. Many backoffs are used against sparse data. We have applied this approach e.g. in BioNLP 2009, and obtained good results (Kaljurand, Schneider & Rinaldi, 2009). Table 2 shows the most frequent counts from a training corpus for gene-disease-drug interactions.

Table 2. Most frequent relevant paths between entities from a training corpus and probabilities of being relevant

Probability	Head	Path1	Path2	Count	Potential
13.62%	associate	subj	pobj-with	53	389
17.82%	associate	subj modpp-in	pobj-with	31	174
18.29%	cancer			30	164
14.57%	effect	modpp-of	modpp-on	22	151
18.92%	effect	modpp-of	modpp-on modpp-of	21	111
20.65%	association	modpp-of	modpp-with	19	92
6.29%	be	obj modpp-of	subj	19	302
17.82%	metabolize	pobj-by	subj	18	101
29.63%	inhibit	pobj-by	subj	16	54
35.71%	associate	subj modpp-in	pobj-with modpp-of	15	42
23.81%	cause	subj modpp-in	obj	15	63
5.02%	be	subj	obj modpp-of	15	299
100.00%	analyze	subj modpp-in	pobj-in modpart pobj-with	14	14

We have also used versions in which the events are typed, thus forming semantic equivalence classes. All events of the same equivalence class can be said to be semantic alternations. Protein-protein interactions are for example often typed into classes like *regulation*, *binding* and *expression*. Although these classes are domain-specific, they have allowed us to construct a repository of semantic alternations for one domain, as a proof of concept for our suggested model of semantic alternations. There is a strong correlation between the head lexeme and the event type. Table 3 lists the four paths that are found in a training corpus for the verb (first 2 rows) and noun (last 2 rows) *influence*. The paths define the envelope of variation. The last two rows coincide with the classical passive alternation. The probability in the first column is straightforward to interpret in an Information Retrieval setting, but in our setting of finding alternations it is not clear how low a probability should be before we reject it a part of a variation envelope.

Table 3. Data-driven alternations of influence

Probability	Head	Path1	Path2	Count	Potential
37.50%	influence	modpp-of	modpp-on modpp-of	9	24
21.88%	influence	modpp-of	modpp-on	7	32
5.88%	influence	subj	obj	6	102
44.44%	influence	subj modpp-of	pobj-by	4	9

Conclusions

We have suggested a model of semantic alternations which does not use the classical precision-centered perspective (apply and overgenerate, filter with constraints; lose on recall anyway) but an approach starting from a recall-centered perspective: aim at collecting and recognizing all utterances that express the same concept and find out which complex set of alternation choices were involved. We have presented an Information Retrieval application for the biomedical genre which implements this perspective, and which has delivered good results. We use this linguistic model for semantic alternations as a proof of concept.

References

Arppe, A., Gilquin, G., Glynn, D., Hilpert, M., & Zeschel, A. (2011). Cognitive Corpus Linguistics: Five points of debate on current theory and methodology. To appear in *Corpora* 5/2.

Baker, Collin F. & Ruppenhofer, J. (2002) FrameNet's Frames vs. Levin's Verb Classes. In J. Larson and M. Paster (Eds.) *Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society*. 27-38.

Bresnan, Joan & Nikitina, T. (2009). The Gradience of the Dative Alternation. In L. Uyechi & L. Hee Wee (Eds.), *Reality Exploration and Discovery: Pattern Interaction in Language and Life*. Stanford: CSLI Publications.161-184.

CARRERAS, X. & MÀRQUEZ, L. (2005). Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. *In Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, Ann Arbor, Michigan, 152-164.

EVERT, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), Corpus Linguistics. An International Handbook, article 58. Berlin: Mouton de Gruyter.

FILLMORE, C. J., JOHNSON, C. R. & PETRUCK, M. R. L. (2003). Background to FrameNet. *International Journal of Lexicography*, 16:235–250.

HUNSTON, S. & FRANCIS, G. (2000). Pattern grammar: A corpus-driven approach to the lexical grammar of English. Amsterdam: Benjamins,

JUCKER, A. H. (1993). The Genitive versus the of-Construction in Newspaper Language. In: A.H. Jucker (Ed). *The Noun Phrase in English: Its Structure and Variability*. Heidelberg: Universitätsverlag Winter. 121-136.

KALJURAND, K, SCHNEIDER, G. & RINALDI, F. (2009). UZurich in the BioNLP 2009 Shared Task. In *Proceedings of BioNLP workshop*, NAACL/HLT, Boulder, Colorado.

KIM, J., OHTA, T., TATEISI, Y. & TSUJII, J. (2003). GENIA corpus - a semantically annotated corpus for bio-textmining, *Bioinformatics* 19 (1), 180-182.

LABOV, W. (1969). Contraction, deletion, and inherent variability of the English copula *Language 45. 4. 715-62.*

LEHMANN, H. M. & SCHNEIDER, G. (2011). Syntactic variation and lexical Preference in the Dative Shift Alternation. Paper presented at ICAME 2010. To appear in *VariEng*.

LEITNER, F., MARDIS, S. A., KRALLINGER, M., CESARENI, G., HIRSCHMAN, L. A. & VALENCIA, A. (2010). An Overview of BioCreative II.5. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 385-399.

LEVIN, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation.* Chicago: University of Chicago Press.

MAIR, C., HUNDT, M., LEECH, G., & SMITH, N. (2002). Short term diachronic shifts in part-of-speech frequencies. A comparison of the tagged LOB and F-LOB corpora. *International Journal of Corpus Linguistics*, 7(2), 245-264.

MANNING, C.D., RAGHAVAN, P. & SCHÜTZE, H. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

PALMER, M., GILDEA, D. & KINGSBURY, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1).

PAWLEY, A. & SYDER, F. H. (1983). Two Puzzles for Linguistic Theory: Native-like selection and native-like fluency. In J. C. Richards, and R. W. Schmidt (Eds.), *Language and Communication.* London: Longman. 191–226.

PUNYAKANOK, V., KOOMEN, P., ROTH, D., & TAU YIH, W. (2005). Generalized inference with multiple semantic role labeling systems. In *Proceedings of CoNLL-2005*.

RINALDI, F., SCHNEIDER, G., KALJURAND, K., HESS, M. & ROMACKER, M. (2006). An environment for relation mining over richly annotated corpora: the case of GENIA. *BMC Bioinformatics*, 7(Suppl. 3).

ROSENBACH, A. (2003). Aspects of iconicity and economy in the choice between the s-genitive and the of-genitive in English. In G. Rohdenburg & B. Mondorf (Eds). *Determinants of Grammatical Variation in English*. Berlin/New York: Mouton de Gruyter. 379-411.

SCHNEIDER, G. (2008). *Hybrid Long-Distance Functional Dependency Parsing*. Doctoral Thesis. Institute of Computational Linguistics, University of Zürich.

SCHNEIDER, G., KALJURAND, K. & RINALDI, F. (2009). Detecting Protein-Protein Interactions in Biomedical Texts using a Parser and Linguistic Resources. Best Paper Award (2nd place). In *Proceedings of CICLing 2009*, Mexico City. Springer LNC 5449: 406-417.

TOGNINI-BONELLI, E. (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Lingüística computacional basada en corpus

"CorpusLem" una herramienta para la conversión de corpus textuales en datos

Gotzon Aurrekoetxea
UPV/EHU

RESUMEN

La herramienta "CorpusLem" es una herramienta online que convierte información textual en datos organizados en una base de datos. Con una interfaz diseñada en distintas lenguas (inglés, español, francés, vasco y catalán), convierte documentos de texto (.doc, .odt o .txt) en datos estructurados en formato MySQL. A su vez, proporciona un índice alfabético de todas las palabras y propone un lema para cada variante juntamente con el contexto.

Las correcciones oportunas de los índices se pueden llevar a cabo, tanto en la misma herramienta como en su propio ordenador, con la opción de descargar, y, una vez corregido, implementarlo de nuevo. La herramienta está diseñada para albergar diferentes proyectos y soporta más de un usuario por cada proyecto, pudiendo acceder cada uno de ellos a más de un proyecto. El programa puede actuar con textos en variedad estándar o variedades dialectales, en grafía actualizada o grafía original de los textos.

ABSTRACT

"CorpusLem" is a Web tool to convert textual information into data, which is organised in a data-base. The interface has been designed in different languages (English, French, Spanish, Basque and Catalan). This tool converts text documents (.doc, .odt and .txt) into MySQL format and, in addition, it provides an alphabetic index of all the words included in the documents. Apart from that, the "CorpusLem" suggests a lemma for each variant and displays the context of each word.

The user can make the corrections in the index, either into the tool or in its computer, after downloading the required information, and afterwards he can upload the corrected index. The tool is designed to house different projects and more than one user for each project. It could be used with documents written in standard or non-standard varieties, even in standard spelling or in original spelling of the texts.

PALABRAS CLAVE: lingüística de corpus, corpus textuales, indexación, base de datos.

KEY WORDS: Corpus linguistic, text corpora, indexation, data base.

1. Introducción

A pesar de que hoy en día haya gran cantidad de herramientas para el análisis de los corpus orales textuales de la variedad estándar, los corpus orales textuales de habla espontánea para el estudio de la variación lingüística presentan más dificultades para su análisis automatizado, por la carencia de herramientas adecuadas para su explotación; más aún cuando se trata de corpus que recogen textos antiguos que reflejan la variedad oral. En este campo lo más común es encontrar corpus estructurados: NECTE, COD, COSER, EDAK… por citar solo algunos de ellos.

Desde una perspectiva variacionista, dialectológica, los textos espontáneos dialectales presentan grandes posibilidades de análisis. Este es el caso que nos ocupa. El grupo de investigación sobre la variedad lingüística del vasco "Eudia" (UPV/EHU) ha creado la herramienta "CorpusLem" para convertir corpus textuales a base de datos, para posibilitar el análisis geolingüístico de unos manuscritos vascos del siglo XIX.

Estos textos provienen de la recogida de dos leyendas pirenaicas (la leyenda de Barbazán y la leyenda de Tantugou) llevada a cabo por Julien Sacaze en la parte continental de las regiones de los Pirineos, desde el Océano Atlántico al Mediterráneo, en 1887, y reunidas en su "Recueil de linguistique et de toponymie des Pyrénées". Este "Recueil" que alberga textos en vasco, occitano y catalán mayoritariamente, contiene leyendas de 156 localidades en vasco.

El grupo tiene experiencia en este tipo de tareas. En un proyecto anterior (desarrollado entre los años 2004 y 2007), analizó desde esta perspectiva los textos vascos recogidos por E. Bourciez en la Aquitania (Aurrekoetxea & Videgain, 2004, 2009). En aquella ocasión el paso de texto electrónico a base de datos fue muy laborioso y supuso un gasto muy importante de tiempo, siguiendo las siguientes pautas:

- pasar del formato .doc a .txt
- sacar un índice de las palabras de los textos localidad a localidad mediante el programa SCP40[152]
- Pasar los outputs del programa SCP a formato .xls
- Dotar de códigos de localidad a cada palabra en formato .xls
- Transportar los datos del formato .xls a formato .mdb
- Lematización de los vocablos (mediante un programa hecho ad hoc).

Por todo ello, cuando comenzamos este proyecto de investigación vimos la imperiosa necesidad de dotarnos de herramientas para convertir formatos de manera más amigable, y de paso agilizar la conversión.

En el análisis del estado de la cuestión analizamos programas de explotación lexicográfica o terminológica de corpus textuales (programas de concordancias), programas de etiquetaje de textos y programas de análisis morfo-sintácticos, programas multilenguas y programas específicos para cada lengua como los lematizadores, etc. Algunos de ellos orientados al

152 Simple Concordance Program Copyright Alan Reed 1997-2003.

análisis (tanto cualitativo como cuantitativo) de corpus textuales (para un resumen de los recursos on-line ver la página de Joaquim Llisterri: http://liceu.uab.cat/~joaquim).

Los programas a las que tuvimos acceso no nos servían: se basan en textos estándares, y ninguno convertía texto en base de datos.

Nuestro punto de partida, sin embargo, eran manuscritos antiguos, que presentaban aspectos o características de la variedad oral. Y necesitábamos alguna herramienta que nos ayudara en la creación de diccionarios o lexicones con todas las formas recogidas en dichos documentos y en la creación de una base de datos para su posterior explotación desde el punto de vista dialectológico.

El objetivo principal era el estudio de la variación lingüística que contenían los textos. No estaba en nuestro interés la anotación de los textos, ni morfológica, ni sintáctica, ni semánticamente. No por desinterés, ni mucho menos, sino por prioridades.

La ingeniería lingüística para el análisis de la variación es uno de los objetivos estratégicos del grupo de investigación "Eudia". Es una apuesta estratégica que nos ha llevado a usar tecnología informática y, en su defecto, a crear herramientas.

2. La herramienta "CorpusLem"

De esa idea nació la herramienta que presentamos por primera vez en este congreso. Herramienta que facilita enormemente el paso de texto digital a dato.

Esta herramienta se ha concebido dentro de una cooperación pluriuniversitaria y pluridisciplinar entre lingüistas e informáticos (de la UPV/EHU y de la UPPA) y ha sido desarrollada por la empresa de ingeniería informática Eleka[153].

A continuación presentamos las características más sobresalientes de la herramienta:

2.1. "CorpusLem" es un programa Online

La herramienta "CorpusLem" es una herramienta online, que funciona con distintos programas de acceso a Internet (Mozila Firefox, Explorer 7 y 8, etc.).

El programa no requiere ninguna instalación en el ordenador personal. Para acceder a ella se necesita únicamente una cuenta y un *password*, que son proporcionados por el administrador de la herramienta.

La herramienta, creada en lenguaje php, MySQL, y Ext JS JavaScript Framework para la interfaz, está operativa en la dirección http://aholab.ehu.es/CorpusLem/login.html.

2.2. Características de la herramienta

La herramienta proporciona mecanismos para subir documentos, crear reglas internas para mejorar los índices, generar los índices de palabras de los documentos de un proyecto,

153 http://www.eleka.net

bajar índices al ordenador personal y subirlos de nuevo y construir la base de datos que puede ser también bajada al ordenador personal.

Cada proyecto funciona de forma separada; cada uno de ellos puede albergar un número indeterminado de documentos. Todos los proyectos soportan más de un usuario, trabajando al mismo tiempo, accediendo desde distintos ordenadores. Si un proyecto tiene asignados más de un usuario, uno de ellos hará las funciones de administrador.

Cada usuario puede descargar toda la información que ha recogido en cada proyecto, en cualquier momento y desde cualquier ordenador.

La herramienta "CorpusLem" parte de unos textos digitales y soporta formatos diferentes (.doc, .docx, .txt, .dot, o .odt).

No hay límites en cuanto a la extensión del documento y tampoco en el número de documentos. Los documentos se agrupan por proyectos. Cada proyecto puede albergar infinidad de documentos.

2.2.1. La interfaz de la herramienta

La interfaz de la herramienta es muy simple y consta de dos pantallas: a) la pantalla de registro y control; b) la pantalla de trabajo.

2.2.2. Pantalla de registro y control

La herramienta es plurilingüe tanto en la interfaz como en las funcionalidades. Los idiomas con que puede trabajar en la versión actual son: vasco, español, inglés, francés, y catalán.

Al acceder a esta pantalla el usuario debe escoger el idioma, que irá asociado a su perfil cuantas veces acceda al programa (EU para el euskara, CA catalán, ES español, FR francés y EN inglés). La primera vez que accede al programa deberá solicitar dar de alta a un nuevo usuario mediante el botón "solicitar nuevo usuario" y rellenar un formulario simple. Con el código de usuario y el *password* suministrados por el administrador podrá acceder al programa.

Mediante el botón "contacto" todo usuario puede ponerse en comunicación con el administrador. El mismo botón proporciona información sobre los autores y patrocinadores del programa.

2.2.3. Pantalla de trabajo: gestión de proyectos

Una vez en la segunda pantalla del programa, el usuario verá el botón de "salir" del programa, la herramienta "proyecto" para crear uno nuevo y el manual de usuario.

Para añadir un nuevo proyecto presionando en "gestionar proyectos" aparece una pantalla en la que se añade el nombre del proyecto presionando dos veces consecutivas el botón derecho del ratón en el recuadro inferior, y posteriormente elegir el idioma

de trabajo haciendo lo mismo debajo del rótulo "idioma"; y eligiendo un idioma de la lista. Finalmente presionará dos veces seguidas en el signo "+" para validarlo. Una vez terminado el proceso debe salir de la herramienta para actualizar los datos. Al acceder de nuevo al programa se deberá presionar el botón "gestionar proyectos" y elegir uno de ellos pulsando el signo "√" del extremo derecho de la ventana.

Al validarlo el administrador tiene la opción de dar permisos a otros usuarios, tanto para ver la información como para hacer los cambios pertinentes.

Todos los proyectos tienen que tener al menos un usuario. Cada usuario puede tener tantos proyectos como quiera en activo. Cada proyecto, a su vez, puede tener más de un usuario trabajando al mismo tiempo (es el responsable o administrador del proyecto el que asigna diversas funciones a los posibles usuarios del mismo.

Al acceder la aplicación presenta siempre el proyecto actual. El usuario puede gestionar su(s) propio(s) proyecto(s): es decir, añadir un proyecto o eliminarlo.

El programa presenta una herramienta para la puesta a punto de los textos que han de ser explotados. Es lo que se denomina "reglas internas". Y para ello el investigador tiene la herramienta para crear las reglas necesarias para "corregir" tanto la grafía como las variantes idiomáticas que presentan los textos del proyecto.

El usuario puede y debe "gestionar las reglas internas" de cada proyecto. Las reglas internas del proyecto son aquellas reglas que el usuario provee al programa para que lea o interprete de una manera correcta las unidades del texto.

Como esta herramienta nació como ayuda para elaborar índices de palabras de unos textos del siglo XIX con grafías diferentes, se utilizaron unas reglas para igualar las diferentes grafías usadas en los textos. Tenemos cuatro opciones en las reglas: forma equivalente, terminación, juntar palabras y separar palabras.

La primera opción (forma equivalente) nos permite igualar palabras con diferente grafía. Por ejemplo, en los textos aparecen las formas *aaci, araci* y *arasi.* Mediante una regla se le indica al programa que interprete todos ellos como "arazi".

La opción "terminación" actúa solo en final de palabra. Si tenemos tres formas diferentes de un mismo sufijo (*agouac, agoac, ago),* mediante una nueva regla indicaremos que son un mismo sufijo y que hay que tratarlos así.

La opción "agrupar palabras" es imprescindible en los casos en que tanto palabras compuestas, como verbos compuestos que se escriben separadamente para que el sistema interprete que es una única palabra. Así la palabra compuesta *aitasso amassouac* se une en *aitaso-amasoak* mediante esta regla. Igualmente *argui arguia* en *argi-argia.*

Por último la opción "separar" permite al investigador separar o partir palabras que en el texto están unidas: *nahisan* > *nahi izan*.

De acuerdo con la clase de textos con las que trabaja, el investigador puede estar en la obligación de crear reglas. De hecho puede añadir cuantas reglas sean necesarias o quitar las reglas a su libre albedrío.

Estas reglas no modifican los textos. Los textos originales se mantienen intactos, sin ninguna modificación. Se trata únicamente de aumentar el número de coincidencias o concordancias entre los textos de un mismo proyecto. Cuantas más reglas internas la herramienta hila más fino y consigue mejores resultados tanto en la creación de lemas como en la creación de la base de datos.

2.2.4. Utilidad para cargar documentos al proyecto

Al añadir un nuevo proyecto, el investigador debe cargar los textos correspondientes, mediante la utilidad "Añadir fichero al proyecto actual". Al accionar el botón "Examinar" se ha de elegir la carpeta del ordenador personal que contiene los documentos de texto del proyecto para subirlos al programa. Una vez elegido el documento y accionando el botón "+" aparecerá en la lista de los documentos del proyecto. Por cada documento cargado se añade una línea con las características siguientes:

- Índice: creado automáticamente por la herramienta.
- Texto: nombre del texto.
- Abreviatura: es asignada automáticamente por la herramienta (pero puede ser cambiada por el investigador). Si se cambia una abreviatura hay que guardar los cambios mediante el botón "guardar abreviaturas".
- Botón para eliminar el documento

2.2.5. Índice de los vocablos del proyecto

Una vez cargada la herramienta con los textos necesarios en un proyecto, se debe crear el índice de los vocablos mediante el botón "crear índice".

En ella aparecen todas las palabras del proyecto agrupadas y ordenadas con las siguientes especificaciones:

- lema (primera columna): la herramienta analiza las semejanzas entre palabras utilizando el algoritmo de Levenshtein.
- palabra-texto (2ª columna): formas de las palabras que aparece en los textos.
- código del texto (3ª columna):
    - código del texto (abreviatura del texto)
    - línea que ocupa en el texto
    - posición que ocupa dentro de la línea

Por ejemplo, si tomamos la primera línea del índice (aberats abeaxac a-zo_20_3) *aberats* es el **lema**, *abeaxac* la **palabra-texto**, que se halla en el texto correspondiente a la localidad Abouet-Sussaute (País vasco-francés) que tiene como **código** su abreviatura *a-zo,* y que se trata de la palabra que se encuentra en la tercera **posición de la línea** 20.

En textos en variedad estándar la creación de los lemas no genera especial preocupación ni dificultad. Cuanto más se aleje el texto de la variedad estándar y menos reglas internas hayamos introducido, tanto más baja será la calidad de la lematización.

En todo caso el usuario tiene la posibilidad de corregir dichos índices, tarea que puede desarrollar tanto dentro de la herramienta como fuera de ella (en una hoja de cálculo).

Para llevar a cabo la corrección de los índices dentro de la herramienta, no hace falta nada más que pinchar dentro de la ventana del índice. Cuando se realice dicha acción se activa un mecanismo mediante el que el sistema visualiza la cadena en la que está insertada la palabra-texto que está analizando. De esta forma puede resolver mucho mejor las dudas que pudiera albergar sobre la palabra analizada, pudiendo lematizar con más tino.

2.2.6. Instrumentos para crear lemas

Otra opción para la corrección de los índices es descargar los datos al ordenador mediante la instrucción "Herramienta para crear lemas". Esta herramienta descarga el índice para poder corregir y crear los lemas en el propio ordenador, en documento creado en formado .xls.

2.2.7. Cargar la versión corregida

Una vez corregidos los lemas y utilizando la herramienta "Cargar corrección manual de lemas" se carga de nuevo en la herramienta "CorpusLem".

2.2.8. Crear el lexicón

Una vez corregidos los índices se puede crear el lexicón de los textos, mediante la instrucción "Crear el diccionario". Esta instrucción crea el lexicón de los vocablos que aparecen en los textos analizados en formato texto (en diferentes formatos .rtf, .doc, .odt o .docx).

Este lexicón tiene las siguientes características:

a) Entrada (en negrita) seguida de un número en paréntesis que indica el número de apariciones de las variantes de este lema en los textos analizados.

b) Las variantes analizadas en los textos en cursiva y ordenadas alfabéticamente.

c) La abreviatura del texto (entre paréntesis) en la que se halla la variante.

2.2.9. La base de datos

La última fase de la herramienta "CorpusLem" es convertir los textos en datos e introducirlos en una base de datos. El formato de la base de datos es MySQL.

La base de datos tiene las siguientes características: código de localidad o de texto, lema, palabra-texto, línea en la que se encuentra la palabra-texto y posición que ocupa en la línea.

3. Futuros pasos

Uno de los retos para el futuro próximo es ampliar el número de idiomas a la herramienta, y aumentar así la accesibilidad para un público más amplio.

Creemos que sería interesante también que la herramienta proporcionara la posibilidad de crear diccionarios bilingües automatizados.

La estructura de la base de datos debe ser enriquecida proporcionando información geográfica, por una parte; por otra debe ser enriquecida con un sistema de consultas que posibilite la extracción de diversa información recogida en la base.

4. Referencias

Aurrekoetxea, G. & Videgain, Ch. (2004). *Haur prodigoaren parabola Ipar Euskal Herriko 150 bertsiotan [La parábola del hijo pródigo en 150 versiones del País Vasco-francés],* Bilbao: Servicio editorial de la UPV/EHU.

Aurrekoetxea, G. y Videgain, Ch. (2009). Le projet Bourciez: Traitement géolinguistique d'un corpus dialectal de 1895. *Dialectologia, 2*, 81-111. Disponible en http://www.publicacions.ub.es/revistes/dialectologia2/

COD (Corpus Oral Dialectal del catalán): http://www.ub.es/lincat/

COSER (Corpus Oral y Sonoro del Español Rural): http://pidweb.ii.uam.es/coser/

EDAK: http://aholab.ehu.es/edak/2/

NECTE (The Newcastle Electronic Corpus of Tyneside English): http://research.ncl.ac.uk/necte/

SCP40 (Simple Concordance Program): http://www.textworld.com/scp/

Anotación semántica del corpus SenSem

Irene Castellón
GRIAL - Universitat de Barcelona

German Rigau
IXA – Euskal Herriko Unibersitatea

Salvador Climent
GRIAL- Universitat Oberta de Catalunya

Marta Coll-Florit
GRIAL- Universitat Oberta de Catalunya

Marina Lloberes
GRIAL - Universitat de Barcelona

La investigación que presentamos consiste en la anotación semántica de los núcleos argumentales del corpus SenSem, un corpus de oraciones etiquetadas a nivel sintáctico-semántico. La anotación ha sido realizada por un equipo de 6 lingüistas y ha proporcionado los siguientes resultados: un análisis profundo del recurso léxico WordNet 1.6 español, una serie de agrupaciones de conceptos y la anotación del corpus. Así SenSem añade la anotación semántica de nivel léxico a la anotación anterior. El corpus SenSem es de libre disposición bajo una licencia GPL.

Palabras clave: Corpus SenSem, WordNet, Anotación semántica

This paper presents the result of a research project about semantic annotation of argumental heads in SenSem, a corpus of syntactic and semantic annotated sentences. The annotation has been developed by a team of 6 linguists and has produced the following results: a deep analysis of Spanish WordNet 1.6, a proposal of sense clustering and the annotated corpus. Thus, SenSem is improved with lexical semantic annotation. The corpus is freely available with a GPL license.

Keywords: SenSem Corpus, WordNet, Semantic Annotation

1. Introducción

La resolución automática de la ambigüedad semántica de las palabras (en inglés *Word Sense Disambiguation*, WSD) es una de las tareas del Procesamiento del Lenguaje Natural (PLN) que más lentamente avanza debido a la dificultad de establecer las unidades básicas de significado (Agirre & Edmonds 2007). Una forma de caracterizar la información semántica asociada a las palabras y a su uso es la anotación semántica de corpus. Disponer de corpus de la lengua anotados a este nivel proporciona una base sobre la que aplicar, por ejemplo, métodos automáticos de adquisición de preferencias selectivas así como métodos de desambiguación semántica automática.

En este artículo presentamos el proyecto de anotación semántica del corpus SenSem[154] (Alonso et al 2007, Vázquez & Fernández 2008), que se centra en la anotación semántica de los sustantivos, concretamente los núcleos de los complementos verbales, con el objetivo final de adquirir preferencias semánticas asociadas a los predicados verbales. Para la anotación hemos utilizado el WordNet Español alineado a WordNet 1.6 e integrado en el Multilingual Central Repository[155] (Atserias et al. 2004a).

2. Anotación semántica de corpus

SenSem (Vazquez & Fernández 2008) es un banco de datos compuesto por un corpus y una base de datos verbal. El corpus, en su primera versión, consta de 25.000 oraciones, correspondientes a los 250 verbos más frecuentes de la lengua española. Estas oraciones están etiquetadas a nivel sintáctico (tipos de constituyentes y funciones sintácticas). Asimismo, se incluye información semántica sobre el núcleo verbal (especificación de sentidos y papeles temáticos), así como información sobre el tipo de semántica oracional que expresa la oración (información aspectual y tipo de construcción).

En esta segunda etapa del proyecto afrontamos la anotación semántica de los argumentos de SenSem, centrándonos en los núcleos nominales, con el objetivo final de adquirir las preferencias semánticas de los predicados verbales. Para ello hemos usado WordNet 1.6 del español (Atserias et al, 2004b), una red lexico-semántica que consta de sentidos agrupados en conjuntos de sinónimos o synsets (del inglés *synonym set*). La red está organizada mediante hiponimia, meronimia y otras relaciones léxicas. Además, se ha usado como base de conocimiento de apoyo el Multilingual Central Repository (MCR) (Atserias et al. 2004a) el cual integra WordNet con múltiples ontologías de propósito general.

En lengua inglesa ya hace tiempo que se apuesta por desarrollar corpus anotados semánticamente. El proyecto PropBank (Palmer et al. 2003) es el caso más paradigmático de anotación de corpus a nivel de predicados verbales. Sin embargo, teniendo en cuenta el recurso que hemos utilizado para la anotación del corpus, WordNet 1.6 español, los proyectos que están más relacionados con la tarea propuesta aquí son Ontonotes (Yu et al. 2007), SemCor (Miller et al. 94) y MultiSemCor (Bentivogli & Pianta 2005).

154 http://grial.uab.es/fproj.php?id=1
155 http://adimen.si.ehu.es/web/MCR

En lo que se refiere a recursos de este tipo en español, podemos citar dos recursos principales que se han desarrollado en la línea propuesta por SenSem: Adesse (García Miguel & Albertuz 2005) y Ancora (Taulé et al. 2008). Ambos recursos, al igual que SenSem, constan de un corpus y una base de datos verbal con correspondencias entre estos dos. La principal característica distintiva de SenSem es que está diseñado y construido desde su inicio para la anotación de datos lingüísticos asociados a la unidad verbal, lo que se concreta en una constitución representativa y equilibrada de los ejemplos asociados a los verbos (100 ejemplos por cada verbo).

3. Metodología de la anotación

La metodología de anotación utilizada incorpora la experiencia del grupo IXA en la creación del corpus Eusemcor (Agirre et al. 2006). Para llevar a cabo la tarea, el grupo IXA desarrolló una interfaz en línea que hemos adaptado para nuestra investigación. Las etapas que hemos seguido han sido las siguientes:

- Anotación morfológica del corpus con la herramienta Freeling (Padró et al. 2010) e identificación de los substantivos que debían ser anotados.
- Adaptación de la interfaz de anotación del proyecto Eusemcor para realizar la anotación de SenSem.
- Prueba de anotación para calcular el acuerdo entre diferentes anotadores y elaboración de criterios iniciales para la desambiguación.
- Anotación real del corpus SenSem. Participaron 6 lingüistas, de los cuales tres desarrollaron la función de anotadores y los tres restantes actuaron como árbitros y anotadores a la vez. El tiempo de anotación fue de 8 personas/mes distribuidos en 6 meses y tres personas con diferente dedicación.
- Sesiones de coordinación para la aplicación y extensión de los criterios iniciales y pruebas de acuerdo entre anotadores.

Las tareas 4 y 5 se realizaron en paralelo. Como resultado de este proyecto se han obtenido:

1. el corpus anotado
2. criterios para la desambiguación
3. una propuesta de agrupación de sentidos del WordNet 1.6 español

Además, modificamos la metodología utilizada por Agirre et al. (2006). Durante dos meses hicimos algunas pruebas de desambiguación (Carrera et al. 2008) y concluimos la importancia de agrupar determinados sentidos de dicho recurso para garantizar una anotación más consistente, incluyendo esta tarea en el plan de trabajo.

3.1 Interfaz

La interfaz[156] se caracteriza por utilizar un método de desambiguación lema a lema. Así, la interfaz permite buscar todas las ocurrencias de un determinado lema y anotarlo

156 http://sisx04.si.ehu.es:8080/spsemcor/

completamente en el corpus. Esta metodología nos asegura una mayor coherencia en la anotación por dos motivos: el lingüista que anota un determinado lema es el mismo en todo el corpus y, a su vez, permite que el lingüista estudie todos los synsets de un lema al completo para luego proceder a su total anotación.

La interfaz permite la búsqueda por lema, categoría y por el identificador de frase. Además, posibilita al lingüista modificar la categoría morfosintáctica de la ocurrencia en el caso de que no esté bien anotada morfológicamente. Una vez seleccionado un lema, la interfaz nos presenta todos los posibles synsets de ese término. Para cada frase se muestran los lemas de los argumentos y el contexto disponible.

4. Criterios de anotación

Los criterios, como hemos comentado anteriormente, fueron elaborados a partir de un experimento previo de anotación (Carrera et al. 2008). Para ello cuatro lingüistas anotaron conjuntamente los sustantivos de 50 oraciones del corpus SenSem. Sin criterios comunes se llegó a un 40% de acuerdo entre anotadores. Mediante discusiones y reuniones se llegó al establecimiento de unos criterios iniciales que permitieron llegar al 80% de acuerdo en la anotación. Estos criterios iniciales fueron ampliados durante la anotación real, siempre con el acuerdo del equipo de anotación.

El primer paso en el establecimiento de criterios consistió en definir el dominio de la anotación:

- No anotar los pronombres.
- Anotar los núcleos semánticos nominales de los argumentos (y no sus complementos, ni los adjetivos que pueden acompañar) anotados previamente en SenSem.
- Para cantidades de dinero, fechas y nombres propios utilizar las categorías MUC, excepto en el caso de nombres propios que contengan el nombre común en su expresión.
- Tener en cuenta las variantes morfológicas (flexiones, diminutivos, etc.) en el momento de consultar WordNet. En caso de duda, buscar la forma y el(os) lema(s) posibles.
- Anotar únicamente los núcleos de los sintagmas, excepto en el caso de partitivos en que se anotan el núcleo sintáctico (partitivo) y el complemento (núcleo semántico).

Una vez establecido el dominio de la anotación, elaboramos una serie de instrucciones y un conjunto de criterios para los anotadores. Las instrucciones básicamente consisten en la comprensión de un lema y cada uno de sus synsets mediante la exploración de todos los elementos relacionados: por un lado, hiperónimos, hipónimos, variantes, rasgos semánticos asociados como el fichero lexicográfico de WordNet, Top Ontology (Álvez et al. 08), SUMO (Niles & Pease 01), y, por el otro, las glosas (definiciones) asociadas tanto del inglés como del español. No obstante, en los criterios se insiste en la mayor importancia de las relaciones léxicas respecto de las glosas. En general, suele ocurrir que la jerarquía de WordNet es más fiable que las definiciones reflejadas en sus glosas.

Los criterios desarrollados incumben aspectos diversos del conocimiento para la desambiguación semántica. Algunos criterios tratan las relaciones como fuente de información fiable. Por ejemplo, los casos de autohiponimia, que son bastante frecuentes en WordNet, se solucionaron mediante el establecimiento del siguiente criterio: siempre que se encuentren dos synsets que representen un mismo concepto (o muy cercano), uno más genérico y uno más preciso, si el contexto no ayuda a la interpretación, seleccionar el synset más genérico.

Otros criterios se ocupan de las multipalabras. En estos casos, el anotador aplica el siguiente procedimiento: consulta WordNet y, si no recoge la multipalabra con el mismo sentido que aparece en el corpus de anotación, se anotan los componentes de la multipalabra. Si está recogida, se incluye un comentario en la anotación.

Con frecuencia, algunas ocurrencias del corpus de anotación aparecen con sentidos que no están definidos en WordNet. Estos casos se documentan y se describen para posteriores revisiones de WordNet español. El mismo proceso se aplica para las variantes (palabras asociadas al synset) que tampoco existen en WordNet español.

Otro caso de falta de sentidos puede solucionarse de otra forma. Si un sentido no existe pero podemos vincularlo a otro metafórica o metonímicamente, lo asociamos y anotamos como tal.

5. Resultados y análisis del MCR

El resultado final de este proyecto es el corpus Sensem anotado semánticamente. El total de formas anotadas es de 23.307 correspondientes a 3.693 lemas (82,6% del volumen total del corpus). Siguiendo la metodología de Agirre et al (2006), no hemos anotado los lemas que ocurrían menos de 5 veces en el corpus. Además, un total de 91 lemas no han sido anotados por no encontrarse el synset o la variante en WordNet. En general, se trata de nombres que no aparecen en la estructura semántica del inglés. A continuación detallamos alguno de estos casos.

a) Distinciones de sentido que no aparecen en WordNet

Un ejemplo lo encontramos en 'sanidad', en el sentido «Conjunto de servicios gubernativos ordenados para preservar la salud del común de los habitantes de la nación, de una provincia o de un municipio» (DRAE[157]), sentido que se asocia a ocurrencias como 'sanidad española'. En WordNet no se expresa este concepto (cuya traducción al inglés seria algo similar a 'health_system').

No en todos los casos falta un synset, más bien, algunos casos están a medio camino entre la falta de synset o de variante. El caso de 'baile' (*baile de cifras*) es un ejemplo. Se puede considerar que un concepto cercano a 'baile' aparece en WordNet ('variación'), aunque este synset no es el concepto exacto para este lema, ya que 'baile' implica una iteración en la variación.

157 http://buscon.rae.es/draeI/

En cuanto a la definición (glosa), no siempre es de la calidad que se quisiera (como ejemplo se puede consultar la glosa de 'Barcelona'). Por este motivo, hemos primado la jerarquía de relaciones léxicas frente a su glosa.

Todo esto nos sugiere que es necesario completar WordNet 1.6 español con algunos nuevos synsets y nuevas variantes, además de mejorar la calidad de la información ya presente, como por ejemplo las glosas.

b) Distinciones de sentidos muy detalladas

También existen ocurrencias de palabras cuyos sentidos, sin el debido contexto, parecen imposibles de distinguir (p.e. 'pista').

Otro caso lo constituyen los plurales cuando entre las dos lenguas existen ciertos desajustes estructurales. Un ejemplo es 'pareja', cuando se refiere a uno de los individuos (en 'mi pareja'). En algunos casos ha sido necesario agrupar sentidos, dado que aunque tienen una diferencia referencial clara, al mantener una relación de implicación entre los dos sentidos, el contexto no suele aportar ninguna evidencia sobre qué sentido es el adecuado. El total de agrupaciones que se han realizado han sido 58, que comprenden un total de 129 synsets.

6. Conclusiones y trabajo futuro

Hemos presentado la metodología, el desarrollo y la aplicación de la desambiguación semántica de los núcleos argumentales de SenSem. El objetivo final de esta investigación, además de disponer de este recurso y dejarlo a disponibilidad de la comunidad, es la adquisición de preferencias selectivas para su inclusión en una gramática de dependencias (Lloberes et al. 2010). Por ello únicamente hemos anotado los sustantivos (teniendo en cuenta que los verbos ya habían sido desambiguados en un proyecto anterior). Además, hemos realizado un análisis de WordNet 1.6 español, detectando errores estructurales, variantes no registradas o synsets necesarios o inexistentes. Todo este análisis nos es muy útil para la construcción de WordNet 3.0 español que estamos llevando a cabo (Fernandez et al. 2009).

Además, este proyecto nos ha enfrentado a la dificultad en el tratamiento de los sentidos. Se ha evidenciado la ausencia de algunos usos metafóricos o metonímicos muy convencionalizados que deberían incluirse. Esto contrasta con la presencia de usos menos convencionalizados, aunque su inclusión no responde a una sistemática clara, seguramente por seguir la estructura inglesa en su construcción.

Por otro lado, hemos detectado variantes y synsets que no existen en WordNet y que serían necesarios para poder representar los sentidos del español, además de glosas de baja calidad. Todos los datos recopilados nos permitirán en un futuro cercano construir WordNet 3.0 español con mayor fiabilidad (Fernández-Montraveta et al. 2008). Además hemos propuesto una serie de agrupaciones de sentidos orientada a la anotación semántica de corpus.

El recurso que hemos presentado, el corpus anotado sintáctica y semánticamente con los núcleos desambiguados con WordNet, es de libre disposición bajo una licencia GPL.

Agradecimientos

Esta investigación se ha llevado a cabo gracias a los proyectos: FFI2008-02579-E/FILO y TIN2009-14715-C04-03 del Ministerio de Innovación y Ciencia.

Referencias

Agirre E., I. Aldezabal, J. Etxeberria, M. Iruskieta, E. Izagirre, K. Mendizabal, E. Pociello (2006). "Improving the Basque WordNet by corpus annotation", *Proceedings of Third International WordNet Conference.* pp. 287-290.

Agirre, A & Ph. Edmonds (ed.) (2007). *Word Sense Disambiguation Algorithms and Applications*. Springer

Alonso, L., J.A. Capilla, I. Castellón, A. Fernández, G. Vázquez (2007). "The Sensem Project: Syntactico-Semantic Annotation of Sentences in Spanish". In Nikolov, K. et al (ed.), *Recent Advances in Natural Language Processing IV. Selected papers from RANLP 2005.* Benjamins Publishing Co, pp. 89-98.

Álvez J., J. Atserias, J. Carrera, S. Climent, E. Laparra, A. Oliver, G. Rigau (2008). "Complete and Consistent Annotation of WordNet using the Top Concept Ontology", *LREC'08*, Marrakesh, Morroco.

Atserias, J., L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, P. Vossen (2004). "The MEANING Multilingual Central Repository", *Proceedings of the Second International WordNet Conference-GWC,* pp. 23-30.

Atserias J., G. Rigau, L. Villarejo (2004) "Spanish WordNet 1.6: Porting the Spanish Wordnet across Princeton versions". *LREC'04*. Lisboa.

Bentivogli, L. & E. Pianta (2005). "Exploiting Parallel Texts in the Creation of Multilingual Semantically Annotated Resources: The MultiSemCor Corpus", in *Natural Language Engineering, Special Issue on Parallel Texts*, Volume 11, Issue 03, pp. 247-261.

Carrera, J., I. Castellón, S. Climent, M. Coll-Forit (2008). "Towards Spanish verbs' selectional preferences automatic acquisition. Semantic annotation of SenSem corpus", *Proceedings of The 6th international conference on Language Resources and Evaluation*, pp. 2397-2402.

Fernández-Montraveta, A., G. Vázquez, C. Fellbaum (2008). "The Spanish Version of WordNet 3.0". In Storrer, A. et al. (ed.), *Text Resources and Lexical Knowledge* Berlin: Mouton de Gruyter, pp. 175-182.

García-Miguel, J.M. & F. Albertuz (2005). "Verbs, semantic classes and semantic roles in the ADESSE project". In Erk, K. et al (ed.), *Proceedings of the Interdisciplinary*

Workshop on the Identification and Representation of Verb Features and Verb Classes.

Niles, I., A. Pease. (2001) "Towards a Standard Upper Ontology". In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems, Chris Welty and Barry Smith, eds, Ogunquit, Maine.

Miller, G.A., M. Chodorow, S. Landes, C. Leacock, R.G. Thomas (1994) "Using a semantic concordance for sense *identification", Proceedings of the ARPA Human Language Technology Workshop.*

Padró, L., M. Collado, S. Reese, M. Lloberes, I. Castellón (2010). "FreeLing 2.1: Five Years of Open-source Language Processing Tools", in N. Calzolari et al. (ed.), *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pp. 931-936.

Palmer, M., D. Gildea, P. Kingsbury (2003). "The Proposition Bank: An Annotated Corpus of Semantic Roles", *Computational Linguistics*.

Taulé, M., M.A. Martí, M. Recasens (2008). "Ancora: Multilevel Annotated Corpora for Catalan and Spanish", *Language Resources and Evaluation-LREC 2008*.

Vázquez, G., A. Fernández (2008). "Annotation de corpus: Sur la délimitation des arguments et des adjoints", *SKY Journal of Linguistics*, 2, pp. 244-269

Yu, L.C., C.H. Wu, A. Philpot, E.H. Hovy (2007). "OntoNotes: Sense Pool Verification Using Google N-gram and Statistical Tests", *Proceedings of the OntoLex Workshop at the 6th International Semantic Web Conference (ISWC 2007)*.

Estudio comparativo de colocaciones en textos originales y en su traducción

Antonio Frías Delgado
Universidad de Cádiz

Resumen: En el presente trabajo se analizan los n-gramas más frecuentes en textos originales y traducidos (i) como rasgos con potencial valor discriminante en tareas de clasificación y (ii) como criterio de evaluación de consistencia en la traducción experta.

Palabras clave: lingüística computacional, lingüística de corpus, colocaciones, traducción, expresiones multipalabra.

Abstract: Collocations and n-grams are analyzed both in original texts and in translation texts. They are used (i) as features in classification tasks, and (ii) as a possible measure of human translation consistency.

Keywords: computational linguistics, corpus linguistics, collocations, translation, multiword expressions

1. Introducción

Aunque no forme parte del núcleo central de problemas de todas las teorías lingüísticas, las colocaciones vienen siendo objeto de estudio desde hace más de 50 años; Firth, Hallyday, Sinclair son algunos de quienes más se han ocupado de ellas. Desde un punto de vista lingüístico se pueden caracterizar las colocaciones como un grupo de palabras con pérdida total o parcial de composicionalidad, sustituibilidad y/o modificabilidad. El de las colocaciones es un conjunto borroso, en el sentido técnico del término *fuzzy*, ya que se trata, más que de una propiedad de todo o nada, de un continuo en el que, si bien los extremos son bastante nítidos, hay zonas intermedias más disputadas. En lingüística computacional las colocaciones fueron objeto de interés muy especialmente por quienes usaban métodos estadísticos o asumían enfoques empíricos. Aunque no es el primer tratamiento extensivo, el hecho de que Manning y Schütze dedicaran todo un capítulo de su influyente manual (Manning & Schütze, 1999) a las colocaciones, las situó en un prominente lugar de la investigación. Desde hace casi 10 años se celebran *Workshops* sobre *Multiword Expressions*. Las colocaciones son importantes en numerosos campos: generación del lenguaje, traducción automática, lexicografía, enseñanza de lenguas, estilometría, etc.

Que un grupo de palabras sea ("vino blanco") o no ("camisa blanca") una colocación es una cuestión lingüística. Pero un análisis estadístico parece que debería mostrar *candidatos* a colocaciones: palabras que coocurren con una frecuencia mayor de la esperable, estadísticamente significativa. Normalmente el análisis se centra en los bigramas, aunque puede generalizarse a *n*-gramas de cualquier longitud, a grupos de *n* palabras consecutivas. Se utiliza alguna métrica que aplique un número a la significatividad estadística del bigrama y se ordenan en un ranking. Cada métrica suele generar un ranking diferente. Pecina y Schlesinger (Pecina & Schlesinger, 2006) analizan 82 medidas de asociación léxica. El análisis más detallado y el que mejor presenta las complejidades del problema es el de la tesis doctoral de Evert (Evert, 2004).

En este trabajo presentamos los resultados de dos tareas de investigación en las que las colocaciones se usan principalmente como elemento caracterizador de textos originales y traducciones.

2. Tarea 1

Si comparamos un texto creativo original en español, una novela digamos, con un texto traducido, se esperarían, en principio, diferencias en ambos lenguajes. El supuesto general de esta primera tarea es el siguiente: los *n*-gramas estadísticamente significativos, aunque no todos constituyan colocaciones en sentido estricto, revelan preferencias o clichés o muletillas lingüísticas que esperaríamos fuesen diferentes en textos originales y textos traducidos. O, más en general, ¿hay características marcadas que distingan los textos originales de creadores de los textos traducidos?

Para abordar esta tarea seleccionamos 120 textos de 70.000 palabras cada uno distribuidos de la siguiente forma: (i) 12 de clásicos del siglo XVII; (ii) 12 de autores del siglo XIX;

(iii) 12 de autores españoles del XX; (iv) 12 de autores latinoamericanos del XX; (v) 48, divididos en 4 grupos de 12, traducciones de bestsellers en inglés; (vi) 12 traducciones de clásicos en inglés; (vii) 12 traducciones de clásicos en lengua no inglesa. El primer grupo tenía como objetivo medir efectos de diacronía. El grupo (v) trata de ver si hay un efecto género, debido al tipo de texto original, con independencia de la lengua origen. Los grupos (vi) y (vii) tratan de medir efectos de lengua origen.

Una vez preprocesados los textos han de tomarse decisiones metodológicas que afectan a los resultados: ventanas de búsqueda, signos de puntuación, filtros de frecuencias, lematizado, etiquetado, etc. Dado que nuestro interés era encontrar indicios de usos de clichés de cualquier tipo, sólo introdujimos un mínimo filtro de al menos dos ocurrencias.

En general, y con independencia de la métrica que se utilice, los ranking de bigramas suelen tener en las primeras posiciones: nombres propios, erratas, palabras extranjeras, marcadores discursivos (preguntar, decir, responder, etc.), términos cuantitativos o cualitativos generales (mucho, bueno, etc.), formas de pocos verbos (tener, poder, hacer, querer, dar, etc.), y sustantivos genéricos (cosa, gente, mundo, etc.), marcas temporales (día, noche, tiempo, etc.) y partes del cuerpo (ojo, cabeza, mano, pierna, etc.).

La pregunta sobre las diferencias entre textos originales de creadores en español y textos traducidos puede abordarse de distintas formas. Nosotros usamos la siguiente línea: si hay diferencias en el nivel de las colocaciones (más exactamente: de los candidatos a colocaciones, medidos en función del ranking de bigramas), un clasificador automático que utilizase este criterio debería obtener buenos resultados.

Como línea base tomamos los resultados de un clasificador automático que use la frecuencia de palabras funcionales. Aunque los resultados difieren según se use un clasificador u otro, un tamaño de fragmentos u otro, aplicaremos siempre un clasificador bayesiano naive (fragmentos de 7.000 palabras, validación cruzada con 10 particiones). Los resultados que obtenemos para esta base comparativa son los siguientes:

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	780	65 %
Incorrectly Classified Instances	420	35 %
Kappa statistic	0.6111	
Mean absolute error	0.0699	
Root mean squared error	0.2527	
Relative absolute error	38.8425 %	
Root relative squared error	84.2245 %	
Total Number of Instances	1200	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.967	0.007	0.935	0.967	0.951	0.998	1
	0.842	0.042	0.692	0.842	0.759	0.958	2
	0.625	0.054	0.564	0.625	0.593	0.908	3
	0.767	0.052	0.622	0.767	0.687	0.938	4
	0.608	0.053	0.562	0.608	0.584	0.901	5
	0.608	0.028	0.709	0.608	0.655	0.944	6
	0.492	0.047	0.536	0.492	0.513	0.856	7
	0.65	0.049	0.595	0.65	0.622	0.918	8
	0.467	0.043	0.549	0.467	0.505	0.867	9
	0.475	0.015	0.781	0.475	0.591	0.893	10
Weighted Avg.	0.65	0.039	0.654	0.65	0.646	0.918	

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	<-- classified as
116	1	0	0	0	0	0	0	0	3	a = 1 (XVII)
1	101	8	4	0	0	1	1	2	2	b = 2 (XIX)
0	9	75	15	1	0	3	5	11	1	c = 3 (XXE)
0	4	9	92	2	0	0	3	10	0	d = 4 (XXLA)
0	0	4	1	73	15	9	15	3	0	e = 5 (BS)
0	0	2	6	13	73	11	12	2	1	f = 6 (BS)
0	0	7	5	21	7	59	12	7	2	g = 7 (BS)
0	2	1	8	13	4	13	78	1	0	h = 8 (BS)
0	13	13	10	6	3	10	2	56	7	i = 9 (CI)
7	16	14	7	1	1	4	3	10	57	j = 10 (CNI)

Figura 1. Clasificador con palabras funcionales.

Si utilizamos como criterio clasificador los bigramas comunes a los textos, los resultados son los siguientes:

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	419	34.9167 %
Incorrectly Classified Instances	781	65.0833 %
Kappa statistic	0.2769	
Mean absolute error	0.1353	
Root mean squared error	0.2967	
Relative absolute error	75.1483 %	
Root relative squared error	98.9102 %	
Total Number of Instances	1200	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.8	0.09	0.497	0.8	0.613	0.933	1
	0.433	0.138	0.259	0.433	0.324	0.779	2
	0.258	0.046	0.383	0.258	0.308	0.754	3
	0.183	0.044	0.319	0.183	0.233	0.717	4
	0.308	0.078	0.306	0.308	0.307	0.802	5
	0.45	0.061	0.45	0.45	0.45	0.845	6
	0.225	0.046	0.351	0.225	0.274	0.777	7
	0.408	0.102	0.308	0.408	0.351	0.786	8
	0.217	0.055	0.306	0.217	0.254	0.721	9
	0.208	0.064	0.266	0.208	0.234	0.756	10
Weighted Avg.	0.349	0.072	0.344	0.349	0.335	0.787	

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	<-- classified as
96	3	2	1	0	0	0	1	4	13	a = 1
19	52	9	7	4	5	4	8	8	4	b = 2
19	21	31	9	3	2	6	15	8	6	c = 3
9	22	4	22	19	3	8	19	7	7	d = 4
0	10	7	5	37	19	5	26	9	2	e = 5
0	18	8	2	11	54	11	10	5	1	f = 6
1	13	2	5	19	21	27	16	3	13	g = 7
1	13	3	9	21	9	5	49	5	5	h = 8
15	25	11	4	4	5	4	8	26	18	i = 9
33	24	4	5	3	2	7	7	10	25	j = 10

Figura 2. Clasificador con bigramas.

Los peores resultados obtenidos al usar bigramas quizás se deban al bajo número de bigramas que comparten los textos: 28. Aunque, por otra parte, 28 propiedades debieran ser suficientes; excepto que pertenezcan a un nivel general de lengua que hace imposible discriminaciones más finas. Las palabras funcionales, al ser más y aparecer en todos los textos, permiten una mejor discriminación si fragmentamos los textos en torno a 7.000 – 10.000 palabras; si consideramos textos enteros (70.000 palabras), los resultados empeoran ya que prevalece la distribución de Zipf por encima de las características del autor, distribución que para la lengua española al menos experimenta variaciones entre siglos.

La siguiente figura muestra la correlación de las palabras funcionales. Se observa (i) la mayor distancia, esperable, de los textos del XVII, (ii) una influencia del género (mayor distancia de los textos traducidos desde lengua no inglesa), que (iii) se amortigua en la traducción de clásicos del inglés por algún efecto de la lengua de origen.

17	19	20E	20L	BS	BS	BS	BS	CI	CNI	
1	0.8	0.7	0.7	0.6	0.6	0.6	0.6	0.7	0.8	17
-	1	0.8	0.8	0.8	0.7	0.7	0.8	0.8	0.8	19
-	-	1	0.8	0.8	0.8	0.8	0.8	0.8	0.8	20E
-	-	-	1	0.8	0.8	0.8	0.8	0.8	0.8	20L
-	-	-	-	1	0.9	0.8	0.8	0.8	0.7	BS
-	-	-	-	-	1	0.8	0.8	0.8	0.7	BS
-	-	-	-	-	-	1	0.9	0.8	0.7	BS
-	-	-	-	-	-	-	1	0.8	0.7	BS
-	-	-	-	-	-	-	-	1	0.8	CI
-	-	-	-	-	-	-	-	-	1	CNI

Figura 3. Correlación de 314 palabras funcionales comunes.

Que hay un efecto género se observa ya simplemente en la frecuencia de palabras funcionales. Los bestsellers contienen un significativamente menor número de palabras funcionales.

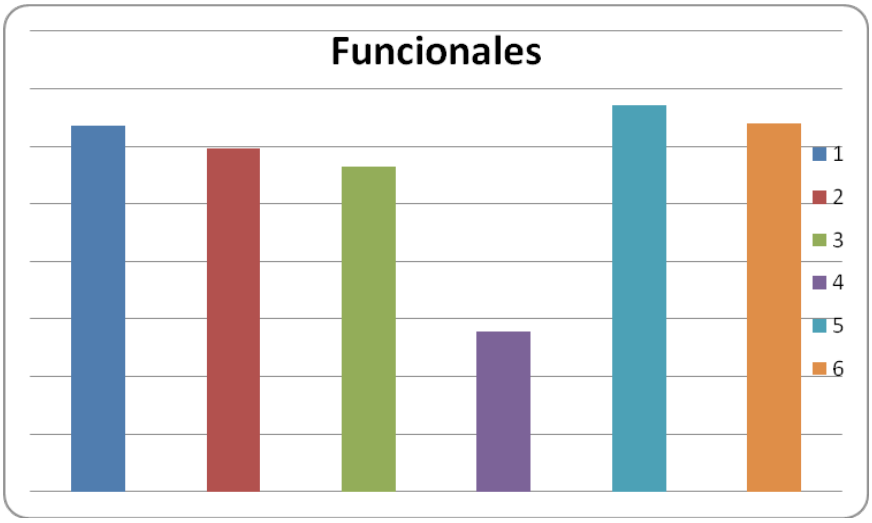


Figura 4. Palabras funcionales 1:XIX; 2:XXE; 3:XXL; 4:BS; 5:CI; 6:CNI

Las figuras siguientes sugieren que existe diferencia entre textos de creadores y textos traducidos en rasgos como el número de hápax y el de bigramas (sin palabras funcionales).

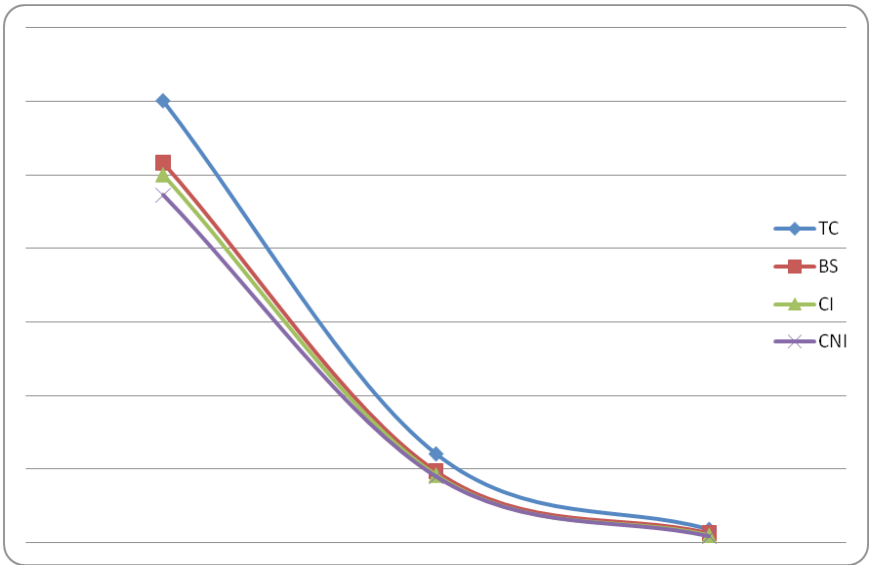


Figura 5. Número de palabras que ocurren 1 (hápax), 2 y 3 veces.

TC: corpus de textos de creadores; BS: bestsellers; CI: clásicos en lengua inglesa; CNI: clásicos en lengua no inglesa.

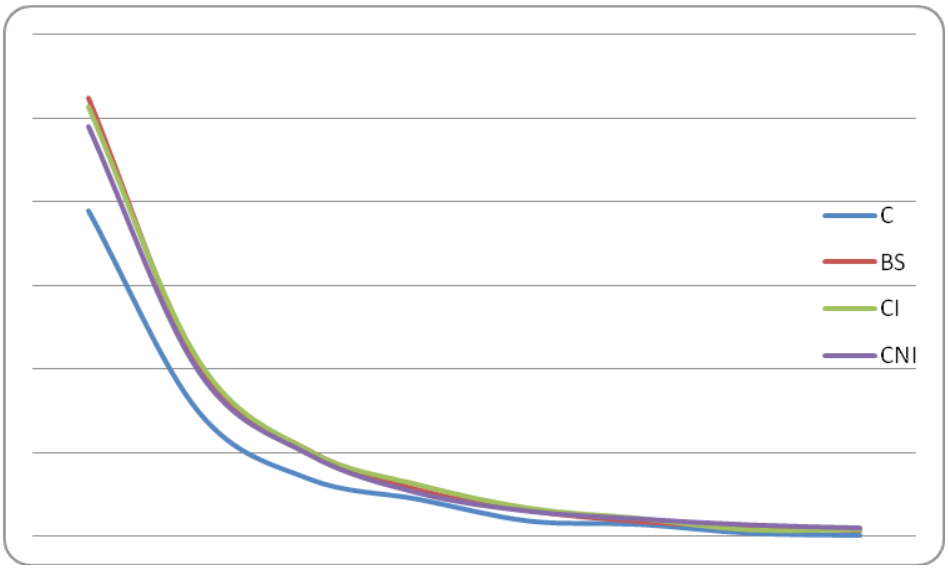


Figura 6. Número de bigramas sin palabras funcionales con frecuencia mayor de 2.

C: corpus de textos de creadores.

El efecto lengua origen puede inferirse también de la correlación de trigramas más frecuentes que muestra la siguiente figura.

XVII	XIX	XXE	XXH	BS	CI	CNI
1	0.29	0.30	0.25	0.28	0.30	0.37
-	1	0.34	0.31	0.34	0.40	0.33
-	-	1	0.44	0.43	0.44	0.32
-	-	-	1	0.43	0.39	0.33
-	-	-	-	1	**0.50**	0.39
-	-	-	-	-	1	0.31
-	-	-	-	-	-	1

Figura 7. Correlación de trigramas más frecuentes.

La siguiente figura muestra la correlación de las 500 palabras no funcionales más frecuentes y que son comunes a los grupos comparados (los resultados van desde 1 –las palabras se usan en el mismo orden de frecuencia- hasta -1 –las palabras se usan en el orden de frecuencia exactamente inverso-). Aunque las palabras no funcionales dependen en buena medida del tema de cada texto, comparar 500 palabras puede amortiguar

parcialmente esa causa y mostrar otros rasgos. Los resultados se muestran en la figura siguiente. Quizás haya que ver en ellos un efecto que tiene más que ver con la cultura (la sensiblemente mayor diferencia de los textos latinoamericanos que los peninsulares respecto a los clásicos del XVII aunque se trate de la misma lengua); esto también explicaría la alta correlación de los textos escritos en inglés, sean bestsellers o clásicos.

XVII	XIX	XXE	XXH	BS	CI	CNI
1	0.14	-0.13	**-0.21**	-0.30	-0.11	0.38
-	1	0.54	0.39	0.39	0.53	0.56
-	-	1	0.58	0.48	0.55	0.34
-	-	-	1	0.52	0.54	0.33
-	-	-	-	1	**0.73**	0.31
-	-	-	-	-	1	0.45
-	-	-	-	-	-	1

Figura 8. Correlación de las 500 palabras no funcionales más frecuentes.

Podemos resumir las conclusiones de esta tarea:

(i) Los bigramas empeoran los resultados de un clasificador automático respecto a las palabras funcionales. No discriminamos mejor los textos si las usamos como criterio alternativo.

(ii) Hay un efecto lengua origen.

(iii) Hay un efecto traducción. El número de bigramas y hápax es menor que en textos de creadores en español.

(iv) Hay un efecto género, observable en la menor ocurrencia de palabras funcionales para los bestsellers.

(v) No hay resultados llamativos sobre un mayor número de clichés en los textos traducidos. Habría que discriminar entre, por ejemplo, frecuencia de marcadores discursivos, meras colocaciones, clichés propiamente dichos, etc. Por otra parte, los textos de los creadores en español muestran altas frecuencias para trigramas y cuatrigramas. A modo de ejemplo, Becquer repite en su prosa hasta 18 veces "en el fondo de" y Cela usa en una de sus obras 62 veces el trigrama "a lo mejor".

3. Tarea 2

Desde hace años (Papineni, Roukos, Ward, Zhu, 2002), los *n*-gramas han formado parte de métricas de evaluación de tareas en campos de lingüística computacional: traducción automática, resúmenes, etc. El objetivo de esta segunda tarea es comparar textos originales y sus traducciones para buscar indicios de colocaciones y, especialmente, para evaluar la traducción desde el punto de vista de la fidelidad y la consistencia.

Cuando se evalúa un traductor automático se compara su traducción con alguna(s) de referencia que sea obra de traductores humanos especializados. La traducción de un experto se supone que es fluida y comprensible para un hablante de la lengua de

traducción. Hay otro criterio, el de la fidelidad a la obra original, que puede ser medido en algunos aspectos, al menos en el de la consistencia. Aunque la fidelidad sea difícil de medir de forma automática, la consistencia puede serlo en algunas variables. Dos nos parecen los mejores candidatos: los nombres propios y las colocaciones o *n*-gramas más frecuentes.

La consistencia es la propiedad de una traducción que respeta lo que dice y cómo lo dice el original. Los nombres propios o nombres de entidades debieran ser traducidas consistentemente a la lengua meta. Dicha consistencia debiera ser, relativamente, fácil de medir al comparar las frecuencias en el texto original y el traducido. Sorprendentemente, suele haber bastantes discrepancias. Los nombres propios, especialmente los geográficos, suelen diferir de una lengua a otra. Obviamente un novelista español escribirá "Londres" y un traductor al inglés "London". Pero la frecuencia de "Londres" en el original y de "London" en la traducción debiera ser la misma si la traducción es consistente. Otro aspecto que debiera respetar una traducción fiel y consistente es el de las colocaciones o giros o clichés que use el autor. Si en el texto original el autor usa 5 veces la expresión "doncellas cobrizas", el traductor, una vez que elija las palabras inglesas para su traducción - "coppery maids", supongamos- debiera usarlas consistentemente en vez de usarlas dos veces y las otras tres usar expresiones alternativas. La discrepancia entre textos originales y sus traducciones por parte de humanos nos hizo añadir un elemento adicional de comparación: la procedente de un traductor automático. Para descontar los efectos que pudieran deberse a la lengua meta, en algunos casos hemos introducido un texto de control arbitrario con el mismo número de palabras y de la misma época.

La siguiente figura muestra una lista de bigramas y trigramas más frecuentes, eliminados los nombres propios, palabras extranjeras y erratas, correspondientes al texto de Blasco Ibáñez *Los cuatro jinetes del Apocalpsis*. La segunda columna corresponde a la traducción inglesa del texto del Proyecto Gutenberg (obra de Charlotte Brewster Jordan) y la tercera al traductor automático de Google.

mansión histórica	acoustic tubes	alarming rate
patriotas exaltados	bell rang	intimate adviser
profeta falso	dagger thrusts	false prophet
vaca floja	printed sheets	perfect state
doncellas cobrizas	barbed wire	tall boy
gaucho fino	wire fences	instinct warned
rudo vaivén	false prophet	riding boots
empresas	alarming manner	reserve artillery
industriales	nothing else	both sides
palidez verdosa	modern warfare	poor deceased
casco puntiagudo	laughed heartily	open spaces
deber cumplido	historic mansion	opposite direction
sordo rodar	irresistible impulse	nothing else
daba cuenta exacta	goings and comings	comings and goings
me da lástima	ups and downs	become less visible
a corta distancia	little sugar soldier	half an hour
sin darse cuenta	in normal times	ray of sunshine
quién sabe si	an evident desire	in normal times
del deber cumplido	a short distance	on both sides
soldadito de azúcar	perish with hunger	who knows if
un modo alarmante	half an hour	and nothing else
cumplir su deber	ray of sunlight	exploring the field
todas las mañanas	on both sides	spent long hours
al día siguiente	no matter how	an hour later
su carácter tenaz	little by little	soon as possible
de este modo	mad with pride	did not hesitate
al aire libre	at any moment	the dining room

Figura 9. Bigramas y trigramas más frecuentes: original, traductor humano, automático.

La consistencia en la traducción de nombres propios es difícil de evaluar automáticamente debido a factores como las normas de cada lengua, acentos, nombres geográficos, etc. Tomando como medida el porcentaje de *tokens* compartidos con distinta frecuencia, hemos encontrado analizando un extenso corpus que el porcentaje de inconsistencia se sitúa como mínimo en torno al 5%.

Cuando se analizan en detalle las posibles fuentes de inconsistencia, puede ocurrir algo tan simple (y extraño) como que el traductor se deje fragmentos enteros sin traducir. La siguiente es una muestra del texto de Blasco Ibáñez citado:

Debían encontrarse a las cinco de la tarde en el pequeño jardín de la Capilla Expiatoria, pero Julio Desnoyers llegó media hora antes, con la impaciencia del enamorado que cree adelantar el momento de la cita presentándose con anticipación. *Al pasar la verja por el bulevar Haussmann, se dió cuenta repentinamente de que en París el mes de Julio pertenece al verano.* El curso de las estaciones era para él en aquellos momentos algo embrollado que exigía cálculos.	They were to have met in the garden of the Chapelle Expiatoire at five o'clock in the afternoon, but Julio Desnoyers with the impatience of a lover who hopes to advance the moment of meeting by presenting himself before the appointed time, arrived an half hour earlier. The change of the seasons was at this time greatly confused in his mind, and evidently demanded some readjustment.

Figura 10. Original y traducción. En cursiva el fragmento no traducido.

Podemos resumir las conclusiones de esta segunda tarea:

(i) La consistencia en la traducción, aunque difícil de establecer, podría medirse por rasgos como la estabilidad en los nombres de entidades y frecuencias de *n*-gramas.

(ii) El porcentaje de inconsistencia, incluso en aspectos simples como el de nombres propios, es mayor del que cabría esperar,

(iii) aunque las causas son muy variadas –algunas tan extrañas como la omisión de fragmentos del original- y sólo un análisis *ad hoc* podría desentrañarlas.

4. Referencias bibliográficas

Church, K.W., and Hanks, P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics*, Vol. 16, No. 1. 22-29.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, Vol. 19, No. 1. 61-74.

Evert, Stefan (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

Manning, Christopher D. and Schütze, Hinrich (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachussetts: The MIT Press.

Papineni, S., Roukos, S., Ward, T. and Zhu, W.J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the ACL 2002*. (pp.311-318)

Pearce, Darren (2002). A comparative evaluation of collocation extraction techniques. In *Third International Conference on Language Resources and Evaluation (LREC)*. (pp. 1530-1536).

Pecina, P. and Schlesinger, P. (2006). Combining association measures for collocation extraction. In *Proceedings of the ACL*. (pp. 651-658)

SCHONE, PATRICK AND JURAFSKY, DANIEL (2001). Is Knowledge-Free Induction of Multiword Unit Dictionary Headwords a Solved Problem? In *Proceedings of Empirical Methods in Natural Language Processing*. (pp. 100-108).

SINCLAIR, JOHN (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Adjunct and complement postmodifiers in popular and academic medical articles: a generative corpus-based approach

Imen Ktari
Research Unit in Discourse Analysis, faculty of Arts and Humanities, Sfax, Tunisia

Abstract

This paper investigates the use of adjunct and complement postmodifiers in academic and popular medical articles, following a generative approach advocated by Chomsky (1970). It aims to show that this use is affected by the genre of these articles. For this reason, the frequency distributions of these adjunct and complement postmodifiers are studied in a medical corpus of 70.000 words equally divided between academic articles (6 articles) and popular articles (11 articles). The quantitative analysis reveals that adjunct postmodifiers are found to be slightly less frequent than complement postmodifiers within the whole corpus and more importantly that academic articles show a preference of complement postmodifiers at the expense of those acting as adjuncts whereas popular articles display a higher frequency of adjunct postmodifiers. These findings lead to the conclusion that the use of adjunct and complement postmodifiers is relatively genre-affected.

Key Words: generative approach, X bar theory, adjunct, complement, popular, academic

0. Introduction

Generative grammar as advocated by Chomsky seeks to reach an analysis of the linguistic item describing its syntactic and phonological structure. According to the X bar theory, which falls under the generative framework, the analysis should comprise three levels of projection: a minimal projection (N), an intermediate projection (N' or N bar) and a maximal projection (N'', N double bar or NP), (Carnie, 2001: 109).

Following a generative approach to a corpus of 70.000 words divided between 6 academic articles (henceforth, AAs) and 11 popular articles (henceforth, PAs) as an example of genre, this paper focuses on the distribution of adjunct and complement postmodifiers. It also aims to show how such a distribution could be genre-affected. The following section reviews postmodification, the generative approach and the X bar theory which distinguishes between adjuncts and complements. The final part describes the corpus and the methods of investigation as well as the findings of the distribution of adjunct and complement postmodifiers within the whole corpus and within each genre along with its relation to the genre.

1. Background

In this section, a review of postmodification, the generative approach and the X bar theory is presented.

1.1 Postmodification

A postmodifier is a linguistic feature that comes after the head noun to add "descriptive" (Quirk, Greenbaum, Leech and Svartvik, 2005: 65) or "additional" (Lynch 2008) information to the head to characterize it (Halliday & Matthiessen, 2004: 324). Others however suggest that modification is meant to restrict and limit the meaning of words. Indeed, Valin and Lapolla (1997: 441) as well as Gleason (1965: 139) use this term to denote respectively "grammatically dependent upon" and "grammatically subordinate to". Because of these various definitions, the latter argues that "modification seems to be one of the fundamental grammatical concepts that cannot be defined, only exemplified" (ibid: 139).

Carnie (2001), following Chomsky's X bar theory of the generative grammar, studies postmodifiers as one constituent within the noun phrase that could be an adjunct or a complement according to several criteria that will be discussed in a next section.

1.2 Generative approach

In phrase structure (PS) grammar, NP analysis follows a flat structure involving two levels: phrase level (NP) and word level (N, AP, D and PP) and following this rule:

NP → (D) (AP+) N (PP+), (Carnie, 2001), where the constituents between parentheses

are optional and those with the "+" sign can be recursive. However, the PS grammar has been unable to account for such features as the tenses, the interrogative form, sentences that share the same structure but have different meanings, etc. (Broderick, 1975: 96; Huddleston, 1989: 47), which led to the emergence of the generative grammar, "a system of rules that can iterate to generate an indefinitely large number of structures" (Chomsky, 1975: 15-6).The aim of this approach advocated by Chomsky is "to generate the well-formed sentences of a language and describe the syntactic and phonological structure of each" (Huddleston, 1989: 82), using "a set of rules or procedures" (Carnie, 2001: 20).

This paper is interested in one theory of the generative grammar: the X bar theory, which will be defined in the following section, and more precisely in the distinction between complement and adjunct postmodifiers as well as their possible association with the genre of the text.

1.3 X bar theory and the adjunct vs. complement distinction

The X bar theory, one of the major rules of generative grammar, "correctly predicts constituency and along with structural relations and the binding theory it also accounts for phenomenon such as how we interpret nominal" (Carnie, 2001: 107). It stipulates three levels of projections: a minimal projection (N), an intermediate projection (N' or N bar) and a maximal projection (N'', N double bar or NP). NP analysis will thus follow these rules:

NP → (D) N'

N' → (AP) N' OR N' (PP)

N' → N (PP)

This theory, therefore, moves from the flat level of analysis to a deeper one that, for instance, distinguishes between the adjunct and the complement postmodifier within the noun phrase rather than representing them at the same level hierarchically. Although this distinction is "difficult" (Herbst, 1988: 268) since it "is not one that you can hear" (Carnie, 2001: 119), some criteria to differentiate between these two linguistic functions are provided by Herbst (1988). They consist in: (i) the possibility of replacing the PP by a one-word adverb, (ii) the question form with "who" or "what, (iii) possibility of the element to appear in a postmodifying relative clause, (iv) the replaceability of the governing noun by "something" or "someone" and (v) the pseudocleft sentence test. Carnie, on the other hand, provides other tests: the preposition they take, location to the head, recursivity, and the "one" replacement test. As for the first test he proposed, he argues that "complement PPs take the preposition *of*. Adjuncts, by contrast, take other prepositions (such as *from, at, to, with, under*, *on*, etc.)", (Carnie, 2001: 119). However, he also admits that "[t]his test isn't 100% reliable" (ibid). The second test consists in the position of the postmodifier to the head. In fact, "the complement PP will always be adjacent to the head. Or more particularly, it will always be closer to the head than an adjunct PP will be" (ibid). The third test, the property of recursivity, is applicable to adjuncts only since "you can have any number of adjuncts, but you can only ever have

one complement" associated to the same head (ibid: 120). The adjunct rule thus "can generate infinite strings of X' nodes, since you can apply the rule over and over again to its own output" (ibid), as shown in the example below:

The book [with a red cover] $_{adj1}$ [from the library] $_{adj2}$ [by Robert John] $_{adj3}$

The complement rule however "cannot apply iteratively" (ibid). This ungrammatical example will explain this idea:

*The book [of poems] $_{comp1}$ [of fiction] $_{comp2}$

The fourth test is the *one*-replacement test stating that "[s]ince complements are sisters to X and not X', they cannot stand next to the word *one*. Adjuncts, by definition, can" (ibid: 122). Consider these two noun phrases illustrating this difference.

The book [of poems] $_{comp}$ → *the one of poems

The book [with the red cover] $_{adj}$ → the one with the red cover

Sister to the head and daughter of the single bar level, the complement is more "lexically specified" (Kroerger, 2005: 88) or in Richard's (2002) terms "much 'pickier'" than adjuncts, in the sense that these complements occur only with some specific words but not with others. The complement "of physics", for example, can be related to the NP "the student" in the student of physics, but not to the NPs "the car" or "the boy".

Hence the complement rule:

X' → X (WP)

Derived from the Latin "to fill", i.e. to complete, the complement is considered as "obligatory" with respect to syntax and meaning (Miller, 2002: 4; Dowty: 2).

The adjunct, on the other hand, is a sister to and a daughter of a single bar level (Carnie, 2001: 117) and "may be freely added to any number of NPs" (Kroeger, 2005: 87). It thus follows this rule:

X' → X' (ZP)

Derived from the Latin verb "join" or "add", the adjunct means "something adjoined, tacked on or not part of the essential structure of a clause" (Miller, 2002: 4), hence its being "optional" (ibid) or "peripheral" (Herbst, 1988: 269). "Not subject to any co-occurrence restrictions apart from 'the need to make sense' (Matthews, 1981: 127)", (Herbst, 1988: 269), the adjunct is seen as "more loosely related to the noun phrase" (Carnie, 2001: 119), which is not the case with complements.

This paper investigates the distribution of adjunct and complement postmodifiers within the two corpora using as many of the above tests as possible. It also verifies if such a distribution is genre-related.

2. Corpus analysis and findings

2.1 Corpus description and methodology

The corpus understudy belongs to the medical genre. It consists of almost 70.000 words equally divided between academic articles (6 articles) and popular articles (11 articles) which are randomly downloaded freely from online journals. The following table provides more information about the corpus.

Table 1. Corpus description

	Academic articles	Popular articles
Sources	Sciencedirect.com Cancer cell, Heart and Lung, Academic Radiology, Cardiovascular Pathology	Time, NY Times, Scientific American, New Scientist
Number of words	34176	35549

Selecting this corpus in particular originates from the fact that the medical register interests almost all people since it is related to their health. Choosing to work on academic and popular medical articles is an attempt to get a comprehensive analysis of the medical genre. Both AAs and PAs are characterized by distinct communicative purposes as well as formal features, which makes them two different genres. . In fact, although they both "report on the same subject-matter and in the same field, one should not be misled into assuming that they belong to the same genre or that one or the other is the sub-genre of another" (Nwogu, 1990: 354).

In the analysis, as many of the tests discussed above as possible will be checked to decide whether a postmodifier is an adjunct or a complement, starting with Carnie (2001)'s four tests (because of their simpler aspect) and then Herbst (1988)'s five criteria. Then the frequency of each is calculated, compared and analysed first in relation to the medical corpus as a whole and then across the two genres of AAs and PAs. Conclusions are then made concerning the possible relation between adjunct and complement postmodifiers and genre.

It is assumed that the high frequency of a structure is understood as a preference of this choice. The analysis also relies on the Chi-square test χ^2 to consolidate the findings.

2.2 Corpus Analysis and Findings

Investigating the frequency distributions of adjunct and complement postmodifiers within the whole corpus as well as across the genres of popular and academic articles is at the heart of this study.

2.2.1 Adjunct and complement postmodifiers within the whole corpu

Table 2 shows the distribution of adjunct and complement postmodifiers within the whole corpus.

Table 2: Adjunct and complement postmodifiers distribution within the whole corpus

	Total
Adjunct Postmodifiers	2019 47 %
Complement Postmodifiers	2315 53 %
Total	4334 100 %

As shown in this table, complement postmodifiers are slightly higher than adjunct postmodifiers in the corpus. On the whole, it can be said that adjunct and complement postmodifiers have nearly equal rates (47% for the adjunct and 53% for the complement postmodifiers). This distribution is surprising with regard to the medical genre. In fact, since "learning science is the same thing as learning the language of science", the scientific genre turns up to be "the discourse of an intellectual elite", of "the expert" (Halliday, 2004: 161). In other words, the characteristics of this genre "suit the experts; and by the same token they cause difficulty to the novice" (ibid: 178). With a high proportion of specialised vocabulary, (Lowe, 2009; Halliday & Martin, 1993; Halliday, 2004), "special expressions", "technical taxonomies", "grammatical metaphor" (Halliday, 2004), the scientific discourse is thus "the prerogative of elite" (ibid: 179). It should then display more items that are syntactically and semantically obligatory and "lexically specified" (Kroerger, 2005: 88) or in Richard's (2002) terms "much 'pickier" than items that are "peripheral" (Herbst, 1988: 269) and "more loosely related to the noun phrase" (Carnie, 2001: 119). It is therefore expected to have more complements than adjuncts.

Example1, AA 3

An analysis , [of two large health plan databases]comp [in the United States]adj

Example2, PA 6

the perils [of alcohol] comp [in school]adj

As can be noted from the two examples above, the complements play a crucial role in completing the meanings of their heads while the adjuncts just add some peripheral information about their heads-here circumstantial data of place- that can be dispensed with. Also, complements are much "pickier" since they reflect some constraints as far as their heads are concerned whereas adjuncts are not. In other words, while many heads can be related to the adjuncts "in the United States" and "in school": a student/ a book/ a picture/ a car, only a few can be associated to the complements "of two large health plan databases" and "of alcohol" In fact, the first complement can have "an investigation"/ "an evaluation," / "a study", etc. as its head but not *a boy/ *a car/ *a picture. Similarly, the second complement can have only a limited number of heads: the dangers, the benefits, etc. but not *a couple/ *a book/ *a picture. For this reason, complements seem to be more defining and specifying of the head while adjuncts can be said to be more describing through providing additional information. Therefore, it seems likely to have more frequent complements than adjuncts in the scientific medical genre, which is not the case in this corpus. A possible explanation may reside in the hybrid nature of this corpus and the differences between AAs and PAs, which will be verified in the next section.

2.2.2 Adjunct and complement postmodifiers across the genres

Table 3 shows the distribution of adjunct and complement postmodifiers across the genres.

Table3. Distribution of adjunct and complement postmodifiers across the genres

	Academic Articles	Popular Articles
Adjunct Postmodifiers	954 41%	1065 52%
Complement Postmodifiers	1336 59%	979 48%
Total	2290 100%	2044 100%

As shown in the table above, in AAs complement postmodifiers (59%) outnumber adjunct postmodifiers (41%) whereas in PAs, the reverse pattern is true. To consolidate these findings, the Chi-square test is calculated. $\chi^2 = 45.72$, which makes it possible to reject the null hypothesis (with 1 freedom degree, χ^2 is insignificant only if it is ≤ 3.84). This implies

that this distribution is not random and that the probability of occurrence of adjunct or complement postmodifiers can be interpreted in relation to genres.

This distribution may be attributed to the difference between PAs and AAs. In fact, Parkinson and Adendorff (2004) assert that academic or research articles show a high degree of formality through their "technical language, nominalisation, passivisation, impersonal tone" (389). For this reason, academic articles are considered as more authoritative (Mahoney, 2009) and more representative of the scientific jargon, as can be seen in the following examples.

Example3, AA 2

relative increase [in EC repopulation rates] $_{comp}$

Example4, AA 2

stratification [of EGFR + and EGFR- Glioma] $_{comp}$

Example 5, AA1:

the αvβ3 receptor [, which may contribute to angiogenesis,] adj

Example6, AA 6

G-CSF pre-treated allograft [at 30 days post-transplant] $_{adj}$

As can be seen from these examples, AAs tend to use a "specialised" and "advanced" English (Lowe, 2009: 1) in the sense that a layperson needs a scientific knowledge in order to understand such words as "EGFR+" "EGFR-" "Glioma", "angiogenesis", "$\alpha v\beta_3$", "G-CSF". For this reason some specification and clarification is needed. Thus the relationship between the head and its postmodifier should be as close (syntactically and semantically) as possible. The fact that complements are syntactically closer to their heads than adjuncts (in the tree) as well as semantically (complements are obligatory) justifies the more recurrent use of complement postmodifiers at the expense of adjunct postmodifiers.

Popular articles, on the other hand, "tend to employ a popular style of writing" (Nwogu, 1990: 126) through "giv[ing] voice to scientists other than those of iconic status (the 'great names' of science such as Einstein)", (Parkinson & Adendorff (2004: 12). In such a way, they help "maintain a vital relationship between researchers and the general public" (Giannoni, 2008: 1) by "correspond[ing] to the needs of less informed groups of readers such as students and laymen" (Varttala, 1991: 2). Some random examples of adjunct and complement postmodifiers in PAs are provided below.

Example7, PA 9

fear [of encountering germs] $_{comp}$

Example8, PA3

its next response [to the same viral strain] $_{comp}$

Example9, PA3

therapeutic genes [, which must find their way into millions of cells,] $_{adj}$

Example10, PA 7

foods [containing saturated fats or pure cholesterol] $_{adj}$

These examples show that the language used in PAs is "common" (Lowe, 2009: 1), simpler and easier to understand. Targeting laypeople lacking scientific background, PAs should thus be as explicit and clear as possible as far as the lexis as well as the grammatical relationships are concerned. In studying explicitness in postmodification, Quirk et al. (2005: 1243) argue that explicitness is greater in the finite relative clause (where the relative pronoun, the action and the tense are given), than in the non finite relative clause (where only the action is given with no relative pronoun nor tense). This in turn is less explicit than the prepositional phrase (where all three elements disappear). PAs are therefore expected to have more finite relative clauses and non-finite relative clauses than prepositional phrases. Since these two forms are considered as adjuncts (Herbst, 1988: 272), PAs are thus more probable to display a higher proportion of adjunct than complement postmodifiers, which is confirmed in this corpus. The distribution of adjunct and complement postmodifiers is thus genre-related.

Conclusion

This paper has attempted to prove that the distribution of adjunct and complement postmodifiers is genre-affected. To this end, a quantitative analysis was conducted to check this distribution within the whole medical corpus and more importantly within the two genres of academic and popular articles. Two main findings were statistically evidenced. First, when investigating the corpus as a whole, complement postmodifiers and adjunct postmodifiers are found to have almost the same frequency distribution with a slight difference, which seems surprising with regard to the medical genre. Second, AAs favour complement postmodifiers than adjuncts, while PAs are characterized by the reverse pattern. This is attributed to the fact that each genre has its own communicative purposes as well as formal features.

For more validation of these findings, this study could be extended in two other ways. Other case studies investigating further linguistic features while keeping the same generic framework could be carried out. Similarly, adjunct and complement postmodifiers may be the heart of another corpus study focusing on other genres.

References

Broderick, J. P. (1975). *Modern English Linguistics: A Structural and transformational Grammar*. New York: Thomas Y. Crowell Company.

Carnie, A. (2001). *Syntax*. Oxford: Blackwell.

CHOMSKY, N. (1970). Remarks on Nominalization. In R. Jacobs & P. Rosenbaum (Eds.), *Reading in English Transformational Grammar* (pp. 184-221). Walthman, Mass: Ginn.

CHOMSKY, N. (1975). *Aspects of the* Theory *of Syntax* (tenth ed.). Massachusetts: The M.I.T Press

DOWTY, D. The Dual Analysis of Adjuncts/Complements in Categorical Grammar. Retrieved from http://www.ling.ohio-state.edu/~dowty/papers/degruyter.8x11.pdf

GIANNONI, D. S. (2008). Popularizing features in English journal editorials. *English for Specific Purposes* (27), 212-232.

GLEASON H A, J. (1965). *Linguistics and English Grammar*. New York: Holt, Rinehart and Winston.

HALLIDAY, M. A. K. (2004). *The language of Science*. London: Continuum.

HALLIDAY, M. A. K., & MARTIN, J. R. (1993). *Writing Science: Literacy and Discursive Power. Great Britain*: The Falmer Press.

HALLIDAY, M. A. K., & MATTHIESSEN, M. I. M. (2004). *An Introduction To Functional Grammar. Great Britain*: Oxford University Press.

HERBST, T. (1988). A Valency Model for Nouns in English. *Journal of Linguistics, 24*(2), 265-301.

HUDDLESTON, R. (1989). *An Introduction to English Transformational Syntax* (6 th ed.). New York: Longman.

KROEGER, P. R. (2005). *Analyzing Grammar: An Introduction*. London: Cambridge University Press.

LOWE, I. (2009). Characteristics of the language of science. Retrieved from www.scientificlanguage.com/esp/characteristics-language-science.pdf.

LYNCH, J. (2008). Guide to Grammar and Style. Retrieved from www.andromeda.rutgers.edu/jlynch/writing

MAHONEY, S. (2004). Academic Journals vs. Popular Magazines. Retrieved from www.mccneb.edu/library/resources/journalsvsmagazines.asp

MATTHEWS, P. H. (1981). *Syntax*. Cambridge: Cambridge University Press.

MILLER, J. (2002). *An Introduction to English Syntax*. Great Britain: Edinburgh University Press Ltd.

NWOGU, K. N. (1990). *Discourse Variation in Medical Texts: Schema, Theme and Cohesion in Professional and Journalistic Accounts*. England: Department Of English Studies, University of Nottingham.

PARKINSON, J., & ADENDORFF, R. (2004). The use of popular science articles in teaching scientific literacy. *English for Specific Purposes (23),* 379-396.

QUIRK, R., GREENBAUM, S., LEECH, G., & SVARTVIK, J. (2005). *A Comprehensive Grammar of the English Language. London: Longman.*

RICHARDS, N. (2002). *Introduction to Syntax: Phrase Structure*. London: Oxford University Press.

VALIN, R. D., & LAPOLLA, R. (1997). *Syntax: Structure, Meaning and Function*. UK: Cambridge University Press.

VARTTALA, T. (1999). Remarks on the communicative functions of hedging in popular scientific and specialist research articles on medicine. *English for Specific Purposes* (*18*), 177-200.

Cognos toolkit: un conjunto de herramientas para la anotación lingüística de corpus

Garazi Olaziregi Gómez. *Universidad Carlos III de Madrid*

Francisco Javier Calle Gómez. *Universidad Carlos III de Madrid*

Esperanza Albacete Garcia. *Universidad Carlos III de Madrid*

Dolores Cuadra Fernández. *Universidad Carlos III de Madrid*

Alejandro Baldominos Gómez. *Universidad Carlos III de Madrid*

David del Valle Agudo. *Universidad Carlos III de Madrid*

Abstract: *In this paper we present Cognos Toolkit, a set of tools (methodology, formalization scheme based on XML and software components) which allow do the integral process of annotation the pragmatic information of a corpus, facilitating the training of Models of Dialogue. To do this, firstly we analyzed the existing software oriented to the Corpus annotation, as well as proposals based on XML encoding to manage the pragmatic knowledge. Next, we also describe Cognos Toolkit (the proposed methodology and software components) and the projects that have used its tools (Domain Interaction and extent of Corpus). Finally, we describe future extensions and improvements of the toolkit.*

Keywords: *Pragmatic annotation, Natural Interaction corpora, Annotation tool*

Resumen: *En esta comunicación presentamos Cognos Toolkit, un conjunto de herramientas (metodología, esquema de formalización basado en XML y componentes software) que permiten realizar de forma íntegra el proceso de anotación de información pragmática de Corpus, facilitando así el entrenamiento de Modelos de Diálogo. Con ese fin, hemos analizado las aplicaciones informáticas existentes orientadas a la anotación de Corpus, así como de los esquemas de codificación propuestos basados en XML para tratar el conocimiento pragmático. Asimismo, describimos Cognos Toolkit (la metodología propuesta y los componentes software) y los proyectos en los que se han utilizado sus herramientas (Dominio de Interacción y extensión de Corpus). Por último, describimos las extensiones futuras y mejoras del conjunto de herramientas.*

Palabras calve: *Anotación pragmática, Corpus de Interacción Natural, herramienta de anotación*

1. Introducción

Cada día son más los sistemas de interacción hombre-máquina que persiguen imitar el comportamiento humano para hacer más cómoda y accesible esa interacción, incluso para personas sin entrenamiento tecnológico específico. Estos sistemas suelen basarse en modelos de conocimiento que aglutinan los procesos cognitivos que intervienen en el proceso interactivo, en cada una de sus facetas (Calle, Martínez, Valle-Agudo y Cuadra, 2009). La mayoría de ellos cuentan con interfaces basadas en Lenguaje Natural y un Modelo de Diálogo para interpretar y reproducir ese comportamiento humano, intercambiándose mensajes verbales y reflejando el efecto que ese intercambio produce en los interlocutores.

Gran parte de las bases de conocimiento que soportan estos modelos de diálogo se nutren de un corpus definido *ad hoc* para ese dominio y adquirido en situaciones reales. Ese corpus necesita ser tratado para poder extraer de él la información útil. Se trata de una tarea costosa que a menudo se desarrolla de un modo rígido, resultando casi impracticable su reutilización, ni siquiera sobre los mismos modelos para los que fue obtenido.

En este trabajo se presenta Cognos Toolkit, un conjunto de herramientas que facilita el análisis y anotación completa e integral Corpus, ofreciendo una metodología, un esquema XML para las anotaciones y un conjunto de aplicaciones software que agilizan este proceso. Incluye herramientas especialmente orientadas a la anotación del conocimiento pragmático en diálogos, permitiendo a los expertos anotar toda la información relevante de este nivel lingüístico, y estructurándola de acuerdo con un esquema definido que permite la utilización de esta información en el entrenamiento de modelos de diálogos futuros. Por su carácter general, independiente del modelo sobre el que se aplique el producto, y por contar con formalizaciones estandarizadas de las anotaciones, favorece la reutilización y compartición de corpus anotado, reduciendo así los costes producidos y aumentando la calidad del resultado al poder observar corpus de mayor volumen.

2. Trabajos relacionados

Este apartado realiza un recorrido por aquellos trabajos previos relacionados con esta investigación, divididos en los que proponen formalizaciones de la anotación pragmática en XML y las herramientas de anotación lingüística de Corpus.

Así, pueden encontrarse formalizaciones como *HCRC Dialogue Structure Coding* (Carletta, Isard, Isard, Kowtko, Doherty-Sneddon y Anderson, 1996) diseñado específicamente para *HCRC Map Task Corpus*. Esta propuesta distingue los siguientes niveles: movimientos conversacionales (unidad mínima interactiva), juegos conversacionales (conjunto de movimientos con un propósito común), y transacciones (conjuntos de juegos que realizan algún propósito general del diálogo).

Si bien es cierto que *HCRC Dialogue Structure Coding* está orientado a un dominio de interacción específico, es de las pocas propuestas que se realizan atendiendo a la estructura profunda del discurso (metas e intenciones subyacentes), junto con VerbMovil-2 (Alexandersson, Buschbeck-Wolf, Fujinami, Kipp, Koch, Maier, Reithinger, 1998),

proyecto Monroe (Tetreault, Swift, Prithviraj, Dzikovska, y Allen, 2004), y DIHANA (Alcácer, Benedí, Blat, Granell, Martiniez, and Torres, 2005).

Otras propuestas como DAMSL (Allen y Core, 1997) únicamente abordan la estructura profunda del discurso para especificar si una expresión dada está orientada al propósito general de la interacción o no. Sin embargo, tanto DAMSL como sus extensiones (SWDB-DAMSL (Jurafsky, Shriberg y Biasca, 1997), ADAM (Cattoni, Danieli, Panizza, Sandrini, y Soria, 2001), SPAAC (Leech, Geoffrey y Weisser, 2003) sí tratan el análisis desde la estructura superficial (actos de habla y pares de adyacencia). Es de especial interés SAPAAC, ya que introduce una taxonomía genérica de actos de habla reutilizable en distintos dominios, con sus 41 actos distribuidos en 5 clases distintas.

La tarea de la anotación de Corpus puede resultar tediosa, y más cuando se trata de un corpus extenso, como se recomienda para el entrenamiento de un Modelo de Diálogo. Por esta razón se han desarrollado distintas herramientas que facilitan la anotación lingüística, cada una de ellas atendiendo a distintas necesidades. A continuación se exponen algunas, clasificadas en tres grupos:

- Herramientas configurables: Las herramientas de este grupo soportan diversas capas o niveles lingüísticos que pueden configurarse en función de las necesidades concretas. Estas herramientas, aunque versátiles, requieren una configuración previa que en ocasiones afecta incluso a la interface gráfica, suponiendo un elevado coste, más cuando el usuario no es un experto desarrollador software. Cabe destacar **ELAN** (Brugman y Russel, 2004), donde las capas de anotación deben configurarse para determinar las dependencias entre ellas y sincronizarse con la línea temporal. Estas características son similares a las de **Anvil** (Neff, 2008), aunque este segundo está orientado a la anotación multimodal, sincronizándose también con los archivos de video y de audio, de la misma forma que lo hacen **NOMOS** (Niekrasz y Gruenstein, 2006) y **NITE XML** (Popescu-Belis, 2010), esta última con una interfaz gráfica incluida.
- Anotación pragmática: Aunque existen herramientas que abordan fases del proceso de anotación pragmática, no parece que ninguna solvente el problema íntegramente. Así, para las primeras etapas (dividir y anotar roles del diálogo), puede emplearse **Transcriber** (Barras, Geoffrois, Wu y Liberman, 2001) y exportar el resultado eligiendo el esquema de anotación pragmática más conveniente.
- Otros niveles lingüísticos: Existen numerosas herramientas de anotación sintáctica, morfológica y semántica, como son **GATE** (Cunningham, Maynard, Bontcheva, y Tablan, 2002), **NLTK (**Loper y Bird, 2002) o **Stanford-Parser** (Klein y Manning, 2003), que no abordando el nivel pragmático pero son un buen referente.

3. Metodología

Cognos Toolkit entre sus herramientas ofrece una metodología de anotación cuyas etapas se muestran en la ilustración 1.

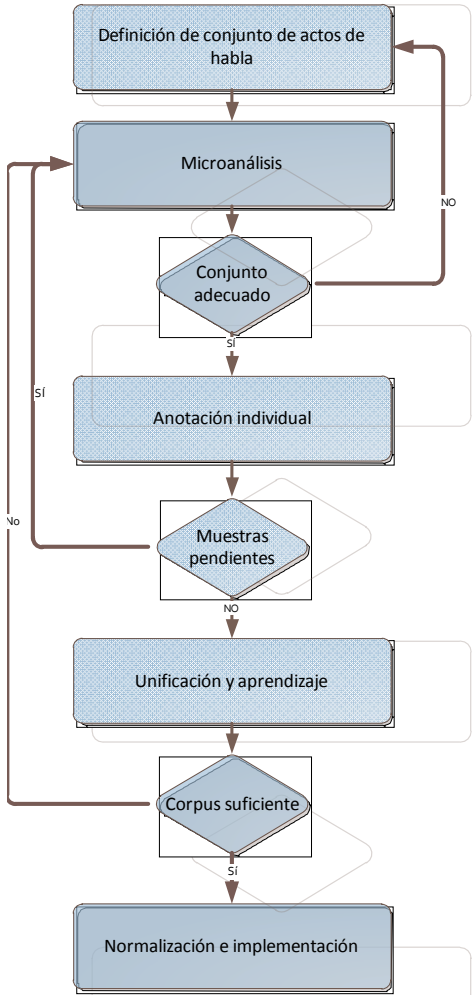


Ilustración 1: Metodología de anotación

La primera etapa consiste en la definición del conjunto de actos de habla que se utilizará durante todo el proceso. En esta fase, opcionalmente, el usuario podrá caracterizar los actos del conjunto, aportando información requerida por el sistema con el fin de poder taxonomizar el conjunto de acuerdo con unos criterios definidos.

La fase de microanálisis consiste en dividir la interacción en las intervenciones que el usuario considere oportunas. Una vez dividido el corpus se debe describir cómo se distribuyen las intervenciones en el tiempo, anotando pausas y solapamientos entre intervenciones. A continuación se verifica si las intervenciones son oraciones válidas en una gramática relajada vinculada al corpus. Esta gramática puede ser ampliada en cualquier momento, incluyendo patrones para las expresiones que inicialmente no sean

válidas en dicha gramática. La realización de esta fase conlleva la comprobación de que el conjunto de actos diseñado cubre todas las necesidades expresivas del dominio de interacción (si esto no se cumple sería necesario completar ese conjunto volviendo a la primera fase).

La tercera fase es la anotación de muestras individuales. En primer lugar se identifican los segmentos –y la relación entre ellos- y sus intenciones. Seguidamente, se estudian por un lado las consecuencias que pueda tener cada evento en la atención de los interlocutores, y por otro las técnicas empleadas por los interlocutores para reforzar el compromiso. Por último, en esta fase se anotan los actos perlocucionarios de cada intervención, esto es, los efectos de las intervenciones tales como la consulta de información.

Una vez se hayan anotado todas las muestras individuales comienza la quinta fase. En esta fase se unifican los actos perlocucionarios anotados en la intervención en función de la similitud en sus entradas, salidas y efectos, y se aprenden la gestión de la atención y la del compromiso. Tras la unificación, deberá comprobarse la cobertura del corpus. Si existieran operaciones o efectos deseados no contemplados en el corpus, deberá revisarse la definición del dominio de interacción añadiendo los escenarios oportunos, y en consecuencia habrá que adquirir Corpus adicional para estos escenarios y reactivar con la fase de microanálisis (fase 2).

La última etapa de la metodología consiste en la normalización de las intenciones identificando aquellas que sean equivalentes, y finalmente se implementan los hilos de diálogos. La implementación de los hilos se realiza de acuerdo con un modelo de diálogo concreto, por lo que se ha retrasado al máximo esta tarea, pudiendo cambiar el modelo de diálogo y reutilizar todas las etapas anteriores de la anotación.

4. Componentes software

Aunque se prevé que en un futuro se incluyan aplicaciones para la anotación de elementos paralingüísticos, en la actualidad Cognos Toolkit cuenta con tres aplicaciones software que abordan la anotación lingüística (Cognos.CA, Cognos.DIAL y Cognos.NL), que complementándose permiten realizar análisis íntegros y completos, ya que cada una de ellas opera sobre un ámbito distinto. Además, Cognos Toolkit cuenta con una base de conocimiento común para todas las aplicaciones, convirtiéndose en un hilo comunicador entre ellas y elemento integrador de las anotaciones que se realizan.

A continuación se describen todos los componentes software, así como los módulos que la integran:

4.1. Base de Conocimiento:

Todas las aplicaciones de Cognos Toolkit acceden a la misma base de conocimiento, tanto para consultarla como para actualizarla. En esta base de conocimiento están almacenados todos los corpus analizados, los escenarios que los integran y cada una de las muestras de dichos escenarios. Una vez conectado, el usuario puede seleccionar una muestra existente

(y por tanto el escenario y el corpus al que pertenezca), o bien crear muestras, escenarios o corpus nuevos, partiendo de transcripciones en ficheros de texto o creándolas desde el propio toolkit.

Toda muestra puede ir acompañada de un archivo de audio que contenga su grabación, posibilitando su reproducción posterior para facilitar distintas fases. Además, toda la información anotada con las distintas aplicaciones de Cognos Toolkit es almacenada en la base de conocimiento, pudiendo guardarse en cualquier momento del análisis, y abrirse posteriormente para ser consultada, o para realizar cambios o ampliar la anotación guardada.

4.2. Cognos.CA:

Para poder realizar una anotación pragmática es necesario definir previamente el conjunto de actos de habla que se utilizará durante el resto del análisis. Cognos.CA es la herramienta de Cognos Toolkit que permite definir este conjunto de actos, posibilitando caracterizarlos conforme a criterios definidos a través de la propia experiencia investigadora y partiendo de la propuesta de (Searle, 1975). Esta caracterización permite taxonomizar el conjunto de actos y ofrece al analista la posibilidad de visualización del conjunto en forma arbórea facilitando la verificación de la cobertura del conjunto, esto es, la detección de solapamientos y ausencias de actos de habla.

Aunque la metodología definida no utiliza un conjunto de actos estándar sino que incluye la creación de un conjunto de actos para cada corpus, Cognos.CA permite adaptar el conocimiento pragmático anotado a un conjunto de actos distinto. Esta adaptación es posible gracias a la realización de la taxonomía en función de criterios comunes que permite equiparar dos conjuntos caracterizados cualesquiera de forma automática en el mejor de los casos, y semiautomática en el caso de que se requiera la toma de decisiones del especialista para la resolución de los conflictos que surjan en el proceso.

4.3. Cognos.NL

La aplicación Cognos.NL permite crear gramáticas relajadas partiendo de las expresiones en Lenguaje Natural del corpus. Para ello se asocia a cada expresión una estructura semántica que recoge el significado literal de la misma. Este proceso se divide en dos subprocesos: el primero consiste en asociar actos comunicativos a las expresiones, almacenándose estos vínculos en la base de conocimiento de forma que puedan ser consultados desde la pestaña microanálisis de Cognos.Dial.Indiv.

El segundo subproceso consiste en determinar los elementos de la expresión en Lenguaje Natural que son característicos para asociar dicha expresión a los actos con que se vincula. Aquellos elementos que sean prescindibles para ese fin pueden ser comodines (completamente irrelevantes) o variables (cuyo valor es contenido proposicional anotado en el acto comunicativo).

Los elementos característicos, por su parte, pueden ir asociados a subpatrones – autónomos para construir una expresión por si solos, y por tanto con su propia secuencia

de actos comunicativos asociada- o *tokens* –que necesariamente requiere estar embebido en una expresión mayor-. Cada *token* se realiza con términos, pudiéndose diferenciar las instancias por elementos circunstanciales, pero siendo equivalentes por lo demás. Cognos.NL permite elegir para cada *token* un término previamente almacenado en la base de conocimiento, o por el contrario crear uno *ad hoc* para la expresión cuya anotación está en curso. Asimismo, se almacenan todos los *tokens* en la base de conocimiento, evitando inconsistencias y redundancias, puesto que Cognos.NL ofrece sugerencias en el análisis basándose en el conocimiento anotado previamente.

Como resultado final –con la totalidad del corpus anotado- se obtiene un conjunto de patrones con una secuencia de actos comunicativos cada uno, que serán asignados automáticamente a las expresiones del corpus que encajen en dichos patrones.

4.4. Cognos.Dial

La aplicación de Cognos Toolkit que permite llevar a cabo las anotaciones de conocimiento pragmático sobre un corpus es Cognos.Dial. Se divide en tres herramientas independientes que abordan las distintas etapas, proporcionándoles autonomía, aunque produciendo un resultado íntegro por operar sobre la misma base de conocimiento.

La primera de las herramientas es Cognos.Dail.Indiv, que aborda las anotaciones de las muestras individuales. La interfaz de esta herramienta cuenta con pestañas que se corresponden con fases del proceso. Cada una de las pestañas irá habilitándose a medida que se finalice y valide la fase anterior, evitando de esta forma errores de inconsistencia entre las distintas fases del análisis. A continuación se enumeran las pestañas de Cognos. Dial.Indiv:

- **Pre-segmentación:**

  Preparación de la muestra para la anotación. El analista puede cargar desde un archivo de texto la transcripción de la interacción o realizarla directamente desde la interfaz. En esta transcripción debe dividirse la interacción en intervenciones asignándole un interlocutor a cada una de ellas. Estas intervenciones serán indivisibles durante el resto del análisis.

- **Realización temporal:**

  Consiste en distribuir en la línea temporal las intervenciones, solapamientos, y silencios (intervalo/apelativo, lapso/anuncio, pausa o pausa oralizada). Asimismo, si se dispone de una grabación de audio de la interacción puede adjuntarse a la anotación.

- **Microanálisis:**

  En esta fase el anotador asigna a cada intervención uno o varios actos de habla del conjunto definido con Cognos.CA, definiendo su contenido proposicional y sus variables.

  Este proceso es complementario al realizado con Cognos.NL, por lo que esta pestaña integra la funcionalidad de asignar automáticamente los actos de habla a las intervenciones utilizando el conocimiento anotado con Cognos.NL.

- **Segmentación:**

  Fase en la que el analista identifica las secuencias de intervenciones que se realizan con una intención común que el analista anotará. Cada una de las intervenciones de un segmento se anotará de acuerdo con su función en el segmento. La herramienta permite anotar segmentos embebidos en otros segmentos, o segmentos discontinuos, así como restricciones de precedencia entre segmentos.

- **Compromiso:**

  Cognos.Dial.Indiv incluye la posibilidad de anotar información relacionada con la gestión de la atención y el compromiso desde esta pestaña. Esta información se refiere tanto a los eventos que afectan a la atención y el compromiso de los segmentos como a las técnicas empleadas por los interlocutores para reforzarlos.

  Este conocimiento anotado se unificará posteriormente con la herramienta Cognos.Dial.Global, con la finalidad de aprender qué eventos afectan a la atención y el compromiso, y la conveniencia de emplear qué técnica de refuerzo y en qué circunstancias.

- **Operativo:**

  En esta pestaña se anotan aquellas tareas no interactivas que realizan los interlocutores, resultantes de la interacción, es decir, sus actos perlocucionarios. Desde la perspectiva de la máquina, estos actos son tareas externas o aplicaciones, y su anotación implica identificar sus entradas y salidas, así como los efectos que los resultados de esas tareas tendrán sobre la interacción.

- **Estructural:**

  Finalmente se permite visualizar el resultado de las fases anteriores como un conjunto de autómatas de estados finitos (uno por segmento anotado). Cada autómata está definido como un conjunto de estados unidos por transiciones. A los estados se vinculan las tareas anotadas y a las transiciones las expresiones lingüísticas de la muestra, y por tanto sus actos de habla asociados. La herramienta permite en esta pestaña consultar la información anotada, y modificarla creando, eliminando y editando estados y transiciones. Todas las modificaciones realizadas en esta pestaña, aunque no se incluirán en la exportación de ficheros XML, serán tenidas en cuenta en la unificación realizada por Cognos.Dial.Global.

Una vez anotadas todas las muestras individuales, el analista podrá unificar la información anotada con Cognos.Dial.Global (segunda herramienta de la aplicación Cognos.Dial), seleccionando las muestras anotadas que desea integrar con el fin de alimentar un modelo de diálogo concreto. Cognos.Dial.Global facilita la identificación de tareas equivalentes, así como segmentos similares, a través de la funcionalidad de sugerencia que calcula automáticamente las parejas candidatas a fusionarse. Además permite el aprendizaje de la gestión de la atención y el compromiso.

La tercera herramienta de Cognos.Dial es Cognos.Dial.Eval, orientada a la evaluación de la concordancia entre dos o más anotaciones de la misma muestra. La herramienta calcula el factor Kappa de cada aspecto de la anotación, y un factor Kappa agregado mediante

un vector de pesos (configurable) como medida ponderada de la divergencia entre las anotaciones de la misma muestra.

5. Resultados

La metodología de anotación propuesta en este trabajo ha sido empleada en los proyectos IntegraTV4All[158], SOPAT[159] y SemAnts[160], y en este último se emplearon también los componentes software. En la actualidad se está empleando Cognos Toolkit en el proyecto Thuban, cuyo dominio es la interacción con un acompañante virtual que ofrece un conjunto de servicios (descripción, guiado, etc.) a usuarios itinerantes (realidad aumentada).

6. Conclusiones y trabajos futuros

Cognos Toolkit tiene como objetivo agilizar la anotación del conocimiento pragmático sobre Corpus, independientemente de su Dominio de Interacción. Además, permite reutilizar Corpus anotado de proyectos anteriores, lo que junto con la agilidad obtenida causa una notable reducción de costes y posibilita aumentar el volumen del corpus tratado, y por ende del conocimiento adquirido sobre el dominio, redundando en una mayor calidad del resultado. Asimismo, si bien es cierto que en la actualidad Cognos Toolkit se restringe a la anotación del conocimiento lingüístico, se prevé que un futuro cercano contemple la anotación de más conocimiento integrado, con nuevos componentes como Cognos.E (conocimiento emocional), Cognos.S (conocimiento de situación) o Cognos. Onto (ontologías) y Cognos.U (conocimiento de usuario), estas últimas ya en proceso de desarrollo.

7. Bibliografía

Alcácer, N., Benedí, J.M., Blat, F., Granell, R., Martiniez, C.D. and Torres, F. (2005). Acquisition and Labelling of a Spontaneous Speech Dialogue Corpus. *Procs of SPECOM 2005,* 583-586.

Alexandersson, J., Buschbeck-Wolf, B., Fujinami, T., Kipp, M., Koch, S., Maier, E., Reithinger, N. (1998). Dialogue Acts in VERBMOBIL-2 Second Edition. *Arbeit*, (July)

Allen, J., & Core, M. (1997). Dialog Act Markup in Several Layers (Draft 2.1).

Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (2001). Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1-2), 5-22.

Brugman, H., & Russel, A. (2004). Annotating Multi-Media / Multi-Modal resources with ELAN. *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (pp. 2065-2068). ELRA.

158 44 diálogos cuyo dominio es de servicios de asistencia, ayuda y contenidos interactivos adaptados a personas con alguna discapacidad sensorial.

159 30 diálogos genéricos y 53 específicos del dominio de guiado, posicionamiento y descripción de lugares.

160 59diálogos de descripción de rutas y solicitud de información turística.

CALLE, F.J., MARTÍNEZ, P., VALLE-AGUDO, D. Y CUADRA, D. (2009). Towards the Achievement of Natural Interaction, In: Redondo, M., Bravo, C., Ortega, M. (eds.) *Engineering the User Interface: from Research to Practice*. vol. 12 (pp. 63–79). Heidelberg: Springer.

CARLETTA, J., ISARD, A., ISARD, S., KOWTKO, J., DOHERTY-SNEDDON, G., & ANDERSON, A. (1996). HCRC dialogue structure coding manual.

CATTONI, R., DANIELI, M., PANIZZA, A., SANDRINI, V., & SORIA, C. (2001). Building a corpus of annotated dialogues: the ADAM experience. *Proceedings of Corpus Linguistics*

CUNNINGHAM, H., MAYNARD, D., BONTCHEVA, K., & TABLAN, V. (2002). GATE: A framework and graphical development environment for robust NLP tools and applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics* (Vol. 54, pp. 168-175).

JURAFSKY, D., SHRIBERG, E., BIASCA, D. (1997). *Switchboard SWBD-DAMSL Labeling Project Coder's Manual*, Draft 13. Technical Report 97-02, University of Colorado Institute of Cognitive Science

KLEIN, D., & MANNING, C. D. (2003). Fast Exact Inference with a Factored Model for Natural Language Parsing. (S. T. S Becker & K. Obermayer, Eds.)*Advances in Neural Information Processing Systems*, *15*, 3–10. Citeseer.

LEECH, G., & WEISSER, M. (2003). Generic speech act annotation for task-oriented dialogues. *International Journal of Corpus Linguistics* (Vol. 16, pp. 1-6). Centre for Computer Corpus Research on Language Technical Papers, Lancaster University.

LOPER, E., & BIRD, S. (2002). NLTK: The Natural Language Toolkit. *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, (July), 8.

NEFF, M. (2008). Gesture Modeling and Animation Based on a Probabilistic Recreation of Speaker Style. *ACM Transactions on Graphics*, 27(1). ACM Press.

NIEKRASZ, J., & GRUENSTEIN, A. (2006). NOMOS: A Semantic Web Software Framework for Annotation of Multimodal Corpora. *Proceedings of the 5th International Conference on Language Resources and Evaluation LREC*.

POPESCU-BELIS, A. (2010). Managing Multimodal Data, Metadata and Annotations: Challenges and Solutions. En *Multimodal Signal Processing. Theory and Applications for Human–Computer Interaction* (pp. 207-228) Science Direct. Academic Press. Elsevier

SEARLE, J. R. (1975). A Taxonomy of Illocutionary Acts. In K. Gunderson (Ed.), Minnesota Studies in the Philosophy of Science Volume VII Language Mind and Knowledge (pp. 344-369

TETREAULT, J., SWIFT, M., PRITHVIRAJ, P., DZIKOVSKA, M., AND ALLEN, J. (2004). Discourse Annotation in the Monroe Corpus. In Webber, B., and D. K. Byron (Eds.), *ACL 2004 Workshop on Discourse Annotation.* Barcelona.

Análisis Léxico de Unidades Léxicas Compuestas

Marc Ortega Gil.
Universidad Autónoma de Barcelona
marc.ortega@uab.es

Resumen

En este artículo se quiere mostrar cómo se realiza el análisis de unidades léxicas compuestas como las locuciones, los tiempos verbales compuestos y las locuciones verbales en español, en el marco del sistema de análisis léxico del proyecto FrameNet Español[161] *basado en un diccionario electrónico formado por 634.500 formas, simples y compuestas, y un conjunto de gramáticas y herramientas construidas tomando las máquinas de estado finito como modelo matemático. El análisis de estos elementos se realiza sobre un corpus de oraciones anotadas léxicamente, de modo que cada unidad léxica (palabra) se anota con su correspondiente categoría léxica y sus características morfológicas, como en el caso de los verbos, nombres y adjetivos.*

Abstract

This article aims to show how to perform lexical analysis of multiword lexical units like compound tenses and verbal phrases in Spanish, in the context of the lexical analysis system developed in the Spanish FrameNet project, based on an electronic dictionary consisting of 634,500 forms, simple and compound, and a set of grammars and tools built to take the finite state machine as mathematical model. The analysis of these elements is performed on a corpus of lexically annotated sentences, so that each lexical unit (word) is annotated with its corresponding lexical category and morphological characteristics, as in the case of verbs, nouns and adjectives.

Palabras Clave: lingüística de corpus, análisis léxico, unidades léxicas compuestas, máquinas de estados.

Key Words: Corpus linguistic, lexical analysis, multiword lexical units, finite state machines.

161 **FrameNet Español**: un recurso léxico para el procesamiento semántico automático del español (FFI2008-0875), es un proyecto de investigación de semántica léxica y lingüística de corpus.

1. Introducción

El sistema de análisis en el que se enmarca este trabajo se realiza sobre un corpus de oraciones anotadas léxicamente, de modo que cada unidad léxica (palabra) se anota con su correspondiente categoría léxica y sus características morfológicas, como es el caso de los verbos, nombres y adjetivos, y permite reconocer tanto formas simples como formas locutivas. Dentro de estas últimas se analizan tanto las que se pueden reconocer a partir de un diccionario, como p. ej. *en todo momento*, como las que requieren un análisis sintáctico posterior al análisis léxico inicial para poder ser reconocidas. Este es el caso de locuciones verbales como *dar por sentado*, que puede aparecer como *da [siempre muchas cosas] por sentado*, o de los tiempos verbales compuestos en español. En estos casos el reconocimiento de la unidad léxica no puede llevarse a cabo únicamente a partir de un diccionario o de procedimientos estadísticos (Collins, 1999) y se requiere un análisis sintáctico más profundo que permita identificar como una unidad las formas que constituyen la unidad léxica locutiva y anotarla con su correspondiente categoría léxica y sus características morfológicas, a la vez que los elementos que no pertenecen a la parte conexa de la locución, como *siempre muchas cosas* del ejemplo anterior, se sitúan en el contexto derecho, o izquierdo si fuera necesario, de la unidad locutiva, p. ej. *[dar/por/sentado] siempre muchas cosas* (Subirats y Ortega 2000).

El análisis de estas unidades locutivas se realiza en el marco de un sistema de análisis textual basado en técnicas de estado finito (*finite state methods*) en el que el análisis de las locuciones y los tiempos verbales compuestos se realiza a partir de un conjunto de gramáticas locales representadas como transductores subsecuenciales (Mohri, 1997) que se aplican, mediante un proceso de transducción, sobre autómatas finitos deterministas que representan oraciones anotadas léxicamente a partir de un diccionario electrónico.

El análisis de estas unidades léxicas locutivas permite, a la vez que son reconocidas y etiquetadas, desambiguar de forma eficiente los casos de ambigüedad (Laporte, 2001) como el que aparece con la forma *sentado* del ejemplo anterior, que se asocia a dos categorías distintas: (1) como forma del participio del verbo *sentar* y (2) como adjetivo. El análisis léxico basado en técnica de estado finito, en el que las oraciones se representan como autómatas finitos, permite representar y manipular de forma eficiente los casos de unidades léxicas ambiguas, es decir, aquellas unidades léxicas que como el caso de la forma *sentado*, están asociadas a dos o más clases de palabra o propiedades morfológicas. El análisis a partir de gramáticas locales, representadas como transductores subsecuenciales, permite eliminar gran parte de estas ambigüedades, con un margen de error prácticamente inexistente, de durante el análisis de las formas locutivas.

2. Análisis léxico

El análisis de formas locutivas que se presenta se enmarca dentro del sistema de análisis léxico desarrollado en el proyecto FrameNet Español (FNE). Este sistema de análisis está formado por:

1. Un sistema de diccionarios electrónicos que contiene 634.500 formas, simples y compuestas,

2. un proceso de análisis léxico que permite, a partir del diccionario, analizar un texto y anotar con información léxica las formas simples y las formas locutivas que pueden ser reconocidas a partir del diccionario, como por ejemplo *bomba atómica*,
3. un proceso de transducción l que permite reconocer aquellas formas locutivas que por su naturaleza no pueden reconocerse a partir del diccionario ya que contiene construcciones sintácticas entre los elementos que forman la parte conexa de la locución.

2.1. El diccionario electrónico

El diccionario electrónico se genera a partir de un conjunto de diccionarios de lemas en los que cada uno de ellos está asociado a una categoría léxica con sus especificaciones flexivas, en los casos en los que estas puedan tener flexión morfológica. El lema es la forma que utilizan los diccionarios como modelo para definir sus entradas, es decir, el verbo en infinitivo y los nombres y los adjetivos, en singular, y en masculino cuando tienen flexión de género. Cuando un lema está formado por una única secuencia de caracteres, se denomina forma simple. Si un lema está integrada dos o más secuencias de caracteres, separadas por espacios en blanco, se denomina forma compuesta o locución. Las cadenas de caracteres que integran las formas compuestas o locuciones forman parte de la lista de formas simples que se integran en el diccionario, de modo que las cadenas que integran las locuciones están construidas sobre las entradas simples del diccionario. El conjunto de diccionarios de lemas que recoge 86.104 formas simples y 25.721 formas compuestas.

A partir de los diccionarios de lemas se utiliza un conjunto de reglas de flexión que formalizan la flexión morfológica del español. En el sistema que se presenta esta flexión se divide en dos grandes clases:

1. La flexión verbal,
2. la flexión que se aplica a nombres, pronombres y adjetivos.

El resultado de esta flexión es un conjunto de entradas en las que cada unidad léxica, simple o compuesta, se asocia a su correspondiente lema e información léxica (Cf. Fig. 1). En el caso de las unidades léxicas ambigua cada entrada lleva asociada un secuencia de dos o más lemas y su información léxica.

Este conjunto de entradas que forman el diccionario electrónico se representa en forma de transductor subsecuencial (Cf. Figs. 2 y 3) en el cual cada camino de estados, desde el estado inicial hasta un estado final, se corresponde con una única entrada del diccionario. Este formato permite representar de forma eficiente el diccionario a la vez que acelera de forma muy significativa el proceso de análisis.

```
ama,ama.N:m:f:s,amar.VPRED:IPRES:3s:IIMPE:2s,amo.N:f:s
amaba,amar.VPRED:IPIMP:1s:3s
amabais,amar.VPRED:IPIMP:2p
amábamos,amar.VPRED:IPIMP:1p
amaban,amar.VPRED:IPIMP:3p
amabas,amar.VPRED:IPIMP:2s
amad,amar.VPRED:IIMPE:2p
amada,amado.APRED:f:s,amado.N:f:s,amar.VPRED:PP:f:s
amadas,amado.APRED:f:p,amado.N:f:p,amar.VPRED:PP:f:p
amado,amado.APRED:m:s,amado.N:m:s,amar.VPRED:PP:m:s
amados,amado.APRED:m:p,amado.N:m:p,amar.VPRED:PP:m:p
amáis,amar.VPRED:IPRES:2p

bomba/atómica,bomba/atómica.N:f:s
bombas/atómicas,bomba/atómica.N:f:p
```

Figura 1. Ejemplos de entradas del diccionario electrónico.

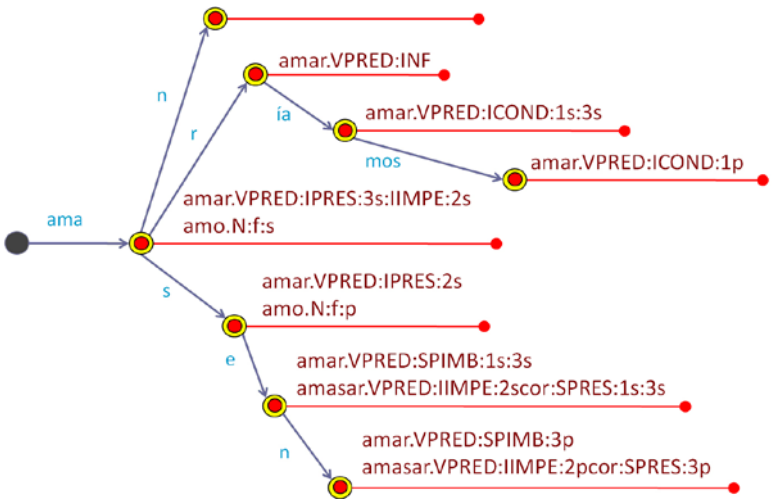


Figura 2. Representación del diccionario electrónico en forma de transductor subsecuencial en la que aparecen parte de las formas del verbo *amar*.

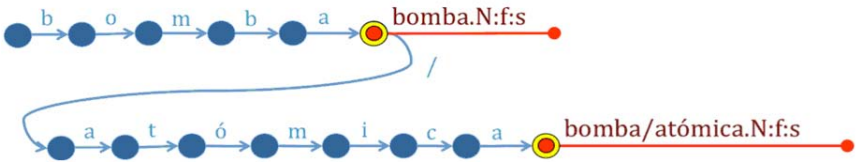


Figura 3. Representación del diccionario electrónico en forma de transductor subsecuencial en la que aparecen las formas de la locución *bomba atómica*.

2.2. Análisis léxico

El transductor subsecuencial que representa el diccionario electrónico se utiliza para transducir texto plano, sin formato ni ningún tipo de anotación. En este proceso el texto a analizar se utiliza como entrada del algoritmo de transducción de modo que cada secuencia de caracteres entre espacios en blanco es tratada como la entrada del transductor del diccionario. El resultado de la transducción de cada una de estas secuencias es una nueva secuencia formada por el lema y la información léxica correspondiente a la unidad léxica presente en el diccionario. Estas salidas no se representan como un texto plano sino que se utilizan como elementos del alfabeto de entrada del autómata finito que finalmente representará el texto analizado.

Como se puede apreciar en la figura 4 el autómata finito que representa el resultado del análisis léxico del texto *al habérselo propuesto a tiempo* contiene la anotación de los elementos simples y locutivos que se pueden analizar a partir del diccionario y permite representar y manipular posteriormente de forma eficiente los casos ambiguos como las formas del verbo *haber*.

Figura 4. Resultado del análisis de *al habérselo propuesto a tiempo.*

3. Transducción léxica

En el análisis de *al habérselo propuesto a tiempo* representado en la figura 4 aparece la forma verbal compuesta del verbo *proponer*, *haber propuesto*, representada como una secuencia cuatro de unidades léxicas en la que entre las formas *haber* y *proponer* que forman el tiempo verbal aparecen dos pronombres clíticos. El hecho de que este tipo de locuciones, a diferencia de lo que ocurre con la forma compuesta *a tiempo*, puedan contener entre las unidades que las forman elementos que requieren un análisis sintáctico (secuencias de pronombres clíticos, grupos nominales, etc.) provoca que no puedan detectarse a partir del diccionario y que por tanto se requiera un proceso de análisis posterior que se realiza sobre el autómata finito que representa el texto.

Dicho proceso se denomina *transducción léxica* y en él se aplican sobre el autómata textual un conjunto de gramáticas léxicas en forma de transductores subsecuenciales (Karttunen 1994) que permiten reconocer los elementos que forman parte de la unidad léxica locutiva, agruparlos en una única entrada dentro del autómata y separarlos de aquellos elementos que no pertenecen a la forma locutiva.

La transducción léxica permite detectar y analizar:

1. tiempos verbales compuestos, como es el caso de “habérselo propuesto”, y
2. locuciones verbales que no pueden analizarse en el paso previo debido a que entre sus elementos pueden aparecer segmentos oracionales que no pertenecen a la

locución y que requieren un análisis sintáctico. Este es caso de la locución verbal "dar por sentado". En ella podemos encontrar elementos no locutivos, entre sus elementos conexos (Bobes 2000), como por ejemplo en "dar siempre muchas cosas por sentado" , donde encontramos la secuencia "siempre muchas cosas" entre las partes conexas "dar" y "por sentado".

El resultado de esta transducción permite eliminar la ambigüedad de las formas transducidas, como es el caso de las formas del verbo "haber" en el ejemplo de la figura 5. Otra de las ventajas de este proceso es que se reduce la complejidad del procesamiento del autómata textual en los procesos posteriores agrupando en una única unidad léxica todos los elementos que forman parte del predicado locutivo y situando los elementos externos en su contexto derecho (Cf. figura 5). De este modo durante un análisis sintáctico posterior la forma verbal compuesta del ejemplo anterior se detecta como un único elemento.

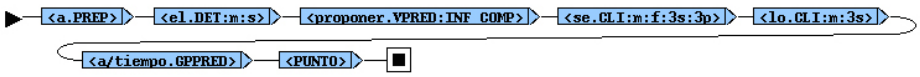


Figura 5. Resultado de la transducción del autómata de la figura 4 en la que se ha reconocido el tiempo verbal compuesto *haber propuesto.*

3.1. Definición de las gramáticas de transducción

Una gramática de transducción se define como una expresión regular o como un conjunto de expresiones regulares que especifican los elementos que forman la parte conexa de la unidad locutiva y los elementos no pertenecen a ella, y cómo se agrupa la parte conexa creando un nuevo lema y asociando a este la información léxica que se deriva del análisis de la unidad léxica.

En el ejemplo siguiente puede verse cómo se define la expresión regular que permite detectar y procesar el tiempo verbal compuesto haber + participio:

(1) (<haber.VAUX:INF\1> + <haber.VAUX:GER\1>) (<E> + <CLI\2> (<E> + <CLI\3>)) <VAR-1.VPRED:PP\4> [VAR-1,4-2,1-3&_COMP |2|3]

En este ejemplo se observa cómo se utilizan variables, \1 y VAR-1, por ejemplo, que permiten ampliar la definición del transductor y recolectar la información de las unidades que forman parte del texto transducido para, posteriormente, construir la salida correctamente si se detecta correctamente la construcción locutiva.

Las variables de tipo VAR-x se denominan **variables posicionales** y se cargan con el valor que se encuentre en esa posición dentro de secuencia de caracteres que define la unidad léxica procesada. De este modo en la salida ([VAR-1,4-2,1-3&_COMP |2|3]) producida por el transductor durante la detección de la *secuencia haber se lo propuesto* la variable VAR-1 se carga con el lema del participio del verbo proponer, VAR-1=proponer.

Las variables definidas como \1 se denominan **variables ligada a la transición** y son variables multicampo que se dividen en tantos campos como campos contiene la información de la unidad léxica que aparece en la posición transición definida por el valor de la variable.

Tabla 1. Ejemplo de carga de variables ligadas a la transición para la transducción de *haber se lo propuesto*.

Unidad léxica	<haber.VAUX:INF>			<se.CLI:m:f:3s:3p>					
Variable	\1			\2					
Elementos de la información léxica	haber	.VAUX	:INF	se	.CLI	:m	:f	:3s	:3p
Campos de la variable	1-1	1-2	1-3	2-1	2-2	2-3	2-4	2-5	2-6

Tabla 2. Ejemplo de carga de variables ligadas a la transición para la transducción de *haber se lo propuesto*.

Unidad léxica	<lo.CLI:m:3s>				<proponer.VPRED:PP:m:s>				
Variable	\3				\4				
Elementos de la información léxica	lo	.CLI	:m	:3s	proponer	.VPRED	:PP	:m	:s
Campos de la variable	3-1	3-2	3-3	3-4	4-1	4-2	4-3	4-4	4-5

En las tablas 1 y 2 puede verse como se definen estas variables y qué valores se cargan durante la transducción de *haber se lo propuesto*. La variable 4-2 se carga con la clase de palabra de la cuarta unidad de la secuencia (*4-1=VPRED* y la variable 1-3 con la información morfológica de la primera unidad, que en esta construcción puede ser *INF* (infinitivo) o *GER* (gerundio), de modo que en el ejemplo *1-3=INF*. Por último las variables 2 y 3 se cargan con el valor de los elementos opcionales que se pueden encontrar en las posiciones 2 y 3 de la secuencia.

El resultado producido por la salida del transductor a partir de los valores cargado en las variables es el que puede observarse en la figura 5. El transductor de la gramática del tiempo verbal haber + participio que se construye a partir de (1) es el que puede observarse en la figura 6.

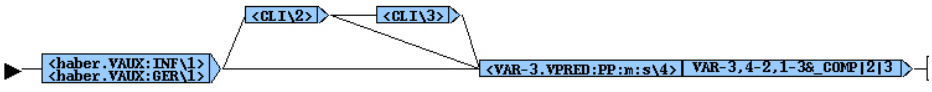


Figura 6. Transductor subsecuencial de la construcción haber + participio.

3.2. Transducción de locuciones verbales

Del mismo modo que en el apartado anterior se ha mostrado el proceso de detección y transducción de un tiempo verbal compuesto se realiza la transducción de locuciones verbales como *correr un riesgo*. De este modo para el texto:

(2) corrió en todo momento un enorme riesgo

su análisis léxico genera el autómata de la figura 7 en el que puede observarse como la locución en todo momento sí se detecta a partir del diccionario, pero en cambio no ocurre lo mismo con la locución correr un riesgo. Obsérvese como esta última contiene entre sus partes conexas elementos que no forman parte de la locución: ... *en todo momento [un] enorme* ...

Figura 7. Autómata resultante de la anotación léxica de *corrió en todo momento un enorme riesgo.*

La anotación de esta locución solo es posible mediante la transducción del autómata textual con el transductor de la figura 8 que formaliza la gramática de detección de la unidad locutiva. Esta gramática se define de forma similar a la gramática del tiempo verbal compuesto vista anteriormente.

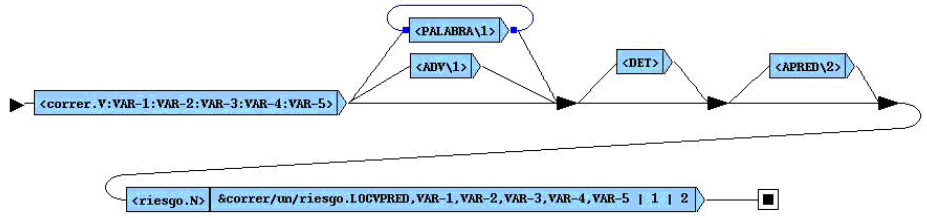


Figura 8. Transductor subsecuencial de la locución verbal correr un riesgo.

El proceso de transducción léxica del autómata de la figura 7 produce como resultado un nuevo autómata textual en el que la unidad locutiva correr un riesgo se anota como una única unidad y los elementos que no forman parte de ella se sitúan en su contexto derecho (Cf. Fig. 9).

Figura 9. Resultado de la transducción del autómata de la figura 7.

4. Referencias

Bobes, E. (2000). *Gramática electrónica de las locuciones verbales*. Laboratorio de Lingüística Informática, Universidad Autónoma de Barcelona.

Collins, M. (1999). *Head-driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania.

Laporte, E. (2001). Reduction of lexical ambiguity. *Lingvisticae Investigationes XXIV:1*, Amsterdam-Philadelphie : Benjamins, pp. 67-103.

Karttunen, L. 1994. Constructing Lexical Transducers, en *Proceedings of the Fifteenth International Conference on Computational Linguistics. Coling 94,* vol. I, pág. 406-411, Kyoto, Japan.

Mohri, M. (1997). Finite-state transducers in language and speech processing. *Computational Linguistics*, *vol. 23(2)*, 269-311.

Subirats, C. y Ortega, M. (2000). Tratamiento automático de la información textual en español mediante bases de información lingüística y transductores. *Estudios de Lingüística Española 10*. Disponible en http://elies.rediris.es/elies10/

Corpus, adquisición y enseñanza de lenguas

The GENTT Corpus: Integrating Genre and Corpus in the Teaching of Language for Specific Purposes

Anabel Borja Albi
Natividad Juste Vidal
Pilar Ordóñez López
Universitat Jaume I (Castellón de la Plana)
GENTT Research Group (Géneros Textuales para la Traducción/Textual Genres for Translation)

Abstract

Since the 1980s, the application of the concept of genre to the teaching of Language for Specific Purposes has become one of the most dynamic research lines in genre studies. The GENTT corpus is a multilingual corpus of specialised genres that, based upon the concept of genre, provides formal, communicative and cognitive information about the genres it contains.

In this paper we highlight how genre-based approaches, and specifically the application of the GENTT corpus, as part of the teaching of Legal and Business English can contribute to overcoming the criticism levelled at the use of bottom-up methodologies in corpus linguistics and of decontextualised corpus data. To illustrate this, we present a set of activities for the Legal and Business English classroom, based on the use of the GENTT corpus.

Keywords: Corpora, specialised genres, GENTT corpus, LSP teaching

Resumen

La aplicación del concepto de género a la enseñanza de lenguas para fines específicos se ha convertido desde la década de los ochenta en una de las líneas de investigación más dinámicas en el estudio sobre géneros. El corpus GENTT es un corpus multilingüe de géneros especializados que, basado en el concepto de género, proporciona información formal, comunicativa y cognitiva de los géneros que contiene.

En este trabajo pretendemos poner de manifiesto cómo la incorporación del corpus GENTT así como del enfoque basado en el género a la enseñanza del inglés jurídico-económico, nos ayuda a superar las críticas dirigidas hacia el uso de metodologías bottom-up en la lingüística de corpus y hacia el uso descontextualizado de los datos contenidos en el corpus. Para ello, presentamos una serie de actividades prácticas, basadas en el uso del corpus GENTT, en el aula de inglés jurídico-económico.

Palabras clave: Corpus, géneros especializados, corpus GENTT, enseñanza de LSP

1. Introduction

The main focus of the GENTT (*Géneros Textuales para la Traducción*/Textual Genres for Translation) Research Group is the multilingual study of genres in professional legal, medical and technical contexts, three domains that occupy a central position in LSP Teaching and Translation Studies. The GENTT project has focused on mapping the textual performances of these fields and compiling a multilingual (Catalan, English, German, Spanish and French) example corpus of specialised discourse texts in the fields of law, medicine and technology. Currently the corpus contains approximately 1,000 full-text professional documents corresponding to the more representative genres of these three fields. The representativeness of this corpus lies in the number of different genre examples it provides, not in the number of texts or words it contains. In fact, the corpus design is intended to create a knowledge management system (a genre tree), similar to terminological knowledge representation systems, structured around the notion of genre, for learners and users of professional specialised genres (Borja, 2005). The GENTT corpus embraces the web 2.0 philosophy, providing a collaborative environment that permits those interested in Specific Language Domains to search, feed and manage the Corpus online. More information about the research group and the corpus can be found at www.gentt.uji.es.

2. Corpora and genre in LSP teaching

Since the 1980s, the application of genres to language teaching and, in particular, to LSP teaching, has become one of the most dynamic lines of research in the area of genre theory. Regarding the process of second language acquisition, Bazerman (1988, 2002), Bhatia (1993, 2004) and Swales (1990, 2004), among others, emphasize the importance of understanding communication codes that are specific to the culture of individual fields of specialisation and the structure of its genres in order to acquire linguistic expertise in a particular field of knowledge.

The application of genres to the teaching-learning of LSP provides both teachers and learners with culture-dependent codes in their progress towards the acquisition of language for special purposes. Each specialised genre presents its own features, purposes and cultural conventions, facilitating the learning process to the extent that once learners know how to recognise and use genres (terminology, phraseology, macrostructure, rhetorical devices, etc.); they are able to develop the necessary strategies to cope with new and unfamiliar text types. This is crucial in the context of current Information Society Technologies since genres, as well as reality, are constantly changing, evolving and appearing in new forms. They reflect the evolution of specific socio-cultural interactions and, therefore, of specialised linguistic performance and expertise.

3. The gentt corpus as a lsp teaching tool

Until recently, language-teaching materials were based mostly on made-up sentence examples. However, the impact of electronic text analysis in the area of Language for

Special Domains is rapidly evolving, especially in qualitative terms. Tribble (1997) shows how even small corpora of fewer than one million words can be of considerable benefit in the LSP context. The applications range from high frequency lexis in a specific domain, collocation, colligation and semantic prosody, grammar and discourse, to the contrastive analysis of lexical items in different domains or contrastive analysis of genres (text-types) in different sublanguages.

Adolphs (2006) points out that a key advantage of the use of corpora within the context of LSP is that they can supply data regarding specific phraseology, word frequencies and distributions in different discourse contexts, thereby providing important information for the language learner and instructor. The co-text reveals information about the specific phraseology that surrounds a particular word and contributes to its functional interpretation. Word frequency information can be used to design syllabuses based on the needs of particular learners with regard to both the sequence of specific vocabulary items that are being taught and the overall size of the restricted vocabulary that is required to achieve an adequate coverage of a specialised domain. This kind of information derived from corpora provides LSP learners with a list of words that make up the core of the language domain, and that can be used to analyse and deal with numerous vocabulary acquisition problems such as polysemous words, "learnability" or interference with the learner's first language and decontextualisation.

The structure of the GENTT corpus incorporates "genre templates" which provide formal, communicative and cognitive information about the genres it contains, e.g. macro- and micro-linguistic features, function, rhetorical devices. Previous research in the area has shown that when in possession of this information, LSP learners can progressively improve their professional competence, both linguistic and extralinguistic, through a self-directed learning process. In this paper, we show how the incorporation of both corpus-based and genre-based approaches into text analysis as part of LSP teaching can, in some respects, overcome the criticism that corpus linguistic analyses apply bottom-up rather than top-down methodologies, and that the use of decontextualised corpus data does not take into account the socio-cultural context.

The GENTT corpus also provides learners of professional languages with text models and patterns to be used as cultural, textual, conceptual, linguistic and terminological reference. The possibilities of applying the work on genre-based corpora to LSP teaching are evident. According to Bhatia (1997), work with genres pertaining to the student's professional background and interests causes learners to develop an explicit desire for conscious participation in the professional community and a feeling of 'shared ownership' of their communicative resources, rather than learning words and structures mechanically and out of context. Bhatia (2008) believes that learners of a specialised language need: (1) to understand the specialist's communication code; (2) to familiarise themselves with rhetorical resources and those that occur in specialised genres; (3) to understand the various socio-cultural contexts in which specialised communication takes place; and (4) to be capable of using specialised genres to respond to new and unexpected situations. All these four skills can be enhanced by using the GENTT corpus and becoming familiar with the textual *mapping* and genre characterisations it provides. Another advantage of this electronic tool, based on genres extracted from real life

communication, is that language is learnt in its true context and learning programmes can be designed with very specific needs in mind. It is possible, for example, to design a course focused on the discourse of a particular professional and communicative situation (e.g. the documents of a judicial process which might be of interest for a student of Legal English), treating it as a single genre (a judgment) or as a system of genres that would include all the documents that accompany that particular genre (claim, counterclaim, injunction, judgment, appeal, etc.), and even the oral genres related to them (witness deposition transcripts).

The GENTT corpus permits users interested in restricted domains (legal, medical and technical) to search, feed and manage a collection of texts on line. Participants in the teaching-learning process can manage their own subcorpus in different languages and in different specialised domains, depending on their aims and needs. As stated by Adolphs (2006), in order to build a suitable corpus or subcorpus for specific needs in the LSP context, it is essential to establish the basis for its design criteria, that is, what the *role of the corpus* is; *who* uses it; what the particular *learning objective* is; or what the particular *genre* that is being explored is.

Bearing these factors in mind, a constructive teaching-learning approach based on the concepts of 'corpus' and 'genre' can be implemented, applying a methodology in which learners actively use language in a given context, monitor their own learning progress and develop new skills and competences such as hypothesis testing and data analysis. Teachers act as facilitators of the student's learning process and they can design their own syllabus ranging from very controlled to more complex learning tasks. This approach enhances data-driven learning, allowing learners to explore language data and to derive patterns of language use, which promotes creativity and innovation in the language classroom.

Within this context learners become the centre of the process as they can improve their linguistic and extra-linguistic competences according to their own learning styles. Teachers, on the other hand, are facilitators of the instruction process and can also tailor their syllabus design towards the needs of a diverse range of learners. Both sides can benefit from the use of corpus technologies by means of a hypothetical deductive approach, helping LSP students learn to communicate effectively and fully understand the realities of the world of specialised discourse.

The following section presents three practical learning activities for Business English using the GENTT corpus, in which the role of the teacher is that of a facilitator instructing students in analytic strategies, both rhetorical and textual.

4. The GENTT corpus in the business English classroom

In this section we will present different corpus-based activities for the Business English classroom. As we have explained, the GENTT corpus is built upon the concept of genre, which constitutes 'an important source of insight' (Swales, 1990: 54) into the specialised area. From a methodological point of view, the use of the GENTT corpus in the classroom makes it possible to adopt a data-driven approach. This approach contributes significantly to the development of autonomous learning, enabling students to identify, in an

autonomous manner, the characteristic patterns of each macrogenre, genre and subgenre, thereby increasing their capacity of performing critical analysis and decision making; browsing the corpus helps students to develop a 'researcher-led' approach.

4.1. Business English: a trendy cover term?

Business English has experienced a steady growth in recent decades, becoming the main expansion area of ESP (Hewings, 2002), due to various factors such as the consolidation of English as a *lingua franca* and the expansion of local and national markets, leading to an increase of international business relationships.

Scholars coincide in the difficulty of establishing a clear definition of 'business English' (Pickett, 1986; Dudley-Evans & St John, 1998). On the one hand, business English is taught/learnt for a wide variety of purposes; on the other hand, there exists considerable overlap between business English (ESP) and general English in strictly linguistic terms. Nevertheless, it is important to establish some defining features of business English in order to maximize the pedagogical outcome of the proposed classroom activities. Given the educational setting in which we are working, the most relevant characteristics are the following:

- Effective communication is the main concern (Dudley-Evans & St John, 1998: 73).
- Language depends on status, power and how well established the business relationship is (ibid. 73).
- Seven core communicative events are identified: telephoning, socializing, making presentations, taking part in meetings and negotiating (oral); and corresponding and reporting (written) (ibid.: 63).
- The communicative events above, especially those that require the written form, are carried out by means of specific genres.
- As well as sharing other pragmatic features with general English, assertion and downtoning, as well as checking and confirming (Duckworth, 1995) are key elements of business English.
- Combination of general, semi-specialized and specialized lexis.

4.2. Corpus-based activities in the teaching of business English

4.2.1. Course description:

The following activities are designed for the subject 'Business and Administrative English for Translators', which is an optional course that undergraduates in their second year can choose to take. This course is aimed at providing students with some basic knowledge of business communication (correspondence, types of associations, reports, CVs and covering letters). For the vast majority of students, this is the first contact with business English. Furthermore, it is important to bear in mind that most of them are not familiar with corpus methodologies. This implies that some initial familiarisation sessions are needed, in which

students get acquainted with the terminology of corpus linguistics and with the GENTT corpus itself. The students' average language proficiency is at an upper intermediate level.

4.2.2. Activity 1: Familiarizing themselves with the lexis of business correspondence

Task: Students are asked to use the GENTT corpus to browse the documents available in the genre 'Letters', and to identify the most typical lexical elements of this genre.

Aims: To conduct a simple search by genre in the GENTT corpus; to extract the most characteristic lexical elements of the genre.

Methodology: Students work in pairs. A list of instructions is provided, explaining the steps that should be followed to carry out this activity. Once students are logged on to the corpus, they should conduct a search to retrieve the information they are required to find. After retrieving the corresponding documents, it is their task to identify the relevant lexical elements, working in pairs. To conclude this activity, each pair is asked to present their results to the other students in the group.

Teaching rationale: The rationale behind this activity is to provide students with a new approach to the analysis of business language and, at the same time, to encourage them to take an active role in the study of the semi-specialised and specialised lexis of business correspondence.

4.2.3. Activity 2: Identifying collocations typical of business correspondence

Task: Students are asked to find the most frequent collocations in which the word 'payment' occurs.

Aims: To introduce concordances; to make students reflect on the relevance of phraseology; to make students aware of genre conventions.

Methodology: Students work in small groups. First, students log on to the corpus. They are provided with instructions for every step. Students need to conduct a keyword search, explore the concordances they retrieve, and analyse these results according to the task. Following this, students are expected to take part in a discussion to share their findings with the rest of the classroom.

Teaching rationale: The motivation for this activity is to raise students' awareness of the fundamental role of collocations, phraseology and genre conventions in specialised communication. Ultimately, the rationale behind this activity is to challenge students' restricted perception of specialised language, which is frequently equated with specialised lexis.

4.2.4. Activity 3: Learning to identify expressions of obligation, permission and prohibition in business agreements

Task: Students are asked to identify all the Business Agreements in the corpus and find clauses expressing three particular categories of language: obligation, permission and prohibition.

Aims: To make students aware of the importance of using the right verbs and categories of language to convey the exact meaning and purpose of the agreement.

Methodology: The students are provided with tables containing one or more examples of a particular category of language, with each example being followed by variations on that example (e.g.: obligation clauses using *shall*, *must*, *agrees to*, *undertakes*, *covenants*, *is obligated to*, etc.). The table is introduced by the teacher and discussed in the classroom to highlight the typical dysfunctions in the use of these expressions and the problems a deficient use may bring about. Then students work individually to identify and download the Business Agreements contained in the corpus and create a subcorpus. Working in small groups, they are asked to find examples in the subcorpus similar to the ones provided. Each group works on one category and presents its findings to the class.

Teaching rationale: Commercial litigation frequently has its roots in mishandled contract language and this activity is aimed at raising students' awareness of the importance of seeking consistency in written usages with very specific purposes.

Additional activities, once students have familiarised themselves with the corpus, are the identification of downtoning structures, the mapping of the business genre by means of identifying different subgenres, and the analysis of more complex phraseological units.

5. Conclusion

The GENTT corpus provides a user-centred interface where learners and teachers alike can proceed to a very specific type of text exploitation. Personalised corpora and subcorpora can be designed in order to select adequate texts for a particular teaching issue and their corresponding analysis. Advanced search criteria or classification may also be implemented according to user's needs (by genre, working language, etc.), thus fostering creativity and collaboration among them. With the implementation of these types of tools we are moving towards new expectations in language teaching, that is, towards a dynamic corpus-based and genre-based approach in which learner and teacher collaborate, participate and interact in the process of acquisition of Language for Special Purposes.

Bibliography

Adolphs, S. (2006). *Introducing Electronic Test Analysis. A practical guide for language and literary studies.* New York: Routledge.

Bazerman, C. (1988). *Shaping written knowledge*. Madison, WI: University of Wisconsin Press.

Bazerman, C. & Russell, D. (2002). *Writing Selves/Writing Societies: Research from Activity Perspectives*. Perspectives on Writing. Fort Collins, Colorado: The WAC Clearinghouse.

BHATIA, V.K. (1993). *Analyzing Genre: Language Use in Professional Settings*, London: Longman.

BHATIA, VIJAY V.K. (1997). Translating Legal Genres. In Anna Trosborg (ed.) *Text Typology and Translation* Amsterdam: Benjamin, 10-15.

BHATIA, V.K. (2004). *Worlds of Written Discourse: a Genre-based View.* London and New York: Continuum.

BHATIA, V. K.; CANDLIN, C. N. & J. ENGBERG, (EDS.) (2008). *Legal Discourse across Cultures and Systems*. Hong Kong: Hong Kong University Press.

BORJA, A. (2005). Organización del conocimiento para la traducción jurídica a través de sistemas expertos basados en el concepto de género textual. In I. García Izquierdo (ed.). *El género textual y la traducción. Reflexiones teóricas y aplicaciones pedagógicas.* Bern: Peter Lang, 40-42.

DUCKWORTH, M. (1995). *Oxford Business English. Oxford: Oxford University Press.*

DUDLEY-EVANS, T. & M. J. ST JOHN (1998). *Development in English for Specific Purposes: a Multi-disciplinary Approach. Cambridge: Cambridge University Press.*

HEWINGS, M. (2002). A History of ESP through English for Specific Purposes in English for Specific Purposes World, 1(3). Disponible en http://www.esp-world.info/Articles_3/Hewings_paper.htm

PICKETT, D. (1986). «Business English. Falling between two stools». *Comlon* 26, 16-21.

SWALES, J. (1990). *Genre Anaylsis. English in Academic and Research Settings. Cambridge/New York: Cambridge University Press.*

SWALES, J. (2004). *Research Genres. Explorations and applications.* Cambridge: Cambridge University Press.

TRIBBLE, C. (1997). Improving corpora for ELT: quick-and-dirty ways of developing corpora for language teaching. In J. Melia and B.

LEWANDOWSKA-TOMASZCZYK (EDS) *PALC '97 Proceedings*. Lodz: Lodz University Press. Disponible en http://www.ctribble.co.uk/text/Palc.htm

Plataforma *GARALEX*: infraestructura tecnológica para la investigación y la didáctica de lenguaje del ámbito de las ciencias jurídicas

Joseba Ezeiza Ramos
Universidad del País Vasco/Euskal Herriko Unibertsitatea. Departamento de Filología Vasca

Resumen

En este artículo se presenta una plataforma Web, para la investigación y la didáctica del lenguaje jurídico basada en metodologías de análisis de corpus. Desde esta plataforma se ofrecerá a estudiantes, profesores y profesionales del ámbito jurídico y administrativo tres tipos de recursos: a) recursos de comunicación; b) recursos de consulta; y c) recursos de formación. Se trata de un proyecto de enfoque jurilingüístico que tiene como finalidad contribuir a dinamizar y armonizar el desarrollo y el uso de la lengua vasca entre los especialistas del área jurídica en el entorno universitario: profesores, estudiantes, investigadores, etc.

Palabras clave: jurilingüística, lenguaje jurídico-administrativo, instrumentos de corpus, TIC

Abstract

This article presents a Web platform for research and the teaching of legal language based on methodologies of corpus analysis. Three types of resources will be offered from this platform to students, teachers and professionals in the legal and administrative fields: a) communication resources; b) consultation resources; and c) training resources. It is a project of jurilinguistic approach which aims to contribute to dynamize and harmonize the development and use of the Basque language among the specialists in the legal area in the University environment: teachers, students, researchers, etc.

Keywords: jurilinguistics, administrative language, corpus instruments, ICT

1. Objetivos del proyecto garalex

El proyecto GARALEX[162] tiene como objetivo desarrollar una infraestructura específicamente diseñada para la investigación y la formación en el ámbito del lenguaje jurídico-administrativo. Se trataría de una plataforma de carácter social, abierta y dinámica, que facilitaría la participación activa –con diferentes funciones y atribuciones en función del proyecto- de profesionales, profesores y estudiantes del ámbito de las Ciencias Jurídicas, y que trataría de promover el trabajo colaborativo entre lingüistas y expertos en el área. También pretende crear un hilo conductor entre el mundo académico y el mundo profesional. En este sentido, GARALEX se proyecta como una *base logística* que pueda servir de vía de conexión entre agentes universitarios, profesionales del ámbito jurídico y administrativo, y lingüistas que trabajan fuera de la universidad en traducción, la terminología jurídica, la estandarización documental, etc.

En última instancia, con todo ello se pretendería contribuir desde la universidad a impulsar y reforzar el desarrollo léxico-discursivo del euskera el ámbito jurídico y administrativo, aprovechando de forma estratégica las posibilidades de intervención que ofrece la institución universitaria. En la figura nº 1 se ha tratado de reflejar la relación sinérgica que el proyecto trata de establecer entre los fines propuestos en cada uno de los niveles en los que pretende realizar algún tipo de aportación: el nivel estructural (plataforma de recursos), el nivel participativo (redes de colaboración) y el nivel sociolingüístico (desarrollo léxico-discursivo).

Figura 1: Objetivos del proyecto GARALEX

162 plataforma GARALEX se ha ido gestando al abrigo de varios proyectos de I+D+I: PREST (DGE06/04), GARATERM (EJIE07/07), DB (OTRI: 2007.0077), EHLB (OTRI: 2008.0368) e HIZLAN (DIPE08/16)

Esta idea comenzó a gestarse el año 2006 coincidiendo la reforma de los planes de estudio universitarios. De hecho, comenzó a ver la luz el curso académico 2006/2007 en el marco del proyecto PREST (DGE06/04). En este proyecto se analizaron, entre otras, las necesidades comunicativas y lingüísticas de los profesionales del Derecho y ello permitió detectar las competencias, habilidades y conocimientos de tipo comunicativo-lingüístico que dichos profesionales deberían desarrollar y/o adquirir durante su estancia en la universidad. También se evidenció una carencia de infraestructuras y recursos para el desarrollo de estas competencias y conocimientos en euskera.

Así pues, el año 2007 se puso en marcha el proyecto GARATERM (EJIE07/07) con el objetivo de analizar con qué recursos de tipo tecno-lingüístico convendría contar para facilitar estos aprendizajes en diversas áreas de conocimiento, tratando de optimizar la convergencia de esfuerzos de los diferentes equipos docentes que intervienen en los planes de estudio, con el fin de garantizar que los estudiantes alcancen un nivel adecuado de competencia comunicativa disciplinar y, en el caso que nos ocupa, un conocimiento suficiente de los rudimentos lingüísticos propios de las Ciencias Jurídicas. Para ir cubriendo algunas de las lagunas detectadas, se estableció un convenio de colaboración entre el Departamento de Filología Vasca de la UPV/EHU[163] y el grupo Ametzagaiña A.I.E.[164], unidad de I+D empresarial de la Red Vasca de Tecnología, que en el periodo 2007-2008 se materializó en dos contratos universidad & empresa (OTRI 2007.0077 y OTRI 2008.0368) que permitieron crear el núcleo central de lo que posteriormente se ha ido materializando como plataforma GARALEX.

No obstante, el avance definitivo se produjo en el marco del proyecto I+D+I HIZLAN (DIPE08/16). Este proyecto se ha llevad a cabo a lo largo de los años 2009 y 2010 y ha contado con financiación del Departamento de Innovación de la Diputación Foral de Bizkaia. Su objetivo principal consistió en desarrollar una plataforma para la gestión del conocimiento lingüístico, el control de la calidad de la comunicación, y la creación de recursos formativos, adaptada -en uno de sus módulos- a las particularidades de las producciones textuales de la administración, la justicia y el derecho.

Dicha plataforma tiene ciertas concomitancias con otras, como las desarrolladas por la Escuela de Lingüística de Valparaíso para el ámbito de la comunicación académica[165]; las de grupos como GENTT o GITRAD,[166] desarrolladas para la traducción jurídica y su didáctica; la de la Universidad Pompeu Fabra para orientar la redacción de textos jurídicos[167]; la del IULA dedicada a la Lingüística Forense;[168] el Servei Lingüístic de l'Àmbit Judicial de Catalunya;[169] o las diversas webs de la Red Canadiense de Centros de Jurilingüística,[170] el Portal Lingüístico de Canadá,[171] la plataforma LEGISTICS,[172] etc.

163 URL: http://www.ef.ehu.es/s0113-home1/es
164 URL: http://www.ametza.com/castellano/index.htm
165 http://www.linguistica.cl/
166 http://www.gentt.uji.es/?q=es y http://www.gitrad.uji.es
167 http://parles.upf.edu/cr/casjur2005b/index.html
168 http://www.iula.upf.edu/recurs08es.htm
169 http://tinyurl.com/6ekkj38
170 Instituto José-Dubuc, el Centro de Estudios Legales, Traducción y Documentación de la Universidad de Ottawa o el Centro de Traducción y de Terminología Jurídicas de la Universidad de Moncton.
171 http://www.noslangues-ourlanguages.gc.ca/decouvrir-discover/outils-tools/oar-wt-fra.html#c13
172 http://canada.justice.gc.ca/fra/min-dept/pub/legis/index.html

En cualquier caso, GARALEX es un proyecto aún emergente que, de momento, será necesariamente bastante menos ambicioso que los que se acaban de enumerar. No obstante, nace con la decidida intención de servir como punto de acceso privilegiado al conocimiento existente en torno al euskera jurídico-administrativo y pretende también hacer alguna contribución a la generación de nuevo conocimiento a partir de la investigación del contexto sociolingüístico que enmarca el uso del euskera en la administración y la justicia, el estudio de los modos y formas de comunicación de la comunidad profesional bilingüe y el análisis de los rasgos lingüísticos más representativos de dichos usos.

2. Focos de interés del proyecto garalex

El proyecto GARALEX se adscribe al marco epistemológico que ofrece la Jurilingüística (Bernard, 2009; Cacciaguidi-Wagner, 2008; Ciuro, 2009; Conru, 2005; Fahy, 2009; Ferran, 2005; Gemar, 1982; Gemar & Kasirer, 2005; Isani & Lavault-Olleon, 2009; Matilla, 2008; etc.). Se trata de un campo disciplinar situada en la intersección de las Ciencias Jurídicas y la Lingüística Aplicada, que reivindica la necesaria colaboración entre operadores jurídicos y lingüistas para el desarrollo de leyes, programas, procedimientos y recursos que garanticen la convivencia ajustada a derecho de las lenguas en el ámbito administrativo y jurídico.

En definitiva, la perspectiva jurilingüística pretende favorecer el desarrollo de marcos conceptuales, procedimientos metodológicos e instrumental técnico adecuado para el estudio, desarrollo y promoción del lenguaje administrativo y jurídico en perspectiva multilingüe. Para ello, la jurilingüística se alimenta, entre otras disciplinas, de la Sociolingüística, la Traductología Jurídica, Jurada y Judicial, la Normografía, la Terminología, Jurídica, el Derecho del Lenguaje, etc. A todas luces, este marco epistemológico aparenta muy adecuado para los para los fines trazados en el proyecto GARALEX.

Los focos de interés de la Jurilingüística serían básicamente cuatro. En el eje vertical se situarían los temas relativos a la legislación lingüística y la planificación que se realiza para desarrollar, armonizar y normalizar los usos de las lenguas presentes en el ámbito jurídico y administrativo. Este eje se cruzaría con otro en el que situarían los temas propiamente relacionados con el discurso y la terminología administrativo-jurídica y la calidad legislativa y documental. De la intersección de estos ejes emergen áreas de gran interés teórico-práctico para los profesionales del Derecho que desarrollan su tarea en contextos bilingües o multilingües, como son la política y criterios de traducción, la técnica legislativa y normográfica, la estandarización documental y la normalización terminológica. Los postulados de la Jurilingüística proponen analizar estas cuestiones tratando de conciliar la perspectiva jurídica (marco legal, procedimientos legislativos, etc.) y la perspectiva lingüística (política lingüística, criterios y recursos de estandarización, etc.). En la figura nº 2 se ha tratado de reflejar el "mapa" que, eventualmente, desplegaría la Jurilingüística, de acuerdo con la interpretación que se hace en el proyecto GARALEX.

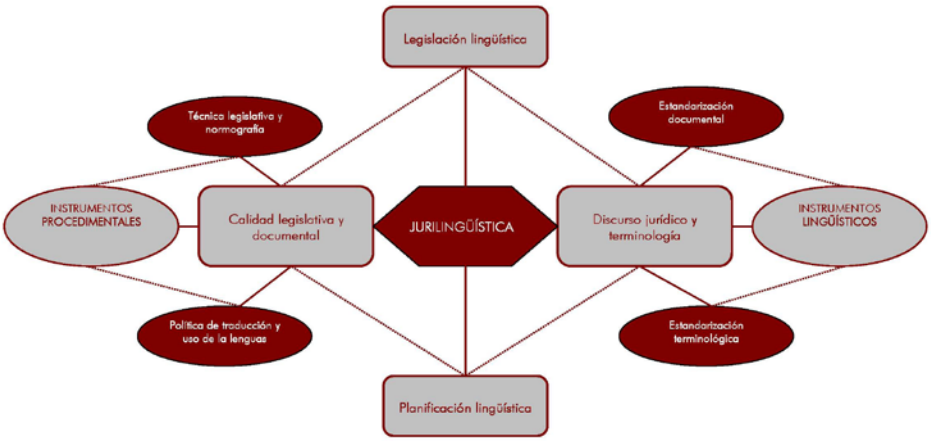


Figura 2. Ámbito de intervención del proyecto GARALEX

En la Facultad de Derecho de la Universidad del País Vasco se dan unas condiciones muy favorables para desarrollar un proyecto en este campo. Por una parte, en algún momento de su trayectoria docente una buena parte del profesorado bilingüe ha participado directamente en alguna iniciativa relacionada con el desarrollo y la normalización del euskera jurídico y administrativo bien en la universidad o bien fuera de ella. De hecho, la Facultad de Derecho constituye un polo de referencia en este ámbito desde los primeros años del proceso de normalización del euskera, a principios de la década de 1980.

Por otra parte, la Facultad de Derecho cuenta con un organismo –el Seminario de Euskera Jurídico- que promueve y coordina el proyecto Legeak/Leyes para la elaboración de versiones bilingües de leyes estatales y que colabora de forma habitual en los proyectos de normalización terminológica en este campo. Además, en la Facultad de Derecho hay una plaza docente permanente de Euskera Jurídico encargada específicamente de la docencia e investigación de esta materia en la universidad.

En este contexto, el proyecto GARALEX pretende ofrecer una plataforma que permita integrar las iniciativas orientadas a la investigación, la didáctica y la normalización del euskera de la Administración y el Derecho, que nacen y se desarrollan en el ámbito universitario y, dentro de sus posibilidades, pretende también actuar de motor dinamizador de futuras iniciativas fuera de la Universidad. Para todo ello, el proyecto desplegaría tres líneas de trabajo: una de vertiente social, otra de carácter didáctico y otra orientada a la investigación.

Así pues, por una parte, el proyecto GARALEX se propone contribuir a crear redes de colaboración entre expertos, estudiantes y lingüistas con el objetivo de facilitar la armonización de los usos lingüísticos y la terminología en el ámbito jurídico y administrativo. Por otra parte, se traza como objetivo reunir y generar recursos formativos adecuados para garantizar que los graduados universitarios y los profesionales

de este campo adquieran un conocimiento sólido del euskera administrativo y jurídico. Finalmente, con el fin de dotar de soporte empírico a las iniciativas y propuestas que se desarrollen para la formación y la dinamización de las comunidades de usuarios, se considera necesario promover líneas sistemáticas de investigación en torno a los usos del lenguaje en el ámbito jurídico-administrativo.

Dentro de este ámbito, el proyecto GARALEX centrará prioritariamente su atención en el discurso administrativo, el discurso jurídico y los lenguajes propios de dichos ámbitos. Pero, tal y como se refleja en la figura nº 3, también se interesará por el discurso empresarial y político-legislativo. Este interés lo justifica el proyecto de implantación de la doble titulación Derecho-ADE (Administración y Dirección de Empresas) en la UPV/EHU, y la afinidad de la oferta que realiza el Departamento de Filología Vasca en las titulaciones de Derecho, Ciencias Empresariales, Relaciones Laborales y Ciencias Políticas. En un futuro próximo está previsto integrar en este esquema todo lo relacionado con la Lingüística Forense, para dar respuesta a las demandas de la titulación de Criminología cuya primera promoción ha comenzado su andadura el curso académico 2010/11.

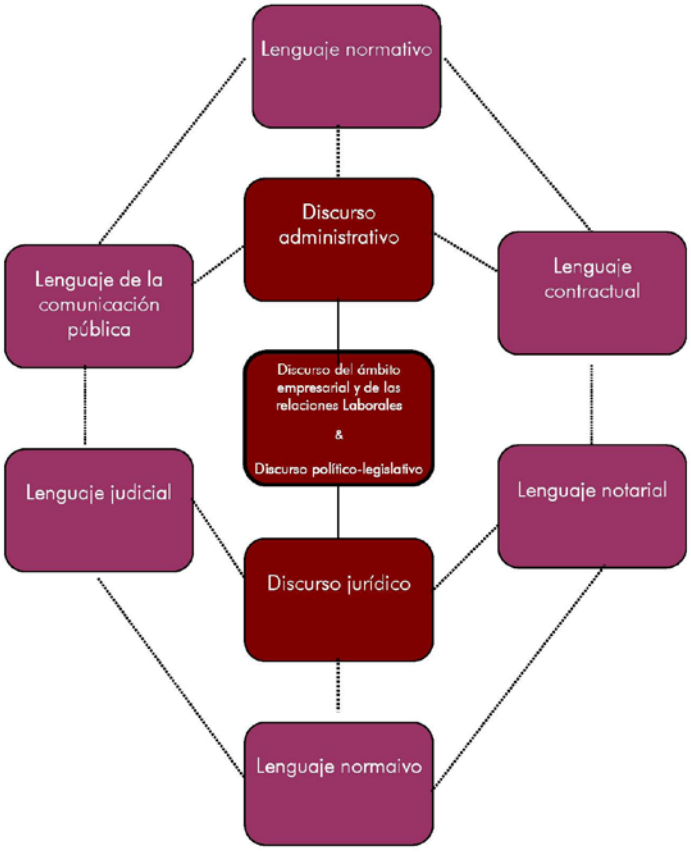


Figura 3. Áreas prioritarias de estudio del proyecto GARALEX

Los aspectos lingüísticos a los que se prestaría especial atención son muy diversos. Tal y como se puede observar en la figura nº 4 la agenda de trabajo prevista es bastante ambiciosa. No obstante, la agenda de trabajo aún no está claramente definida (por ello la esbozamos deliberadamente con un claro aire de provisionalidad). De hecho, en este momento estamos tratando de concretarla de manera más precisa. Pero, fundamentalmente, se trataría de estudiar a partir de las producciones de diversos agentes (estudiantes universitarios, profesores, profesionales, legisladores, políticos, etc.), qué correlaciones existen entre aquellos rasgos de carácter individual o social que pudieran estar incidiendo en los usos lingüísticos y los rasgos pragmático-funcionales y formales de dichos usos. De este modo, se espera poder realizar un diagnóstico del estado actual del euskera en el campo jurídico-administrativo, y poder, así, proponer intervenciones o recursos que permitan revitalizar o, en su caso, reconducir el proceso de desarrollo social y funcional de la lengua en este ámbito.

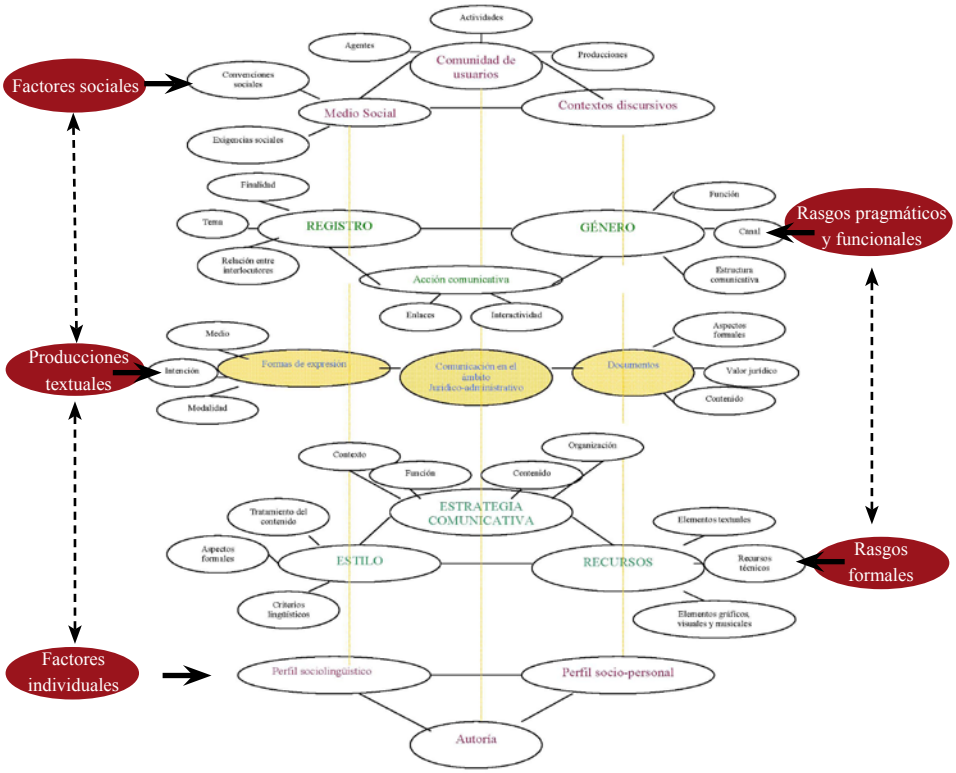


Figura 4. Focos de interés del proyecto GARALEX

3. Estructura de la plataforma GARALEX

Para facilitar esta tarea, el proyecto GARALEX articulará su actividad en torno a un portal web[173] que ofrecerá diversos tipos de recursos. Como se puede observar en la figura nº 5, dicho portal contaría con tres secciones: a) una primera sección de recursos de información y comunicación (blogs, revistas digitales, foros, wikis...), pensados para contribuir a articular, gestionar y dinamizar la red social de estudiantes, profesores, expertos y lingüistas interesados en colaborar con el proyecto; b) una segunda sección que ofrecerá recursos de formación y consulta sobre cuestiones relacionadas con la comunicación y usos del lenguaje en el ámbito administrativo y jurídico (glosarios, fichas y recomendaciones de estilo, modelos textuales, cursos telemáticos, etc.); y c) un tercer apartado desde el que se podrá acceder a recursos para el análisis y la evaluación de la calidad de los textos doctrinales, legales, profesionales y académicos (manuales de redacción legislativa, redacción académica, redacción de documentos profesionales, oratoria jurídica, etc.).

Figura 5. Arquitectura de la plataforma GARALEX

El núcleo de coordinación de esta estructura web lo constituiría HIZLAN (Ezeiza 2009a, b y c; Ezeiza, 2010), una plataforma concebida para cargar, procesar, clasificar y analizar producciones representativas de las diversas comunidades de práctica que desarrollan su actividad en este campo que, además, cuenta con diversas funcionalidades que permiten crear, administrar y ofrecer dentro del mismo entorno recursos de formación y consulta desarrollados a partir del estudio de los corpus alojados en ella. Para ello se cuenta con una base documental preparada y configurada para generar y administrar un número indeterminado de corpus dentro del sistema.[174] Sobre los corpus integrados en dicha base

173 http://www.ehu.es/ehusfera/ejm [Fecha prevista de puesta en marcha: 30-06-2011]
174 *Dokumentu Biltegia* ® SS-202-10

documental se puede operar selectivamente mediante una serie de módulos de gestión y de análisis lingüistico y extracción terminológica (motor de búsqueda de concordancias, análisis de frecuencias, bigramas y trigramas, detección de patrones morfosintácticos, contraste léxico de textos, etc.) que permiten obtener selectivamente datos y comparativas intra e intercorpus. También permite discriminar aquellos documentos que cumplen un determinado conjunto de rasgos y generar con dichos documentos un nuevo corpus o subcorpus en el sistema.

Esta plataforma la integran, además, una base de datos léxico-terminológica[175], una base de datos lingüística[176] y una base de datos bibliográfica[177]. Todos estos intrumentos constituyen la base de conocimiento que dará soporte al proyecto. En la figura nº 6 se destacan los módulos principales de HIZLAN. En la figura nº 7, por su parte, se puede apreciar una captura de pantalla correspondiente al motor de detección de frecuencias de uso[178].

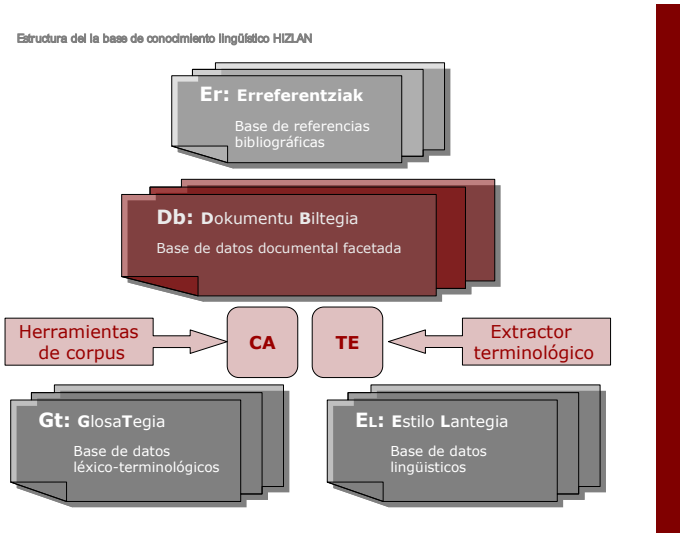


Figura 6. Estructura del elemento nuclear de la plataforma GARALEX

175 *Glostegia* (Gt) ® SS-29-11
176 *Estilo Lantegia* (EL) ® SS-28-11
177 *Erreferentziak* (Ef) ® SS-27-11
178 En este ejemplo se observa la distribución de lemas constituidos con el sufijo derivativo *–le* (lista de la izquierda) presentes en los corpus activos en el sistema (distribuidos en columnas)

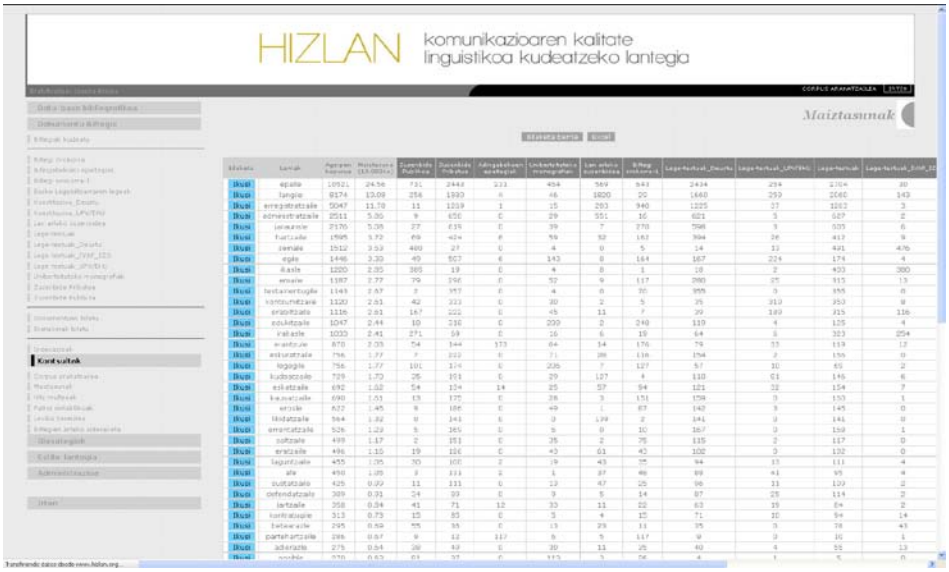


Figura 7. Captura de pantalla de una consulta de distribución de frecuencias

A la fecha de esta comunicación, la arquitectura informática se encuentra completamente desarrollada[179]. En la base documental hay procesadas más de 4 millones de palabras. El objetivo inmediato es alimentar el corpus con textos doctrinales, académicos, legislativos y jurídicos hasta alcanzar un volumen de 8-10 millones de palabras. Además del taller de trabajo, en este momento están plenamente activos el blog de la asignatura de Euskera Jurídico[180], en el que los estudiantes hacen públicos los resultados sus reflexiones, investigaciones, discusiones, etc. y en el que se pueden encontrar ya diversos glosarios, listados de expresiones y orientaciones de estilo elaborados en el marco de la asignatura (ver figura nº 8). También se encuentra plenamente activa la cuenta Twitter de GARALEX,[181] que sirve como lista de distribución de todas aquellas noticias y novedades de interés, relacionadas con los temas del proyecto. Además se encuentran accesibles en la red diversos materiales didácticos multimedia,[182] elaborados para que aquellos estudiantes, profesores o profesionales que lo deseen, puedan utilizarlos para su propia formación o para que aquellas entidades que organizan cursos de capacitación lingüística en este campo puedan tenerlos a su disposición.

179 URL: http://www.hizlan.org/db-zuzenbidea
180 URL: http://www.ehu.es/ehusfera/garalex
181 URL: http://twitter.com/#!/GARALEX_EHU
182 http://tresnak.kapsula.com/komunikazioa_euskaraz/administrazio_orokorra/
http://tresnak.kapsula.com/komunikazioa_euskaraz/justizia_administrazioa1/
http://tresnak.kapsula.com/komunikazioa_euskaraz/justizia_administrazioa2/
http://tinyurl.com/5s9s3r4

Todos estos elementos se integrarán próximamente en el portal GARALEX, que a la fecha de la redacción de este texto (05-04-2001) se encuentra en fase de construcción. La versión de prueba se está desarrollando sobre la plataforma de blogs de la Universidad del País Vasco (EHUSFERA),[183] utilizando la aplicación WordPress. No obstante, una vez consolidado el diseño, se tratará de obtener financiación para dotar al portal de un diseño más sólido, atractivo y funcional. Una vez finalizada esta fase, está previsto elaborar un plan de lanzamiento, que incluirá algunas acciones para involucrar a aquellos agentes tanto de la propia universidad como de otras instituciones. El objetivo es crear una red de colaboración estable entre el mundo académico y el profesional y los operadores del ámbito de las Ciencias Jurídicas y la Lingüística Aplicada para avanzar juntos en este proyecto. Los contactos realizados hasta la fecha han resultado muy prometedores, y avanzan que el proyecto pueda echar a andar ya con paso firme muy próximamente.

Referencias

Bernard, T. (2009). Jurilingüística: Las equivalencias interlingüísticas en Derecho. En Álvarez, M R., Álvarez, C. & Leonel, M. (eds.) Actas *del XI Simposio Internacional de Comunicación Social,* 58-61.

Cacciaguidi-Fahy, S. (2008). Quelques réflexions sur la linguistique juridique ou la jurilinguistique. *International Journal of the Semiotic of Law* 21, 311–317.

Ciuro, M. A. (2009). El verbo en el antecedente de la norma jurídica (un aporte a la "Jurilingüística" con especial referencia a la lengua española). *Revista del Centro de Investigaciones de Filosofía Jurídica y Filosofía Social* 32, 17-26.

Conru, G. (2005). *Linguistique juridique*. Paris: Montchrestien, Domat.

Ezeiza, J. (2009). Herramientas para la compilación, estudio y gestión de la producción

183 URL: http://www.ehu.es/ehusfera/ejm

lingüística en la universidad: una aproximación didáctica y social. En Caridad de Otto, E. & López de Vergara (comp.). *Las lenguas para fines específicos ante el reto de la Convergencia Europea.* La Laguna: Universidad de la Laguna, 553-567

Ezeiza, J. (2009b). *Criteris, metodologies i eines per a la gestió, l'estudi i la dinamització de la producció lingüística a la universitat: una aproximació socia*l. Bartzelona: UAB. Disponible en

http://www.slideshare.net/sdl_uabidiomes/josebaezeizaproducciolinguistica

Ezeiza, J. (2009c). *Dokumentu Biltegia*: *Proiektuaren diseinua eta protokoloa jasotzen duen txostena*. Informe de investigación no publicado. ® SS-239-09.

Ezeiza, J. (2010). DB (*Dokumentu Biltegia*): corpus akademikoak sortzeko eta kudeatzeko azpiegitura teknologikoa. En Salaburu, P. eta Alberdi, X. (arg.), 2010. *Euskararen garapena esparru akademikoetan*, 168-190.

Ferran, E. (2005). La traducción del documento jurídico negocial fundamentada en las funciones jurilingüísticas. En Monzó, y Borja, A. (eds.) *La traducción y la interpretación en las relaciones jurídicas internacionales*. Castelló de la Plana: Universitat Jaume I, 243-254.

Gémar, J. C & Kasirer, N. (eds.) (2005). *Jurilinguistique: entre langues et droits / Jurilinguistic : between law and language*. Montréal: Éditions Thémis.

Gémar, J. C. (dir.) (1982). *Langage du droit et traductio : essais de jurilinguistique / The language of the law and translatio : essays on jurilinguistics*. Québec: Linguatech.

Isani, S. & Lavault-Olleon, E. (2009). A la Confluence des Langues, des Cultures et du Droit: Jurilinguistique et Traduction. *International Journal of the Semiotic of Law* 22, 451–458.

Mattila, H. E. S. (2008). Les abréviations juridiques: méthode de recherche jurilinguistique. *International Journal of the Semiotic of Law* 21, 347–361.

Wagner, A. (2008). Les fondements de la construction du savoir du jurilinguiste. *International Journal of the Semiotic of Law* 21, 377–384.

Implementing an academic corpus in the English Language classroom in tertiary education.

Miguel Fuster-Márquez

Begoña Clavel-Arroitia
IULMA-Universitat de València

In this article we propose a model that allows teachers to integrate corpus linguistics (CL) in their English language courses at university level. This research is still in progress since we need to assess the results at the end of this academic year. The subjects of this study are our own students in the second year of the compulsory module English Language IV of the new degree of English Studies at the Universitat de València. It is precisely in the new university paradigm, in which students are required to be active participants of their own learning and solve problems autonomously, that the deployment of CL can be employed to enhance students' potential in this direction. The purpose of our project is no other than to offer a coherent procedure to promote corpus exploitation, either by teachers through the design of corpus-based activities, or by students themselves. We believe that the inductive approach through corpus-driven awareness-raising activities that we have applied is in conformity with the main guidelines being implemented in higher education pedagogy.

Key words: Corpus Linguistics, Second Language Acquisition, Academic English, higher education

En nuestro artículo proponemos un modelo que permita a los profesores integrar la Lingüística de Corpus (LC) en cursos de lengua inglesa a nivel universitario. Esta investigación se halla en curso ya que necesitamos evaluar los resultados al final del año académico. Los sujetos del estudio son nuestros propios estudiantes de segundo curso de una asignatura obligatoria, Lengua Inglesa IV, del recién implantado grado de Estudios Ingleses en la Universitat de València. Es precisamente en el nuevo paradigma universitario, en el que a los estudiantes se les requiere ser partícipes de su propio aprendizaje y que resuelvan problemas con autonomía, donde la implementación de metodologías de corpus puede contribuir a la mejora del potencial de los estudiantes en esa dirección. El propósito del proyecto no es otro que ofrecer un procedimiento coherente para fomentar la explotación de corpus, bien por parte de los profesores a través del diseño de actividades basadas en corpus, o bien por parte de los alumnos. Creemos que un enfoque inductivo mediante actividades de toma de conciencia mediante corpus se halla en línea con las directrices pedagógicas fundamentales puestas en marcha en la enseñanza superior.

Palabras clave: Lingüística de Corpus, Adquisición de Segundas Lenguas, Inglés Académico, educación superior.

1. Introduction

In this article we propose a model that endeavours to integrate corpus linguistics (CL) in the teaching of *English Language* subjects in tertiary education. We are reporting research in progress and the initial results can only be assessed at the end of this academic year. The recipients of this study are our students in the second year of a compulsory module, *English Language IV*, of the newly implemented degree of English Studies at the *Universitat de València*. This project focuses on the use of corpus in the development of productive written skills in ESL. It is our contention that if teachers wish to embark on CL as part and parcel of their teaching methodologies there is no need to resort exclusively to large corpora like the BNC or *The Bank of English*. Our proposal is more modest and envisages compiling smaller corpora which can immediately be put to work offline in the classroom by means of software tools such as *AntConc* or *Wordsmith tools.*

When the project has been completed it will contain three corpora. One of them will consist of a database of twenty million words drawn from recent articles published by quality newspapers from the UK and the USA. Ideally, this corpus would represent "General written English". A second corpus will contain recent academic articles and reviews published in leading journals, as well as other samples of academic writing but exclusively in the field of *English Studies*. For this corpus, we deem that one million words should suffice; at present it contains well over 350 thousand words, from 32 texts. And the third one will be a smaller learner corpus, whose size and make-up are still being discussed. This last corpus, containing our students' production, should provide a more accurate picture of students' progress at their different learning stages. The three corpora may be improved, updated or enlarged according to pedagogical criteria.

A number of corpus linguists have pointed out the appropriateness of collecting smaller targeted corpora (Franca, 1999; Kind & Wright; O'Keeffe & Farr, 2003; Partington, 1998; Stubbs, 2002). Mendikoetxea, Murcia-Bielsa & Rollinson (2010) state that a growing number of researchers are collecting their own pedagogical corpora. In their view, compiling a corpus where errors are identified can help teachers design more suitable pedagogical materials.

The three corpora in our project can contribute to the development of both receptive and productive skills. Here, however, we will focus on our recent experience where the academic corpus is being used to develop students' writing skills. The corpus has been compiled to meet the demands of the curriculum in our degree, since our students' learning goals include the attainment of competence in Academic English (AE) in the fields of English Linguistics and Literary Criticism.

2. Defining the teaching context:

Eight compulsory English Language Courses are offered during the eight semesters of the new degree in *English Studies* at the *Universitat de València* which has been launched during the academic year 2010-2011. In this degree, the role of *English Language* subjects is currently more central than in the past. *English Language IV* is a course that students take during the second semester of their second academic year.

The course is conceived in such a way that students are not just given input in what has been termed 'General English'. Our students are required to be conversant with AE in order to write papers in this language. Moreover, they may need these skills in their professional careers. In short, these students are expected to be proficient in the academic environment of their own studies: *Linguistics and Literary Criticism*.

To our knowledge, insufficient attention has been given to this issue. The impression is that in the past students majoring in English studies had little specific training in academic writing as lecturers were more inclined to put greater emphasis on discipline-specific content. Also, not infrequently, AE has been delivered without clear coordinated principles or without consideration of the students' background. It should be clear that the academic skills these students require cannot be successfully tackled in a semester, an academic year, or by using a specific course book. Instead, what is required is a more complex sustained and coordinated effort which involves various courses.

Our decision has been to introduce this newer approach on a pilot basis to two small groups of students in *English Language 4*. At this stage, our primary goal is that students attain a B2+ level. The *Common European Framework of Reference for Languages* (CEFR) quite explicitly establishes that B2 and C1 should test learners' academic language skills. For instance, C1 learners' production should show that a student:

Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices. (p. 24)

The implication is that AE becomes bound up with the specific purposes English is put to. In our case, these relate very straightforwardly to disciplines such as Language, Linguistics or Literary Criticism.

3. The presence of corpus linguistics in language teaching

Even though we were keen to adopt CL, how it could be rendered useful necessitated some discussion. On our side was that an experiment we had carried out a year earlier showed that students had a high regard for corpus implementation in the classroom to the extent that they considered it even more interesting than other ICTs they had been exposed to up to then. Nevertheless, for that experiment students had been exposed to two sessions of CL so the results were not applicable here.

A large number of studies have dealt with the desirability of introducing CL by way of the so-called *direct approach* (DA) either by way of *data driven learning* (DDL) (Johns, 1991, 1994; Sinclair, 2004), or *discovery learning* (Bernardini, 2004). Yet, to our knowledge no cases have been mentioned in the literature where a DA has been fully integrated in courses. Therefore, little can be said about the effectiveness of this methodology. However, what lies ahead of us is an entirely unexplored area of future research which should be of concern to corpus practitioners and foreign language teachers. The lack of

a more integrated corpus approach in language teaching has been variously justified as due to:

(1) A confusion over the distinction between what is 'scientifically interesting' and what is 'pedagogically useful' (Kennedy, 1992: 364–367).

(2) The fact that many CL experiences reported teachers' and learners' problems when working with large corpora (Lavid, Arús Hita & Zamorano-Mansilla, 2010).

(3) The lack of suitable corpora and software (see Ädel 2010: 44-51; Fuster-Márquez, 2010: 271).

(4) From the perspective of the students, unfavourable factors associated with the complexities posed by CL exploration (Ädel 2010: 44-51).

The overall impression was that we were embracing a methodology with many obstacles, but which, for those who believe in its benefits, was worth testing.

4. Implementing academic english in an english language course

We had observed that whereas the course book we had chosen addressed different kinds of writing, AE was not a central concern. To redress this imbalance, we introduced some additional tasks based on a thorough examination of an article in the field of English Linguistics that had been selected. After suggesting a careful reading of the article, students were asked to detect salient features of the language therein and give their opinion.

Our policy towards corpus implementation was that CL should be practised together with exposure to full authentic oral and written texts. We hoped to provide an answer to the criticism levelled against CL, namely that it offers a limited bottom-up approach to language through decontextualised concordances (see Flowerdew, 2009). We agree that the DA is rendered serviceable only when complemented with exposure to authentic texts. Therefore, a desirable combination of bottom-up and top-down approaches can only be achieved through the use of full academic texts coupled with the authentic language offered in a select corpus which represents that genre.

In our opinion, AE discourse requires a more detailed treatment. Our approach does not imply ignoring the larger freely available corpora which also contain samples of academic speech, for instance COCA and BNC [see Mark Davies' site]. On the contrary, we have started to integrate *BAWE* and *MICUSP*, both of which contain academic essays produced by native students. Consequently, these are deemed to be much closer to the needs of our students. Briefly, the British Academic Written English corpus (BAWE) contains 2761 pieces of proficient student writing, varying in length from 500 to 5000 words. The Michigan Corpus of Upper-level Student Papers (MICUSP) is a collection of around 830 A grade papers (roughly 2.6 million words) from a range of academic disciplines of the University of Michigan. Nevertheless, it was decided that the compilation of a smaller corpus with a special focus on features of AE representative of disciplines in English Studies was extremely useful.

So far, in the pilot study, we have had the impression that CL has been quite useful to address the needs of our own students. A caveat, due to copyright restrictions we cannot solve, we do not allow students to manipulate this corpus outside the classroom. This is always done under the instructor's supervision. In order to perform corpus searches outside the classroom, we suggest that they explore the BNC and COCA. Searches that have been proposed could be as simple as analysing the frequencies of common adjectives in our academic context which contain the suffix *–ic/-ical*:

analytic	*analytical*
theoretic	*theoretical*
phonologic	*phonological*
syntactic	*syntactical*

Peters (2004: 264) observes that a large number of *ic-/-ical* English adjectives may appear in two forms. Often, there is no linguistic reason why one should be preferred over the other, as for instance *analytic/analytical, geographic/geographical.* Our corpus exploration allowed students to unveil the different frequencies. In *analytic* (6) vs *analytical* (33), a greater preference for *analytical* arose. In the contrast between **phonologic* (o) vs *phonological* (124), only the latter was possible. In sum, this simple awareness-raising activity highlighted that individual pairs had to be examined carefully one by one. After the session, the students were asked to explore larger corpora at home and draw conclusions.

Other more sophisticated and complex searches included an examination of frequency, usefulness and formal aspects of discourse markers, linking elements or conventions about stance. We examined the different styles displayed in the essays proposed in the students' textbook and that of the article they were asked to read. They could observe that their textbook encouraged them to use formulas which were hardly ever found in the research article, for instance, in the expression of stance, a stylistic feature used to convey the author's personal attitude (Biber, Johansson, Leech, Conrad & Finegan, 1999: 966). The following expressions were proposed in the textbook: *In my opinion*, *I think, I believe*, *I feel.* By contrast, through corpus searches and the article at hand, they were able to detect that these were typically avoided. The option *I feel* yielded two cases, one of which belonged to the genre of reviews. Also, in this respect, Biber et al. (1999: 360) confirm that *I know, I think, I mean,* are informal or conversational language features. Nevertheless, our analysis detected that the word-form *feel* was present in the corpus for various other purposes. However, two instances of *we feel*, seemed to confirm that it was a minor stance option authors might resort to in written academic prose. Naturally, this led to later discussions about the use of '*we*'. In contradistinction to our findings, BAWE showed that *I think*, *I feel,* or *In my opinion* were quite regularly used by British students [see figures 1, 2 and 3]:

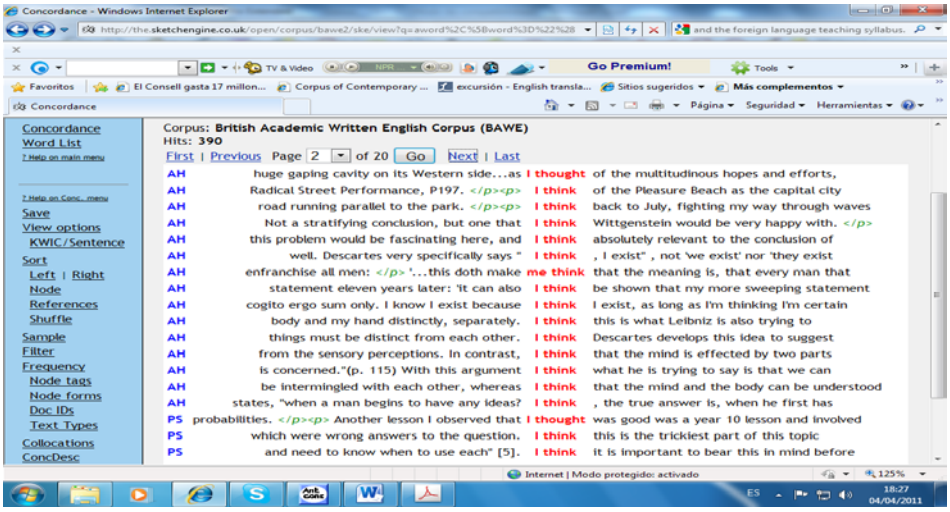


Figure 1. Search in BAWE corpus.

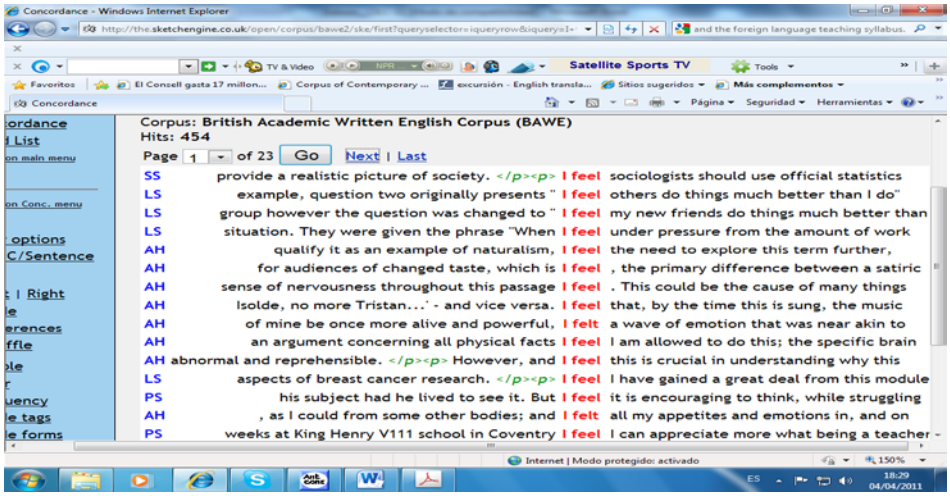


Figure 2. Search in BAWE corpus.

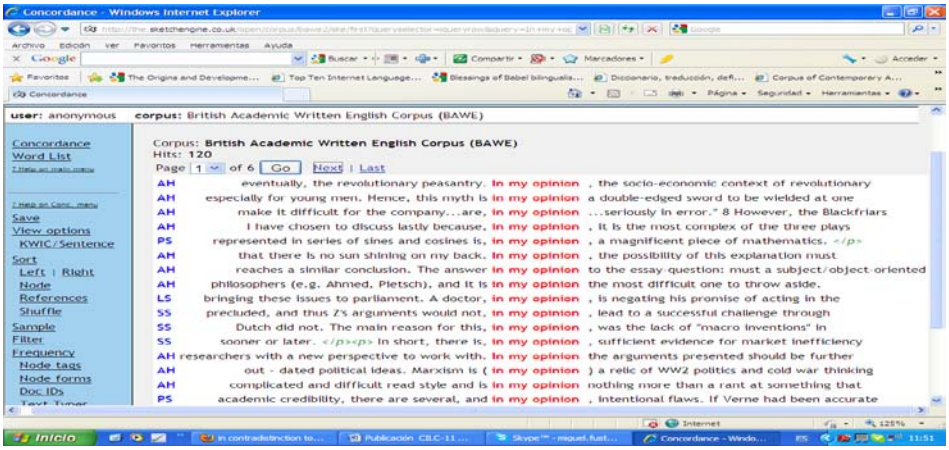


Figure 3. Search in BAWE corpus.

What emerged was that professional academic writers are taught to approach, present or report their research much more 'tactfully', so to say. Of course, this is not what happens with the type of academic writing undergraduate native students had to produce. They are asked to express their opinion on the subject, as our searches of *BAWE* revealed. The high number of hits for *think-thought, feel-felt* proved that these options were encouraged in their own writing.

Consequently, we recommended that our students make use of these more 'personal' options in the individual academic essay they had to write. But that they should take into account that the AE represented by articles, or textbooks tends to be more formal, and that these options were avoided as a rule in these contexts.

Various discourse markers and linking words dealt with and practised in different sections of their textbook, with a wide range of purposes, were common and important in serious academic writing. For instance:

a. Ways of expressing causes and results: *as a result, one of the main reasons is that, as a consequence, because.*

b. Ordering ideas: *firstly, secondly, finally, also, furthermore, what is more…*

c. Expressing contrast: *however, on the one hand, on the other hand*.

In this respect, we found somewhat puzzling that whereas various ways of concluding were offered in the text, e.g. *summing up,* or expressing one's opinion as a way of structuring the last paragraph, other simple options like *to conclude, in short,* etc. were not found. Although we had noticed that former students not infrequently made use of *as a conclusion,* an expression avoided by native writers, we decided that we would not comment on this feature unless it was observed as an error in their writing (Clavel and Fuster, 2010).

Discourse markers containing more than one word presented some difficulties. The case of *on the other hand* was mentioned to illustrate that this was the only option, so that **in*

the other hand was not acceptable. We also indicated that *on the other hand* was used to indicate contrast, even if *on the one hand* was not present. A quick corpus exploration in our corpus confirmed this point: the expression *on the one hand* was much more infrequently used than *on the other hand.* Students raised their questions and so searches were performed in order to detect possible variants of *on the one hand,* as for instance *on one hand.* The students themselves suggested exploring other options. Also, we observed that some apparently simple markers like *despite* caused problems because of the type of sentence construction required. Part of the problem was caused by interference of Spanish *a pesar de*.

Even though the examination of frequencies was revealing, we always assumed that other questions regarding academic writing required a view of the wider context, and that CL was not called for in those cases. A case in point is paragraph structure. Our prior observation was that students' essays often contained one-sentence paragraphs, perhaps because they did not think about the overall structure they wished to impose on the text or, most likely, because they had not been taught to consider paragraph structure. Attention to essay organisation and paragraph structure was achieved by examining the sample essays, articles and reports in the textbook, but supplemented with specific questions which were not being addressed in it, together with the close examination of paragraph structure in the academic article at their disposal.

Conclusions

Although this pilot project is not completed, and students' acceptance of corpus exploration in the classroom is a key factor about which we still have little information, we believe that CL should play a more significant role in ELT. We have used a small academic corpus to address problems within the area of AE and formal writing. In our view, the class textbook we set had shortcomings in this respect mostly due to the fact that textbooks in general do not address the needs of students in *English Studies*.

We believe that there is no reason why CL should not be used regularly, as we have done. However, as this approach becomes a reality, teachers need to address some of the challenges or drawbacks mentioned in the literature. Moreover, they should have an open mind and invite students to explore for instance, the capabilities offered by *WebCorp,* or the search engines available on the internet.

References:

Ädel, A. (2010). Using Corpora to Teach Academic Writing: Challenges for the Direct Approach. In M.C. Campoy-Cubillo, B. Bellés-Fortuño & M. L. Gea-Valor (Eds.), *Corpus-Based Approaches to English Language Teaching* (pp. 39-55). London-New York: Continuum.

Bernardini, S. (2004). Corpora in the classroom. An overview and some reflections and future developments. In J. McH. Sinclair (Ed.), *How to Use Corpora in Second Language Teachin* (pp. 15-36). Philadelphia: John Benjamins Publishing Company.

BIBER, D., JOHANSSON, S., LEECH, G., CONRAD,S. & FINEGAN, E. (1999). *Longman Grammar of Spoken and Written English.* Harlow: Pearson Education.

COMMON EUROPEAN FRAMEWORK OF REFERENCE FOR LANGUAGES (CEFR). Retrieved from http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

FLOWERDEW, L. (2009). Applying corpus linguistics to pedagogy: A critical evaluation. *International Journal of Corpus Linguistics 14*(3), 393–417.

FRANCA, V. B. (1999). Using student-produced corpora in the L2 classroom. In P. Grundy (Ed.), *IATEFL 1999 Edinburgh Conference Selections* (pp. 116-117). Whitstable: IATEFL.

FUSTER-MÁRQUEZ, M. (2010). The challenges of introducing corpora and their software in the English lexicology classroom: some factors. In I. Moskovich, B. Crespo, I. Lareo & P. Lojo. (Eds.), *Language Windowing Through Corpora. Visualización del lenguaje a través de corpus* (pp. 269-288). Coruña: Universidad de Coruña.

FUSTER-MÁRQUEZ, M & CLAVEL-ARROITIA, B. (2010). Second Language Vocabulary Acquisition and its Pedagogical Implications. In L. Pérez Ruiz, I. Parrado Román & P.Tabarés Pérez (Eds.), *Estudios de Metodología de la Lengua Inglesa (V)* (pp. 205-212). Valladolid: Secretariado de Publicaciones Universidad de Valladolid.

JOHNS, T. (1991). Should you be persuaded – two examples of data-driven learning materials. *English Language Research Journal, 4*, 1-16.

JOHNS, T. (1994). From printout to handout: grammar and vocabulary teaching in the context of data-driven learning. In T. Odlin (Ed.), *Perspectives on Pedagogical Grammar* (pp. 293-317). Cambridge: Cambridge University Press.

KENNEDY G. (1992). Preferred ways of putting things. In J. Svartvik (Ed.) *Directions in corpus linguistics* (pp 335-73). Berlin: Mouton de Gruyter.

KINDT, D & WRIGHT, M. Integrating Language Learning and Teaching with the Construction of Computer Learner Corpora. Retrieved from http://www3.nufs.ac.jp/~kindt/media/corpora.pdf

LAVID, J. (2007) Global and local attention in conversation: the case of task-scheduling dialogues. In C. Butler, R. Hidalgo & J. Lavid (Eds.), *Functional perspectives on grammar and discourse: In honour of Angela Downing* (pp. 313-326). Amsterdam: John Benjamins.

LAVID, J., ARÚS HITA, J. & ZAMORANO-MANSILLA, J.R. (2010). Designing and Exploiting a Small Online English-Spanish Parallel Textual Database for Language Teaching Purposes. In I. Moskovich, B. Crespo, I. Lareo & P. Lojo (Eds.), *Language Windowing Through Corpora. Visualización del lenguaje a través de corpus* (pp. 138-148). Coruña: Universidad de Coruña.

MENDIKOETXEA, A., MURCIA-BIELSA, S. & ROLLINSON, P. (2010). Focus on Errors: Learner Corpora as Pedagogical Tools. In M.C. Campoy-Cubillo, B. Bellés-Fortuño & M. L.

Gea-Valor (Eds.), *Corpus-Based Approaches to English Language Teaching* (pp. 180-194). London-New York: Continuum.

O'Keeffe, A. & Farr, F. (2003). Using Language Corpora in Language Teacher Education: pedagogic, linguistic and cultural insights. *TESOL Quarterly,* 37(3), 389-418.

Partington, A. (1998). *Patterns and Meaning*. Philadelphia: John Benjamins.

Peters, P. (2004). *The Cambridge Guide to English Usage.* Cambridge: Cambridge University Press.

Sinclair, J. McH. (Ed.) (2004). *How to Use Corpora in Language Teaching.* Amsterdam: John Benjamins.

Stubbs, M. (2002). *Words and Phrases: Corpus Studies of Lexical Semantics*. Malden-Oxford: Blackwell Publishing.

La adquisición de alemán como lengua extranjera. Una aportación basada en un corpus de aprendices

Daniela Gil Salom
Universidad Politécnica de Valencia

Este estudio se centra en la adquisición de Alemán como Lengua Extranjera ab initio en el caso de estudiantes universitarios españoles durante su primer año de aprendizaje formal. En primer lugar, revisamos investigaciones previas sobre la adquisición de la sintaxis y/o morfología en el aprendizaje de Alemán como Segunda Lengua. También ampliamos el campo de estudio incluyendo nuevas variables. Por medio del análisis de textos escritos por seis grupos de estudiantes, llegamos a la conclusión, de que la secuenciación en la adquisición de Alemán como Segunda Lengua es diferente en el caso del aprendizaje formal de Alemán como Lengua Extranjera.

Alemán como Lengua Extranjera, secuenciación, adquisición, análisis de textos escritos

This study focuses on the acquisition of German as a Foreign Language ab initio in the case of Spanish university students during their first year of formal learning. First, we review previous investigations on the syntactic and/or morphologic acquisition of German as a second language. We also extend the field of study including some new variables. Through the analysis of a number of texts written by six groups of students, we conclude that the sequencing in the acquisition of German as a Second Language is different from that of the acquisition of German as a Foreign Language in a formal context.

German as Foreign Language, sequencing, acquisition, written texts analysis

1. Introducción

El punto de partida de este trabajo es la reflexión sobre cuestiones básicas relacionadas con la enseñanza de Alemán como Lengua Extranjera, que estudiamos en nuestra tesis doctoral. En primer lugar, nos preguntamos qué adquiere el estudiante en realidad de todos los contenidos que los docentes queremos introducir en las aulas. En segundo lugar, si lo hace siempre siguiendo el mismo orden y, en tercer lugar, hasta qué punto influye nuestra práctica docente en esa adquisición.

Para abordar estos temas, nos planteamos tres objetivos básicos: confirmar las secuencias de adquisición observadas en anteriores estudios, verificar la posible correlación entre la adquisición de la sintaxis y la adquisición de la morfología y comprobar la influencia de la enseñanza en las aulas.

Para ello, revisamos primero anteriores investigaciones relativas a la adquisición del alemán como segunda lengua (L2) y como lengua extranjera) LE y a continuación, ampliamos el campo de estudio (siempre que nos ha sido posible) a la adquisición de elementos, que no habían sido incluidos en las investigaciones revisadas, procurando contar con el mayor número posible de sujetos.

Nuestra investigación se ha centrado pues en responder a las tres preguntas de investigación que concretan los pasos de nuestro trabajo: Primero, hemos comparado la adquisición del alemán como segunda lengua (L2) y como lengua extranjera (LE); a continuación hemos descrito el desarrollo de la adquisición en el caso del aprendizaje guiado de alemán como lengua extranjera (LE) por discentes con español como primera lengua (L1); y, por último, hemos comprobado si la alteración en el orden de introducción de los elementos en el aula influye en esta adquisición.

2. La adquisición de Alemán como Lengua Extranjera

La mayoría de los estudios consultados, se centran en la adquisición del orden de palabras en la oración alemana, iniciado por el proyecto *ZISA* en 1983 (Clahsen, Meisel y Pienemann, 1983). Los resultados de dicho proyecto indicaban que después de una fase inicial de palabras aisladas, los inmigrantes de origen italiano, español y portugués, seguían una secuencia de adquisición en 5 etapas:

1º Orden canónico (sujeto-verbo-objeto) representado por SVO.

2º Anteposición del adverbio (adverbio-sujeto-verbo-objeto) o ADV.

3º Separación de verbos (paréntesis oracional) o SEP.

4º Inversión del orden (sujeto-verbo por verbo-sujeto) o INV.

5º Colocación del verbo al final de la frase (*verb final*) o VEND.

La secuenciación en la adquisición de estas estructuras sintácticas se formuló como universalmente válida para todo sujeto independientemente de su L1 o contexto. Como desde entonces han surgido muchos trabajos y distintos enfoques sobre este tema, nosotros los hemos clasificado según los resultados obtenidos.

Así encontramos publicaciones que confirman el paralelismo entre adquisición natural y formal (Boss, 1996; Clahsen y Muysken, 1986; Ellis, 1992; Jansen, 1991; Pienemann, 1987; Pienemann, 1989; Tschirner, 1992; Westmorland, 1983) como hallamos otras, que lo niegan o dudan (Boss, 2004; Lund, 2004). Además, hemos destacado aquellos que parten de sujetos con español como L1 (Ehlers, 2001; Grümpel, 2009; Martínez Adrian, 2008;) y aquellos que incluyen la adquisición de la morfología en su análisis (Ballestracci, 2008; Boss, 1998; Clahsen, 1988; Clahsen y Muysken, 1989; Diehl, Christen, Leuenberger, Pelvat y Studer, 2000; Jordens, 1988; Klein-Gunnewiek,1997; Meerholz-Härle y Tschirner, 2001; Tschirner, 1999; Vainikaa y Young-Scholten, 1994).

Observamos que durante la década de los 80 y 90 predominan los estudios a nivel sintáctico, la lengua materna era el inglés, los textos analizados eran casi todos de producción oral, pero a partir del año 2000 se incluyen ya elementos morfológicos, se analizan también textos escritos y empiezan a surgir trabajos con discentes con español como lengua materna.

Nuestra aportación estriba en analizar tanto aspectos sintácticos como morfológicos, de textos escritos de producción libre, de seis grupos de un total de 66 estudiantes que corresponden a tres momentos del segundo semestre de aprendizaje y en valorar los resultados, no solo a nivel de grupo, sino también individualmente. De esta forma ampliamos el campo de estudio en dos sentidos: por un lado, analizamos aspectos sintácticos y morfológicos (dentro de los cuales incluimos tres elementos más, que no se contemplan en la bibliografía más reciente, y que son: el plural, el acusativo y el dativo) y por otro lado, analizamos textos de un mayor número de individuos. Aquí describimos el estudio de los elementos sintácticos, los aspectos morfológicos se describen en Gil Salom (2010).

3. Materiales y método de estudio

Pasamos ahora a describir el corpus y el método de estudio seguido. Hemos procurado ser lo más rigurosos posible a la hora de seleccionar a nuestros sujetos. Analizamos textos de 66 estudiantes de la UPV que asistieron a clases de lengua alemana como segunda o tercera LE como asignatura Optativa o de Libre Elección, durante dos semestres, distribuidos en seis grupos y con cinco profesores distintos. De 139 estudiantes prescindimos de 73 para que cumplieran las mismas características, como la misma L1, los conocimientos previos o la producción de dos textos como mínimo (uno al principio y otro al final del segundo semestre). La variedad de grupos **y** docentes nos ofrece más objetividad en los resultados, así como la posibilidad de hallar distintas progresiones en la enseñanza. También se dan diferencias en la temporización, elementos que en unos grupos se introducen en el primer semestre, son introducidos en otros grupos, en el segundo. En cuanto a los textos, fueron escritos en el aula, bien como pruebas de diagnóstico o bien como exámenes. Si estos textos son auténticos o no, era fundamental para confirmar que estábamos analizando conocimientos implícitos y no explícitos, es decir, que se trataba de una producción natural, libre y no dirigida en exceso (como ocurre con pruebas de oraciones deshidratadas o traducciones). Para su validez nos remitimos a Ellis y Barkhuizen (2005:50) cuando afirman:

It might be questioned whether samples of written language produced in the context of an examination are 'natural'. We argue they are, on the grounds that an examination constitutes a 'natural' context for learners to use the L2 and that data so obtained have not been designed for purposes of resarch. (Ellis & Barkhuizen, 2005:50).

Dada la cantidad de elementos a analizar creamos una base de datos para cuantificarlos y clasificarlos. Utilizamos la aplicación informática *File Maker* fundamentalmente por su fácil manejo. Para conocer las características individuales de cada sujeto, creamos campos dedicados al grupo, la edad, los conocimientos en otras LE, el tiempo de aprendizaje de la lengua alemana, la posible motivación extra, los posibles refuerzos fuera del aula y la continuidad del aprendizaje

En cuanto a los elementos gramaticales, en sintaxis, analizamos sólo 4 estructuras: las oraciones enunciativas (SVO), el paréntesis oracional (SEP), la inversión en las oraciones enunciativas (INV) y las oraciones subordinadas (VEND). Y en morfología, recogimos datos de la congruencia sujeto-verbo, la elección correcta del auxiliar y la forma correcta del participio perfecto para el *Perfekt*, del nominativo (en realidad del género), del plural, del acusativo y del dativo.

Las secuencias ofrecidas por Clahsen y otros (1983) eran cinco. Nosotros prescindimos de ADV, por no considerarla estructura a ser adquirida, sino como primer paso para adquirir la estructura de la inversión. En realidad ADV equivale a las producciones erróneas de INV. Esta fase está por tanto incluida en la misma estructura INV.

4. Análisis de los datos

Para describir al análisis de nuestra investigación es necesario aclarar previamente, que en las tablas que mostramos a continuación, no aparecen datos del Texto 2 de todos los grupos, porque solo pudimos recogerlos en las dos ocasiones mínimas (al principio y al final del semestre, Texto 1 y Texto 3, respectivamente). Únicamente obtuvimos un segundo texto en los grupos B y F.

4.1 Oraciones enunciativas

La primera estructura analizada, SVO, aún siendo la más sencilla de adquirir (coincide con la L1 de nuestros sujetos), sufre un descenso en la corrección en tres grupos (B, D y F). Advertimos aquí ya la variación de la interlengua, es decir, no se mantiene una adquisición completa de un 100 % (Tabla 1).

Tabla 1. Índices de corrección de SVO en los tres textos (≥ tres contextos obligatorios).

	A	B	C	D	E	F
Texto 1	99 %	99,5 %	96 %	100 %	99 %	100 %
Texto 2	-	96 %	-	-	-	88 %
Texto 3	96 %	89 %	100 %	86 %	100 %	92 %

Los grupos que bajan su rendimiento en SVO, son de los que tienen mejores índices en INV (D, B, F). Esto es debido a que los sujetos perciben las conjunciones de coordinación como señales de inversión, fenómeno que antes desconocían. Coincidimos, pues, en la observación que hacen Grümpel (2004) y Martinez Adrián (2004).

4.2 Paréntesis oracional

Para el estudio del paréntesis oracional, en nuestro trabajo diferenciamos entre verbos separables, verbos modales y *Perfekt.* Solo podemos analizar verbos separables en el Texto 1, ya que no hallamos contextos obligatorios suficientes en el Texto 2 ni en el Texto 3. El Texto 1 nos ofrece un 33% de corrección en esta estructura con este tipo de verbos. En cuanto al paréntesis oracional con verbos modales (Texto 1 – 80%, Texto 2 – 100%, Texto 3 – 78%) y con el *Perfekt* (Texto 2 – 98%, Texto 3 – 95%) obtenemos en ambos casos porcentajes de corrección altos y similares (por encima del 75%), por esta razón, nos hemos permitido presentarlos de forma conjunta en la Tabla 2.

Tabla 2: SEP por grupos y textos (≥ tres contextos obligatorios).

	A	B	C	D	E	F
Texto 1	-	-	-	25 %	100 %	78 %
Texto 2	-	100 %	-	-	-	98 %
Texto 3	92 %	97 %	93 %	96,9 %	92 %	84 %

Rieck en su estudio de 1989 para el proyecto *DiGS*, publicado en Diehl y otros (2000) observa dos fases en la adquisición de este elemento por los escolares de Ginebra: 1º v. modal – infinitivo y 2º v. auxiliar – participio. Nosotros no advertimos tal progresión. En cualquier caso, esta estructura parece homogéneamente adquirida.

4.3 Oraciones enunciativas con inversión

Sin embargo, la Tabla 3 nos muestra que la inversión no es adquirida por todos los grupos: Los grupos A (74%) y E (58%) no alcanzan el 75% y el grupo F justo el 75%, porcentaje que marca el umbral de adquisición. Los datos nos muestran una evolución distinta según los grupos. El grupo B y el grupo D mejoran sus resultados a lo largo del semestre. El grupo C sufre un pequeño retroceso al final de semestre, pero aún así mantiene los mejores resultados. En el grupo F se produce una recaída, desde un alto índice, hasta casi el nivel del inicio del semestre. Por último, los grupos A y E no producen suficientes contextos al inicio del semestre para poder observar alguna evolución, lo que corrobora sus índices inferiores.

Tabla 3. INV a lo largo del segundo semestre de instrucción (≥ tres contextos obligatorios).

	A	B	C	D	E	F
Texto 1	0 %	-	100 %	33 %	-	79 %
Texto 2	-	85 %	-	-	-	98 %
Texto 3	74 %	95 %	96 %	94 %	58 %	75 %

Ellis (1992) también obtuvo resultados distintos en esta estructura entre los grupos que analizó. La explicación sugerida por este autor fue la falta de motivación por parte de los estudiantes. Ésta podría ser también la explicación para nuestros resultados, pero lo que no deja de ser evidente y relevante es que para parte de nuestros estudiantes la estructura es asimilable y para otros no. Los datos ya no son tan homogéneos.

4.4 Oraciones subordinadas

En cuanto a la estructura VEND, observamos en la Tabla 4 que el grupo B en el Texto 2 alcanza un índice de corrección altísimo, un 95%, para bajar a un 70% en el Texto 3:

Tabla 4: VEND por grupos y textos (≥ tres contextos obligatorios).

	A	B	C	D	E	F
Texto 1	-	-	-	-	-	-
Texto 2	-	95 %	-	-	-	100 %
Texto 3	-	70 %	100 %	90 %	67 %	-

Este descenso tan marcado lo interpretamos como una reestructuración en la interlengua, el foco de atención de este grupo se dirige en este momento a un elemento muy concreto del que han de demostrar sus conocimientos: el *Perfekt*. Los grupos C y D alcanzan unos resultados muy buenos. En estos grupos se había adelantado la introducción de esta estructura al semestre A. Apreciamos claramente que esta estructura puede adquirirse simultáneamente a la de la inversión (INV). Autores como Diehl y otros (2000), Boss (2004), y Lund (2004) ya lo indican, aunque no siempre, como nosotros, cuentan con suficientes contextos obligatorios para demostrarlo.

4.5 Análisis general de la sintaxis

Revisadas pues las estructuras sintácticas, podemos afirmar que la adquisición en este aspecto es bastante homogénea en la primera etapa de aprendizaje, sin embargo, conforme va aumentando la dificultad y la cantidad de estructuras a tener en cuenta, aparecen las diferencias grupales (Tabla 5). Es evidente que los grupos C y D son muy buenos. No sólo obtienen los mejores resultados, sino que han ido compaginando las 4 estructuras desde el primer semestre.

Tabla 5. Valoración de la sintaxis.

	A	B	C	D	E	F
SVO	96	89	100	86	100	92
SEP	92	97	93	97	97	84
INV	74	95	96	94	58	75
VEND	-	70	100	90	67	-

Sin embargo, el análisis más detallado de las cifras a nivel individual y no por grupo, nos permite advertir que existen casos en los que se obtienen mejores índices de corrección en la estructura de la INV que en el paréntesis oracional. Como también se dan casos en los que la corrección de ambas estructuras es idéntica.

Tabla 6. Valoración de la sintaxis (cont.).

	A	B	C	D	E	F	Total individuos
SEP > INV	6	9	0	2	4	1	22 (39%)
SEP = INV	2	7	2	6	1	1	19 (34%)
INV > SEP	3	4	3	3	2	0	15 (27%)

La Tabla 6 muestra que de 56 sujetos que producen 3 o más contextos obligatorios de SEP y de INV en el Texto 3, 22, es decir el 39% demuestra mejor competencia de SEP que de INV, lo que siempre se ha afirmado en la bibliografía. Pero 19 estudiantes, que suponen el 34% alcanza el mismo nivel en ambas estructuras. Y, por último, 15, un 27%, utiliza INV con una mayor corrección que SEP. Lo que indican estos datos es que no siempre se cumple el carácter implicacional en la adquisición de estas estructuras tal y como se asumía en la bibliografía.

Como se aprecia en la Tabla 7, estos alumnos no pertenecen al mismo grupo. Nos preguntamos, entonces, si tienen algo en común. Observamos que hay sujetos de todos los grupos excepto de uno de los dos menos numerosos, el F. Es decir, están repartidos de forma más o menos proporcionada. No se trata de estudiantes con especial experiencia en LE, la mayoría sólo tiene conocimientos de inglés. Sí que se observa que gran parte de ellos muestra cierta motivación especial por aprender alemán, ya que desean trasladarse a un país de lengua alemana para continuar sus estudios o para realizar su PFC o prácticas en una empresa. Casi ninguno cuenta con refuerzos en su aprendizaje (como centros de estudio, contacto directo con nativos, etc.). Y muchos de ellos han seguido su aprendizaje de alemán de forma continuada, pero tampoco todos.

Tabla 7. Características individuales.

sujetos	edad	otras LE	motivación	refuerzo	continuidad
A 4	20	inglés, francés	2	—	sí
A 6	20	inglés	1	—	sí
A 12	21	inglés, francés	—	1	sí
B 1	20	inglés	2	—	—
B 2	22	inglés	1	—	—
B 11	23	inglés	2	—	—
B 17	24	inglés	1	1	sí
C 2	20	inglés	1	—	sí
C 4	19	inglés, noruego, italiano	1	—	sí
C 5	23	inglés	2	—	sí
D 2	23	inglés	1	—	sí
D 9	23	inglés	1	—	sí
D 10	25	inglés	—	—	sí
E 2	23	inglés	—	—	—
E 3	25	inglés, francés	—	1	—

Resulta pues evidente, que no siempre se sigue el mismo orden de adquisición en sintaxis, ni tan siquiera en lo que a las primeras fases se refiere. Existen diferencias individuales que no cumplen la escala implicacional considerada universalmente válida.

5. Resultados

En cuanto a una posible temporización, los datos nos indican que después de 15 semanas de clase pueden adquirirse los siguientes elementos: en sintaxis, las oraciones enunciativas y, en morfología, la concordancia sujeto-verbo y el plural. Después de 22 semanas ya puede adquirirse la estructura del paréntesis oracional, de las oraciones enunciativas con inversión y de las oraciones subordinadas. En morfología los discentes ya pueden haber adquirido el participio perfecto. Pasadas 30 semanas pueden adquirir también el género y el plural. Estas tres fases corresponden a la tres recogidas de datos (Textos 1, 2 y 3) y son los obtenidos siempre teniendo en cuenta los datos medios de los grupos.

6. Conclusiones

Con todos estos resultados, podemos responder a nuestras preguntas de investigación, planteadas al inicio del estudio:

A la pregunta relacionada con el paralelismo entre el aprendizaje natural y el dirigido, nuestra respuesta es afirmativa teniendo en cuenta los resultados por grupo, pero no

considerándolos individualmente. La bibliografía ha debatido entre el orden de adquisición de INV y VEND; ahora comprobamos que SEP tampoco es tan fácil de adquirir, o que INV no es tan difícil. Ambas estructuras son igual de ajenas a la L1 de los sujetos. En definitiva, si se centra la atención en un elemento más que en otro, puede acelerar su adquisición. Puede que influya el refuerzo de trabajos individuales, o simplemente, el hecho de que es una estructura diferente (como nos indica la *Teoría de la marcación*).

En cuanto a nuestra segunda pregunta de investigación, las secuencias que siguen discentes adultos siguiendo un aprendizaje dirigido, no observamos una escala implicacional clara ni en sintaxis ni en morfología. Y, por último, observamos que sí influye la progresión del libro de clase *Themen aktuell* (Aufderstrasse, H. 2003) utilizado por la mayoría de los profesores. Y confirmamos que el énfasis sí condiciona la adquisición, puesto que los resultados del dativo son mejores que los del acusativo. Estos elementos son introducidos en dicho libro, en ese orden.

7. Futuras investigaciones

Habrá que comprobar de nuevo los resultados con otros sujetos. Por ello, consideramos necesario contrastar estos resultados con datos nuevos (de otros sujetos en el mismo contexto). Es importante observar y analizar cómo se desarrolla ese énfasis, cómo se materializa esa atención a la forma: ¿sólo en el aula? ¿con el trabajo no presencial del alumno?

Para ello, necesitamos recabar de nuevo información de docentes y alumnos sobre el trabajo autónomo del alumno, sobre el refuerzo en determinados elementos y sobre las causas de los errores. Por encima de todo, creemos que es importante, como nos dicen Edmonson y House (2006:328), que en el proceso de investigación en Enseñanza de Lenguas estén implicados, explícitamente, tanto docentes como discentes.

8. Bibliografía

Aufderstrasse, H. (2003). *Themen 1 aktuell*. Ismaning: Max Hueber Verlag.

Ballestracci, S. (2008). Überindividuelle Merkmale des Grammatikerwerbs im Unterricht des Deutschen als Fremdsprache durch italophone Studierende. Ergebnisse einer empirischen Untersuchung. *DaF, 3*, 160-169.

Boss, B. (1996). German grammar for beginners – the Teachability Hypothesis and its relevance to the classroom. En C. Arbones- Sola, J. Rolin-Ianziti y R. Sussex (Eds.), *Who's Afraid of Teaching Grammar?,1,* (pp. 93-100). The University of Queensland Papers in Language and Linguistics.

Boss, B. (2004). Wann ich habe Freizeit, ich koche gern. Zum Erwerb der deutschen Inversion und Nebensatzwortstellung durch australische Studierende. *DaF, 1*, 28-32.

Clahsen, H. (1988). Parameterized grammatical theory and language acquisition: a study of the acquisition of verb placement and inflection by children and adults. En: S.

Flynn & W. O'Neil (Eds.), *Linguistic theory in second language acquisition*. (pp.47-75). Dordrecht: Reidel.

Clahsen, H., Meisel, J. & Pienemann, M. (1983). *Deutsch als Zweitsprache. Der Spracherwerb ausländischer Arbeiter*. Tübingen: Narr.

Clahsen, H. & Muysken, P. (1989). The availability of universal grammar to adult and child learners: A study of the acquisition of German word order. *Second Language Research, 2,* 93-119.

Diehl, E., Christen, H., Leuenberger, S., Pelvat, I., Studer, T. (2000). *Grammatikunterricht: Alles für der Katz?* Tübingen: Max Niemayer Verlag.

Edmonson, W.J. & House, J. (1993). *Einführung in die Sprachlehrforschung*. Tübingen: Narr Francke.

Ehlers, C. (2001). Parámetros del orden de palabras y transferencia. Factores internos de la adquisición del alemán como lengua extranjera. En S. Pastor y V. Salazar (Eds.), *Estudios de Lingüística. Universidad de Alicante. Anexo 1: Tendencias y líneas de investigación en adquisición de segundas lenguas.* (pp.73-94). Alicante: Universidad de Alicante.

Ellis, R. (1992). Are Classroom and Naturalistic Acquisition the Same? A Study of the Classroom Acquisition of German Word Order Rules. En R. Ellis (Eds.), *Second language acquisition and language pedagogy* (pp. 75-100). Clevedon.

Ellis, R. & Barkhuizen, G. (2005). *Analysing Learner Language*. Oxford: Oxford University Press.

Gil-Salom, D. (2010). La heterogeneidad en la interlengua de principiantes. Trabajo presentado en el *VII Congreso de la FAGE.* Valencia, Septiembre 2010.

Grümpel, C. (2004). El papel de los principios de la gramática universal y de la lengua primera en la adquisición del orden de palabras del alemán por adultos hispanohablantes. En C. Ehlers y A. Haidl, A. (Eds.), *Actas del III Congreso de la FAGE. El Alemán en España: Motivaciones y perspectivas* (pp. 48-61). Málaga: Servicio de Autoedición e Impresión.

Grümpel, C. (2009). Theoretischer und empirischer Vergleich zum deutschen Spracherwerb. *Revista de Lingüística y Lenguas Aplicadas, 4*, 89-101.

Jansen, L. M. (1991). The development of word order in natural and formal German second language acquisition, *Australian Working Papers in Language Development, 5*, 1-42.

Jordens, P. (1988). The acquisition of word order in L2 Dutch and German. En P. Jordens & J. Lallemann (Eds.), *Language Development* (pp. 149-180). Dordrecht: Foris.

Klein-Gunnewiek, L. (1997). Gibt es eine bestimmte Erwerbssequenz bei Deutsch als Fremdsprache? *Materialien Deutsch als Fremdsprache*, *46,* 434-447.

Lund, R. (2004). Erwerbssequenzen im Klassenraum, *DaF 2,* 99-103.

Martínez Adrián, M. (2008). La adquisición del parámetro de ascenso verbal en el alemán como tercera lengua, *Magazin*, *18,* 28-35.

Meerholz-Härle, B. & Tschirner, E. (2001). Processability Theory: eine empirische Untersuchung. En K. Aguado & C. Riemer (Eds.), *Wege und Ziele. Zur Theorie, Empierie und Praxis des Deutschen als Fremdsprache (und anderer Fremdsprachen)*. (pp. 155-175). Baltmannsweiler.

Pienemann, M. (1987). Determining the influence of instruction on L2 speech processing, *Australian Review of Applied Linguistics, 10(2),* 83-113.

Pienemann, M. (1989). Is language teachable? Psycholinguistic experiments and hypothesis, *Applied Linguistics, 10/1*, 52-79.

Tschirner, E. (1999). Lernergrammatiken und Grammatikprogression. En B. Skibitzki & B. Wotjak (Eds.) *Linguistik für Deutsch als Fremdsprache. Festschrift für Gerhard Helbig* (pp. 227-240). Tübingen: Niemeyer.

Vainikaa, A. & Young-Scholten, M. (1994). Direct access to X'-theory: evidence from Korean and Turkish adults learning German. En T. Hoekstra & B. Schwartz (Eds.), *Language acquisition studies in generative grammar*. Amsterdam: Benjamins.

Westmoreland, R. (1983). *L2 German acquisition by instructed adults.* MS, Universidad de Hawaii.

La lengua y la cultura del vino en la enseñanza de lenguas extranjeras

María José Labrador-Piquer
Universitat Politècnica de València

Pascuala Morote Magán
Universitat de València

Resumen

Aunque hay mucha bibliografía en torno al lenguaje relacionado con el vino y a su elaboración desde distintos enfoques (Lehrer, A., Hommerberg, C., Chateau, C…) en la adquisición y enseñanza de la lengua partiendo de la cultura y el léxico del vino, los estudios son escasos. La innovación de este trabajo radica en la posibilidad de ser aplicado en las Facultades y Escuelas de Viticultura y Enología extranjeras, en las que además, los estudiantes necesitan dominar una lengua extranjera.

Debido a que la lengua y la cultura van íntimamente unidas, nuestro trabajo va a versar sobre la fusión entre ellas. En una primera fase se ha recopilado léxico, expresiones, dichos, refranes, canciones, etc. en torno al vino a partir de textos literarios escritos u orales; en una segunda, en la que estamos trabajando en la actualidad, el corpus se va a centrar en la parte cultural que abarca la historia, la geografía, el arte, la música y la literatura (popular y de autor). En este trabajo de investigación presentamos una selección de muestras de este corpus. Se destaca su aplicabilidad didáctica, ya que nos sirve de herramienta en las aulas para el aprendizaje tanto de la cultura como de la lengua, dentro del marco de la enseñanza-aprendizaje de segundas lenguas.

Palabras clave: cultura, vino, aplicación didáctica

Summary

Although there is a lot of literature on language related to wine and its elaboration from different approaches (Lehrer, A., Hommerberg, C., Chateau, C…), the studies about language acquisition and teaching rooted on wine culture and lexicon are scarce. The innovation of this work stems from the possibility to apply this focus to the foreign Faculties and Schools of Viticulture and Oenology, in which students also need to have a good command of a foreign language.

Our work will deal with the close link between language and culture. First of all, lexicon, expressions, proverbs, songs, etc. related to wine have been compiled from written and oral texts; the second aspect, which is the focus of our current research, is a corpus centred around the cultural elements, which include history, geography, art, music and literature (popular and auteur). In this research paper we will present a selection of samples from this corpus. We will emphasize its didactic applicability as a useful classroom tool for the learning of culture as well as language within the second language teaching-learning framework.

Key words: culture, wine, didactic application.

1. Introducción

Diversos autores ofrecen estudios con distintos enfoques en torno al lenguaje del vino y a su elaboración; el libro *The language of wine: an english coursebook* (2008) aproxima al lector, a los amantes del vino o a aquellos que deseen aprender o mejorar el dominio de la lengua inglesa, a unos conocimientos que le van a facilitar el aprendizaje del idioma inglés en aspectos relacionados con el vino. Adrianne Lehrer (2010) de la Universidad de Arizona estudia los aspectos novedosos en el lenguaje del vino, tales como la creatividad, las palabras usadas en marketing, las etiquetas, etc.; Charlotte Hommerberg, (2010) de la Universidad de Linnaeus analiza la argumentación en los textos sobre vinos; Carmela Chateau (2010), por su parte, estudia la enseñanza del lenguaje del vino a un grupo de alumnos de nivel B1-B2 que se preparan para ser expertos en vinos. Los trabajos mencionados anteriormente y muchos otros tratan el tema del vino desde diferentes aspectos si bien en la adquisición y enseñanza de la lengua partiendo de la cultura y el léxico del vino, los estudios son escasos.

Entendemos que la lengua y la cultura van íntimamente unidas, esencialmente cuando se trata de aprender una lengua extranjera, por ello, nuestro trabajo versa sobre su fusión, razón por la que destacamos el concepto de cultura de Marvin Harris (2000:17): «Mi postura personal es que una cultura es el modo socialmente aprendido de vida que se encuentra en las sociedades humanas y que abarca todos los aspectos de la vida social incluidos el pensamiento y el comportamiento».

Martín Sánchez (2002:14) señala: «La vida del hombre está llena de matices y circunstancias que, de obviarse, la harían en gran parte incoherente y casi siempre incomprensible, pues si la historia de la humanidad es la suma de todas las historias de las diferentes culturas que han surgido y desaparecido a lo largo del tiempo, a su vez, todas ellas son el resultado de las historias individuales de las gentes que vivieron en esas culturas».

La cultura (geografía, arte, literatura, medicina gastronomía, fiesta…) es un medio excelente para trabajar las lenguas, en nuestro caso ELE; cuando hablamos de *lengua*, hablamos de *cultura*, noción que no podemos relegar; quien aprende una lengua, ha de aprender con ella civilización y costumbres, en suma, *cultura* concebida como un todo que abarca aspectos histórico-artísticos, literarios, antropológicos, folclóricos… a través de los cuales se asemejan o se diferencian unos pueblos de otros. La cultura es también un fenómeno dinámico de comunicación. Si los estudiantes desconocen la cultura de la legua objeto de aprendizaje, sus capacidades compresivas y expresivas, posiblemente quedarán mermadas y escasamente desarrolladas y como consecuencia, no alcanzarán las competencias que les corresponden.

El Plan Curricular del Instituto Cervantes (2006), plantea tres dimensiones o perspectivas del alumno complementarias e interdependientes entre sí. Como agente social, como hablante intercultural y como aprendiente autónomo. La dimensión de hablante intercultural, según el Plan, debe capacitarlo para identificar los aspectos relevantes de la nueva cultura a la que accede a través de la lengua y desarrollar la sensibilidad necesaria para establecer puentes entre la cultura de origen y la nueva (2006:33, volumen I).

2. Objetivos

Al investigar el tema del vino, pretendemos ampliar la creación de un corpus intercultural, ya en iniciado en nuestros trabajos anteriores en el que cabe todo (el docente verá cuándo, y cómo lo utiliza y qué competencias quiere desarrollar en cada momento) a fin de destacar en nuestras clases una especial sensibilidad hacia la Cultura Universal, a la que contribuyen las de otros países, pueblos y épocas hasta la actualidad.

Como objetivos nos hemos planteado seleccionar nuevos campos culturales; literatura, pintura, música, escultura, geografía, etc., aplicarlos0 a la enseñanza-aprendizaje de segundas lenguas mediante técnicas y metodologías activas y posteriormente estudiar la motivación del alumnado.

La cultura tendrá de esta forma un carácter globalizador e integrador que motivará a los estudiantes, en especial a los de Viticultura y Enología al manejo de la lengua y al disfrute del sentido estético de cualquier creación artística, para después dominar las competencias esenciales: comprender y reconocer las convenciones específicas de los discursos, la historicidad que hay en ellos y los saberes necesarios para poder construirlos, interpretarlos y evaluarlos.

3. Vino, cultura y literatura

La cultura del vino es antiquísima; según Ángel R. del Valle (2003: 49) parece que el paso de la vid como planta silvestre a su cultivo y vinificación tuvo lugar en el Neolítico tardío en áreas de Asia Menor (Anatolia o Armenia) y desde allí se extendió por el Mediterráneo gracias al comercio de fenicios, griegos y romanos durante toda la antigüedad. Con el cristianismo, su presencia en la liturgia y la conversión de Roma, se extendió por toda Europa y en la Edad Media los monjes se convertirían en los auténticos difusores de su cultivo.

La invasión de los árabes supuso una auténtica catástrofe para los viñedos españoles. La mayoría de ellos fueron devastados o arrancados porque en el Corán se prohíbe el uso del vino. Sin embargo, la Reconquista puso las cosas en su sitio y trajo para España el renacimiento del viñedo.

Muchos escritores antiguos y modernos hacen referencia al vino, como hemos señalado en estudios anteriores, tanto en la civilización griega y romana como en la literatura española desde la Edad Media a nuestros días. También se conocen refranes, canciones leyendas en torno al origen del vino como la de Noé o la de Babilonia. Esta última, cuenta que el rey persa Dsemsit almacenó uvas en un sótano de su palacio para poderlas comer en cualquier época, pero al cabo de cierto tiempo, fermentaron y desprendieron anhídrido carbónico, intoxicando a los que las cuidaban. Esto les llevó a pensar que las uvas eran venenosas. Una de las concubinas del rey intentó suicidarse por sentirse despreciada por el soberano; bebió de este zumo envenenado y al contrario de lo que se suponía, en vez de desear la muerte, se sintió muy feliz y contenta. Al presentarse así de radiante y alegre, el rey la prefirió a todas las demás. Desde entonces se dice que esta mujer fue la descubridora de las bondades del vino.

Del poeta persa Omar Jhayyam (siglo XI) hemos hallado la siguiente alabanza al vino, llena de expresividad e interrogaciones retóricas para exaltar sus valores:

> « ¿Por qué vendes tu vino, mercader?
>
> ¿ qué pueden ofrecerte a cambio de tu vino?
>
> ¿ dinero… ? y ¿qué puede darte el dinero…?
>
> ¿ poder… ? ¿pues no eres dueño del mundo cuando tienes en tus manos una copa?
>
> ¿ riqueza… ? ¿hay alguien más rico que tu, que en tu copa tienes oro, rubíes, perlas y sueños?
>
> ¿amor…? ¿no sientes arder la sangre en tus venas cuando la copa besa tus labios…? ¿no son los besos del vino tan dulces como los ardorosos de la hurí…? Pues si todo lo tienes en el vino, dime, mercader, porqué lo vendes.
>
> Poeta, porque haciendo llegar a todos mi vino doy poder, riqueza, sueños, amor…Porque cuando estrechas en tus brazos a la amada me recuerdas, porque cuando quieres desear felicidad al amigo, levantas tu copa, porque Dios cuando bendijo el agua la transformó en vino y porque cuando bendijo el vino lo convirtió en sangre…
>
> Si te ofrezco mi vino… ¡poeta…! no me llames mercader».

Y si nos trasladamos a la Literatura Española, es obligado citar a Cervantes: solo desde *El ingenioso hidalgo D. Quijote de la Mancha* se pueden extraer referencias textuales, para trabajar en las clases de ELE competencias léxico-sintácticas y retóricas, como por ejemplo en el siguiente diálogo con Sancho Panza:

> «¿No será bueno, señor escudero, que tenga yo un instinto tan grande y tan natural, en esto de conocer vinos, que, en dándome a oler cualquiera, acierto la patria ,el linaje, el sabor y la dura, y las vueltas que ha de dar, con todas las circunstancias al vino atañederas? Pero no hay de qué maravillarse, si tuve en mi linaje por parte de mi padre los dos más excelentes mojones que en luengos años conoció la Mancha, para prueba de lo cual les sucedió lo que ahora diré:
>
> Diéronles a los dos a probar del vino una cuba, pidiéndoles su parecer del estado, cualidad, bondad o malicia del vino. El uno lo probó con la punta de la lengua, el otro no hizo más de llegarlo a las narices. El primero dijo que aquel vino sabía a hierro, el segundo que más sabía a cordobán. El dueño dijo que la cuba estaba limpia, y que el tal vino no tenía adobo alguno por donde hubiese tomado sabor de hierro ni de cordobán. Con todo eso los dos famosos mojones se afirmaron en lo que habían dicho. Anduvo el tiempo, vendióse el vino, y al limpiar la cuba hallaron en ella una llave pendiente de una correa de cordobán ».
>
> (Capítulo XIII, 2ª parte)

Infinidad de dichos, refranes, máximas, relatos, coplas y letras de juegos infantiles están plagadas de referencias al vino y a la vid que se relacionan con el campo semántico de la vitalidad humana como alegría, amor, victoria, evasión, diversión…

4. Desarrollo del trabajo

Trabajar el tema del vino partiendo de textos de escritores españoles, clásicos y modernos, ha sido objeto de la primera fase de nuestro corpus; en este estudio nos centramos en la

parte cultural que abarca la pintura, la música, la escultura, la geografía y la literatura (popular y de autor), ya que nos sirve de herramienta en las aulas para la enseñanza-aprendizaje tanto de la cultura como de la lengua, y de base de una investigación posterior sobre la motivación del alumnado, dentro del marco de un máster de Viticultura.

4.1 Metodología

En primer lugar se ha buscado en diversas fuentes ejemplos representativos de los diferentes campos; posteriormente, se han enriquecido con información pragmática, con el estudio del componente cultural de las imágenes, etc.; en segundo lugar, se ha creado una ficha técnica de cada una de las obras; en tercer, lugar, se ha establecido el vocabulario necesario (nivel B1) para el estudio de dichos campos, prestando especial atención a su uso actualizado y por último, se han diseñado unidades didácticas basadas en metodologías activas para la actividad en el aula.

El vino y sus símbolos han inspirado a lo largo de la historia a innumerables artistas que han plasmado en la pintura, la escultura y la música sus obras. Como un primer ejemplo nos detenemos en tres lienzos del Museo del Prado que coinciden con el tema objeto de nuestro estudio: la *Bacanal* de Tiziano, la *Bacanal* de Poussin y *Los borrachos* de Velázquez y mostramos un ejemplo de ficha técnica (la *Bacanal* de Tiziano) creada para el campo de la pintura:

Autor. Tiziano Vecelllio (1488-1576)

Título: La Bacanal de los andrios

Fecha: 1523-1526.

Encargado por. Alfonso I D´Este para su palacio de Ferrara

Localización actual: Museo del Prado

Temática: escena mitológica. Fiesta en honor al dios Dionisos / Baco que representa un canto a los placeres de la vida. Este tema era habitual en la decoración de los palacios de la época.

Comentario de la obra: es un lienzo al óleo de pintura renacentista del Cinquencento con rasgos manieristas, estilo de pintura veneciana del siglo XVI

Son numerosas las obras donde encontramos temas sobre el vino, la vendimia, las uvas, etc. como *La vendimia o el otoño* de Goya; *Niños comiendo melón y uvas* de Murillo; *Comida de los remeros* de Renoir o el último descubrimiento artístico del artista flamenco Pieter Bruegel el Viejo *El vino en la fiesta de San Martín*. Pero en temas de pintura el vino llega aún más lejos convirtiéndose en técnica pictórica, bien utilizada directamente de la botella o bien hervido y unido a aglutinantes orgánicos, de este modo como afirma Casanova Sorolla «el vino tiene un final largo en boca, pero eterno en el papel». De la misma forma se podrían trabajar famosas esculturas como el *Baco* de Miguel Ángel ubicado en el Palacio Bargello de Florencia

Si vinculación hay del vino con la pintura y la escultura, también la hay con la música. En las siguientes óperas se encuentran fragmentos de brindis o escenas en las que aparece algo relacionado con el vino:

L´elisir d´amore de Donizetti; escena en la que Nemorino cree que bebe un licor y Dulcamara explica al público que es vino.

Lucrecia Borgia de Donizetti; Lucrecia envenena a los que no le gustan con vino (Varias escenas).

En Marina de Arrieta; la canción del brindis "A beber, a beber y a apurar/las copas del licor/que el vino hará olvidar las penas del amor."

Don Giovanni de Mozart, aparecen varias escenas relacionadas con el tema del Don Juan y la cena macabra en la que se invita a cenar y a beber a los difuntos que él mismo había asesinado. Tema cultural de la literatura occidental utilizado entre otros por Molière, Tirso de Molina, Zorrilla…

Otello de Verdi, basado en la obra teatral de Shakespeare del mismo nombre; hay una escena en a que se emborrachan con vino. En *La Traviatta* de Verdi, el famoso brindis en el no se especifica qué beben, si champagne o vino. También se trata el vino en *Falstaff de Verdi*, *El murciélago* de J. Strauss, etc.

El oratorio profano de *Las Estaciones* de Haydn, en el libreto se encuentra un brindis de exaltación a la tierra que produce vino y a las personas que lo trabajan:

[…] Alabemos el producto de la uva.
El amigo de la vejez.
El remedio para las cuitas y el amor.
Alabemos en voz alta de alegría
el liquido generoso
y digamos: viva el vino.
Viva la gente que lo cultiva.
Y viva la tierra donde se cría.

Asimismo, la música ligera recoge infinidad de canciones referentes al vino como *Al pan, pan y al vino, vino* de Chayanne (refrán tradicional), *Camarero champagne* de Luis Aguilé, *Copa de vino* de Lola Flores, *Copa rota* de Los Rodríguez, *Días de vino y rosas* de Revólver, *El vino y el pescao* del G5 (Kiko Veneno), *Fiesta y vino* de Duncan Dhu, *Hasta el vino de la copa* de Juanito Valderrama, *Mujeres y vino* de Manolo Escobar, *Sin vino no se anda el camino* de Paco Bello (título basado en el refrán *con pan y vino se anda el camino*), *Soy un truhán soy un señor* de Julio Iglesias, *Una guitarra y un vaso de vino* de Paul Anka, etc.

5. Actividades didácticas

Entre las múltiples técnicas y métodos cooperativos que se pueden utilizar hemos seleccionado a modo de ejemplo la técnica del puzle o rompecabezas para trabajar la

escultura del dios Baco de Miguel Ángel. Una vez formados los equipos de trabajo se realizan las siguientes tareas que se asignan a un miembro de cada uno de los grupos:

- Descripción de la obra
- Biografía del artista y entorno artístico del escultor
- Obra escultórica y pictórica
- Ubicación de sus obras

Una vez reunidos los grupos de expertos y terminado el informe correspondiente, cada experto lo expone a sus compañeros de equipo. Como tarea final se propone inventar una historia sobre lo que les sugiera esta escultura que puede ser narrada y leída, ya que el uso creativo de la lengua conmueve, seduce y despierta el interés por la cultura.

6. Conclusiones

L a cultura es un con junto de saberes de las personas que les sirve para conocer mejor el mundo que les rodea.Trabajar el tema del vino partiendo de la pintura, la música, la escultura, canciones populares, etc. implica conocimiento y motiva el aprendizaje de la lengua y la cultura

Trabajar en las aulas con géneros literarios de tipo tradicional sumerge al estudiante en la filosofía del habla y en el sentido del humor relacionado con el contexto histórico geográfico y las costumbres

Utilizar técnicas activas habladas y escritas basadas en las Bellas Artes supone el fomento de la investigación-acción y el desarrollo de las competencias o destrezas que vinculan la lengua a la cultura

El vino y la vid como parte de la cultura universal, mediterránea y española es una buena herramienta para que el alumnado aprenda refranes y dichos sobre la bebida, con la finalidad de intercalarlos en sus conversaciones habituales, de ahí su valor intercultural, pragmático y didáctico.

Bibliografía

Casanova, L. (2009) Disponible en
http://www.viagourmet.com/noticias/gourmet/luis-casanova-sorolla-pintar-con-vinos.html

Chateau, C. (2010).“English as the Language of Wine-tasting in Burgundy? Corpus Evidence for Collocational paradigms in English and French”. En *Proceedings of the First International Workshop on Linguistic Approaches to Food and Wine Description* Margarita Goded y Alfredo Poves (Coords.). Madrid: UNED, Colección Arte y Humanidades, núm 002.

Goded Rambau , M. y Valera Menéndez, R. (2008): *The language of wine: an English coursebook*. Madrid: Ediciones Académicas.

HARRIS, M. (2000). *Teorías sobre la cultura en la época posmoderna*. Barcelona: Crítica.

HOMMERBERG, C. (2010).“ Argumentation in wine written” en *Proceedings of the First International Workshop on Linguistic Approaches to Food and Wine Description* Margarita Goded y Alfredo Poves (Coords.). Madrid: UNED, Colección Arte y Humanidades, núm 002.

INSTITUTO CERVANTES, PLAN CURRICULAR (2006). Madrid: Biblioteca Nueva.

LABRADOR PIQUER, M.J., Y MOROTE, P. (2010). “El vino y la literatura” en *Proceedings of the First International Workshop on Linguistic Approaches to Food and Wine Description* Margarita Goded y Alfredo Poves (Coords.). Madrid:UNED, Colección Arte y Humanidades, núm 002.

LABRADOR PIQUER, M.J. (2006): “La expresión oral y escrita en la cultura del vino y los toros”. En *Actas del XLI Congreso Internacional de la Asociación Europea de Profesores de Español. 125 años del nacimiento de Picasso en Málaga*. Málaga: Asociación Europea de Profesores de Español y Universidad de Málaga.

LEHRER, A. (2010). “What’s new in wine language” en *Proceedings of the First International Workshop on Linguistic Approaches to Food and Wine Description* Margarita Goded y Alfredo Poves (Coords.). Madrid: UNED, Colección Arte y Humanidades, núm 002.

LOTMAN J, Y USPENKIJ, B. A. (1979). “Sobre el mecanismo semiótico de la cultura” en *Semiótica de la Cultura*. Barcelona: Cátedra.

MARTÍN SÁNCHEZ, M. (2002). *Seres míticos y personajes fantásticos españoles*. Madrid: Edaf.

MOROTE MAGÁN, P. (2006). “Vino y toros. Mito y poesía”. En *Actas del XLI Congreso Internacional de la Asociación Europea de Profesores de Español. 125 años del nacimiento de Picasso en Málaga*. Málaga: Asociación Europea de Profesores de Español y Universidad de Málaga.

MOROTE MAGÁN, P.Y LABRADOR PIQUER, M.J. (2010): “El vino, nexo intercultural”. En *Akten of the 9th International Conference of the European association of Languages for Specific Purposes*. Hamburg: University of Hamburg, Germany.

Error coding in the TREACLE project[184]

PENNY MACDONALD *(Universitat Politècnica de València),* SUSANA MURCIA *(Universidad Autónoma de Madrid)*, MARÍA BOQUERA *(Universitat Politècnica de València),* ANA BOTELLA *(Universitat Politècnica de València)*, LAURA CARDONA *(Universitat Politècnica de València),* REBECA GARCÍA *(Universidad Autónoma de Madrid)*, ESTHER MEDIERO *(Universidad Autónoma de Madrid)*, MICHAEL O'DONNELL *(Universidad Autónoma de Madrid)*, AINHOA ROBLES *(Universidad Autónoma de Madrid)*, KEITH STUART *(Universitat Politècnica de València)*

Abstract

This paper presents the approach to error analysis within the TREACLE project, the aim of which is to profile learner proficiency to help inform teaching curriculum design. We will introduce the error annotation methodology used on a corpus of texts written by Spanish learners of English at University level. After a short introduction concerning Computer Learner Corpora and error analysis, we will discuss the underlying principles of the error coding scheme and then provide more details about the coding scheme itself. To ensure coders are annotating the texts in the same way, two steps were followed. Firstly, we developed a comprehensive coding criteria description giving full details as to how to code particular instances. Secondly, we performed two inter-coder reliability studies to help us identify areas where coders were differing, so that we could address these areas. We will present the preliminary results of the error analysis and discuss the repercussions of these results for grammar teaching.

Keywords: learner corpora, error analysis, English as a Foreign Language

Resumen

Este artículo presenta el enfoque del análisis de errores del proyecto TREACLE, cuyo objetivo es proporcionar perfiles de competencia de aprendices con el objeto de poder informar el diseño curricular. Se presente la metodología utilizada en la anotación de errores de un corpus de aprendices escrito por estudiantes españoles de inglés a nivel universitario. Después de una breve introducción a los corpus computerizados de aprendices y al análisis de errores, se relacionan los principios subyacentes al esquema de codificación de errores y se proporcionan más detalles sobre el esquema de codificación mismo. Para asegurar que los codificadores anotan los textos de la misma manera, se siguieron dos pasos. En primer lugar, se realizó una amplia descripción de los criterios de codificación que contiene detalles sobre cómo codificar ciertos casos. En segundo lugar, realizamos dos estudios de fiabilidad inter-codificadores que nos permitieran identificar aquellas áreas en las que los codificadores discrepaban, de modo que pudiéramos aclararlas. Se presentan los resultados preliminares del análisis de errores y se discuten las posibles repercusiones de estos resultados para la enseñanza de la gramática.

Palabras clave: corpus de aprendices, análisis de errores, Inglés como Lengua Extranjera

184 The TREACLE project is funded by the Spanish Ministry of Science and Innovation (research grant: FFI2009-14436/FILO)

1. Introduction

The development of Computer Learner Corpora (CLC) in the early 1980s prompted the creation of a new and lasting relationship which has served to bridge the gap between the field of Corpus Linguistics and foreign language learning research and pedagogy. One of the first CLC projects was the Danish PIF - Project in Foreign Language Pedagogy (Faerch et al., 1984). In the early 1990s, The International Corpus of Learner English (ICLE), founded and coordinated by Sylvianne Granger at the Université Catholique de Louvain in Belgium (Granger, 1993; 1998) was developed and has become the most cited in the literature. However, since then, the development of new learner corpora has increased tenfold as can be observed on the 'Learner Corpora around the World' web page created by the Centre for Corpus Linguistics at Louvain.

Learner corpora, which tend to be much smaller than native corpora, have been used in the literature to provide evidence concerning what language learners have acquired and not acquired, sometimes comparing the output with native speakers (NSs), and other times comparing it with other mother tongue groups. Granger et al. (2002) identify two main ways in which corpora have been studied from a linguistic-based methodological perspective: Contrastive Interlanguage Analysis and Computer-aided Error Analysis. The first method deals with those studies that compare the output of either two or more non-native speaker (NNS) groups or two or more NS and NNS groups. Computer-aided Error Analysis involves the study of learner output (i) in order to detect the difficulties that learners have, which in turn contributes to finding out more about the processes involved in learning, (ii) for identifying the instances of crosslinguistic influence, (iii) for the development of new materials, etc.

Despite the affirmation of one of the most prestigious researchers in the field of Error Analysis, Carl James, that Contrastive and Error Analysis (EA) "are still going strong" (1994:179), there has not been much evidence of this in the literature in recent years. Indeed, since the heyday of Error Analysis in the 60s and 70s, there have been fewer and fewer studies dedicated to the analysis of errors in learner output in foreign language learning environments, which, as Leech (1998: xvii) points out, may be due to the fact that "...the negative attitudes to Error Analysis inherited from that period have coloured many people's thinking ever since".

Yet on a day to day basis, teachers are doing it continually. And in the research on second language acquisition and teaching methodologies, studying learner output, both correct and incorrect forms, is still central to the agenda although the shift has been more towards finding ways of being successful with corrective feedback in order to improve the learners' linguistic and communicative competence. This may involve teacher-centred feedback; peer feedback, or computer-mediated feedback in the form of grammar checkers, online tutorials, and so on.

From a different point of view, although, with pedagogical goals as the main focus, there are a number of more recent studies which are analysing errors in order to create specific profiles of learner competence (Capel, 2010; Granger and Thewissen, 2005a; 2005b). In the case of the English Profile Project (Hawkins & Buttery, 2009) the aim is to develop

Reference Level Descriptions for English linked to the Common European Framework of Reference for Languages (CEFR).

The TREACLE project (O' Donnell et al., 2009),[185] is also concerned with the creation of a methodology for producing grammatical profiles of Spanish university students' written English language, linking them to the CEFR levels, with the aim of redesigning the English Language grammar curriculum to improve its efficiency and projection.

The project is described in detail in O'Donnell in the present volume, and therefore the purpose of this paper is to present only those aspects of the project related to error-annotation, discussing its methodology and some of the preliminary results, as well as their possible repercussions on the teaching and learning of English in the Spanish higher education context.

2. *The TREACLE project and the corpus for error annotation*

The TREACLE[186] project (O' Donnell et al., 2009) involves the development of an annotated corpus of learner English for pedagogical application. One of the aims of the project is to carry out a computer-aided error analysis on the corpus to find out what students tend to get wrong at each level of proficiency.

This project uses corpora from two Spanish universities: Universidad Autónoma de Madrid (UAM) and Universidad Politécnica de Valencia (UPV). Since these universities have different degree courses on offer, the written output of these students can be said to represent the kind of English as a Foreign Language which learners in both Humanities and Technically-oriented universities produce in Spain.

The two corpora we are using for our study –the Wricle Corpus (UAM) and the UPV Learner Corpus have been developed following strict design criteria.

The UPV Learner Corpus is part of the MiLC corpus[187] (Andreu et al., 2010). The UPV Learner Corpus consists of 950 written compositions (180,000 words) from Spanish students of all levels, mostly centring on the topic of Immigration.

The WriCLE Corpus (Rollinson and Mendikoetxea, 2010), consists of 750 essays written by Spanish learners of English. For the TREACLE project, we use 521 of these essays (about 500,000 words) written by Spanish students of all levels of proficiency. The essays deal with the topics of immigration, homosexual marriages and traffic problems.

In both cases, the metadata is carefully recorded (i.e. details of sex, age, year of study, mother tongue and other languages spoken/learnt, etc.) and all students gave their permission for the texts to be used for research purposes and did the Oxford Quick Placement Test (UCLES 2001) close to the time of writing in order to identify their levels according to the CEFR levels.

185 More information on this project can be found at http://www.uam.es /treacle/index.html

186 ***TREACLE*** *stands for* ***T****eaching* ***R****esource* ***E****xtraction from an* ***A****nnotated* ***C****orpus of* ***L****earner* ***E****nglish.*

187 MiLC is a multilingual learner corpus involving the written work (formal and informal letters, summaries, essays, reports, translations, simulations, computer-mediated communication, etc.) of students learning English, Spanish, French, and German as a foreign language, and also Catalan, as a first, second or foreign language.

The length of the texts varies – the UPV compositions tend to be much shorter (between 200 and 250 words) as the levels, in general, are lower. In the case of the WriCLE essays, these are mostly 1,000 word essays produced by students of English Philology. In this way we have a collection of essays that represent the different levels from A1 through to C2.

3. *Methodology*

The error annotation of the corpus is carried out manually with the *UAM CorpusTool* (O'Donnell, 2008), which uses an error coding scheme devised by the researchers for this particular purpose and which can be modified as needed.

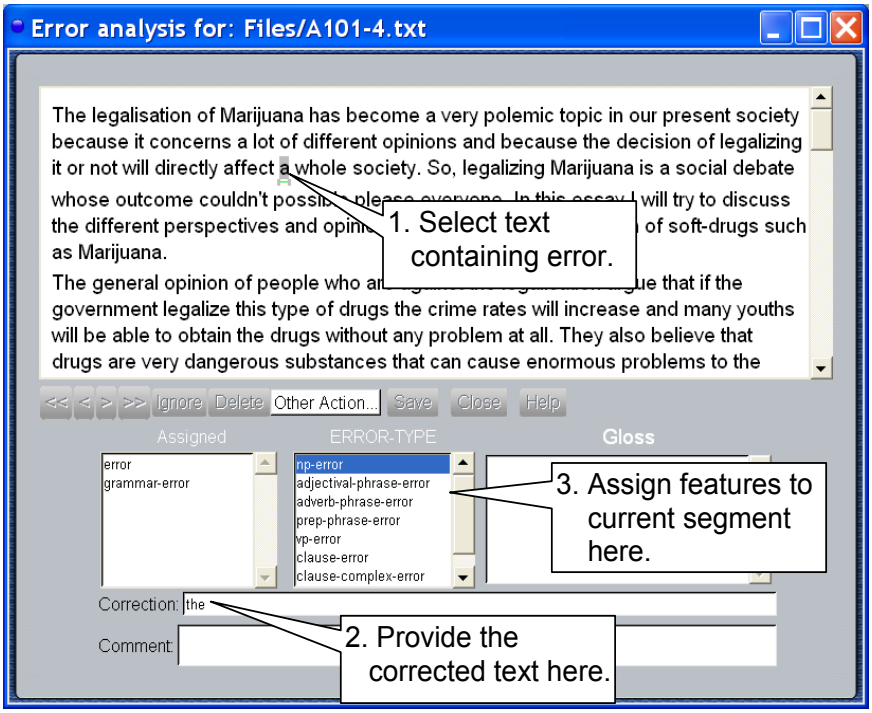


Figure 1. Error coding a text

As shown on the error annotation window in Figure 1, the UAM CorpusTool allows the coder to select the text of the error (step 1 in the diagram), provide the correction for that error (step 2), and assign error codes to it (step 3). The system is provided with a hierarchically organised set of error codes (the coding scheme) which the user walks through to assign a code, e.g., selecting first "grammar-error", then "np-error", then "determiner-error" and then "determiner-choice-error". This process of gradual refinement of error codes facilitates the coding process, because often the coder does not know what leaf of the error tree they should code, but can make a series of decisions on more general grounds (e.g., Is it a grammatical or lexical error?).

To facilitate the coder's job, the error scheme incorporates coding criteria ('glosses') with each feature in the scheme, which are displayed in the coding window. These glosses are also included in a lengthy Coding Criteria Manual (20 pages long) which provides clear guidelines for determining which of the categories is appropriate for a given error.

3.1. *The Error Scheme*

The TREACLE error scheme has been designed from the start to integrate into a University level language teaching programme. So far, the coding scheme contains 113 different errors at the most delicate level in the hierarchy. Figure 2 shows the more general categories, as the whole scheme does not fit here.[188]

The main design principle has been to ensure that the error scheme should map cleanly onto the organization of grammar topics which are taught within EFL courses. The main reason for this is that our goals are pedagogical, and we later want to be able to recover the errors relevant to each grammatical structure, so as to inform our teaching of that topic.[189]

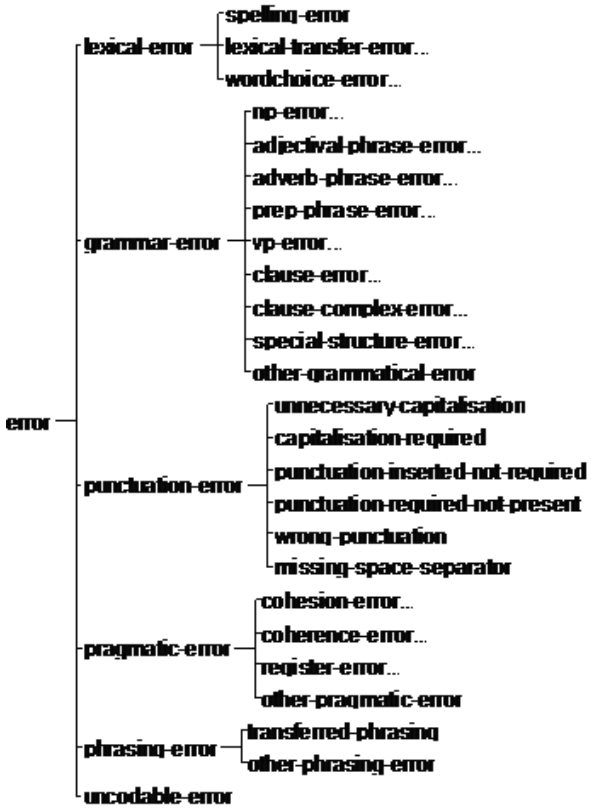


Figure 2. The coding scheme

188 The three dots after an error category indicate that sub-categorization has been omitted.

189 Our scheme, however, includes errors other than grammatical ones, as shown in Figure 2.

To demonstrate our approach, we will focus on one particular part of the error scheme. Let us assume the student has written "*this results are…". At the root of the error hierarchy, we distinguish between various types of error, including lexical errors, grammatical errors, pragmatic errors, etc. (see Figure 2).

Assuming the error is grammatical, one next chooses the type of grammatical error. While some error coding systems are more oriented to coding errors in terms of the part of speech of the word concerned, our approach is more focused on the grammatical phrase in which that word occurs. Thus, while teaching about adjectival phrases, we can find errors within adjectival phrases, whether they involve the adjective itself, or any adverbial premodifier of the adjective, e.g. "very browner".

Continuing with our example, assume the error is within the selection of the determiner. We thus select *np-error*. This leads to the next level of delicacy, as shown in Figure 3. Our division of error codes within the NP reflects the fact that, in many courses, the teaching of the Noun Phrase is divided into topics: determiners, pre-modifiers, the head, and post-modifiers.

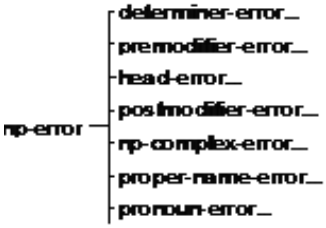


Figure 3. NP Error sub-classes

Figure 4. Subtypes of determiner errors

Selecting determiner-error, we are presented with the next set of choices, as shown in Figure 4. Note that the error codes shown here are not exhaustive; the coder can add new error codes as examples are encountered in the learner texts. The particular example we started with, "*this* results are…", would be coded as *determiner-agreement*.

Our error coding is still at an early stage. So far, we have coded 128 texts containing 57,000 words, with 7,428 errors. However, this trial period has established that the error scheme and annotation process is viable, and that it caters for almost all of the errors encountered. We intend to double the number of errors coded by the end of 2011.

To ensure that the coding scheme and the coding criteria manual were valid instruments for our coding purposes and that they can render reliable coding among the researchers taking part in the project, we carried out two inter-coder reliability studies which will be explained in the following section.

4. Intercoder reliability study

According to Polio (1997: 102), error analysis studies 'have rarely reported intra- and interrater reliabilities, which can call into question' the conclusions reached in the research results, and certainly 'make replication of a study' in a different context, somewhat problematic.

We performed two inter-coder reliability studies with the aim of:

1. Refining the error scheme and coding criteria document.
2. Ensuring all coders were complying with the coding criteria.

One of the first stages in the error analysis was for the coders to become accustomed to the programme and the different levels and types of error codes. Several texts were coded on an individual basis and meetings were then held to discuss either problematic phrases or structures, or doubts concerning how to actually code the errors that had been detected. The second stage in the inter-coder reliability study (ICRS) involved the coding by all researchers of several texts, followed by consensus meetings in order to detect any individual 'doubts' that needed to be clarified, to add on new codes, or to make any necessary changes to the error hierarchy.

The coders had to familiarize themselves with the guidelines concerning segmentation with the UAM Corpus Tool. The rule is that we use minimal segmentation, i.e. only select the amount of text that would be necessary in order to make the correction of the erroneous form. So, in the case of a frequent error such as: **in the other hand* we would only select the preposition '*in*'. The reason for this is that the automatic syntactic analysis[190]

will locate this word within the phrase on another level of analysis, and also, as the rest of the connector is correct, the learner has got most of it right and therefore it is not necessary to highlight the whole phrase. However, to begin with, it was found that several researchers tended to select more words than was necessary, since the error could be seen to be affecting the whole phrase.

We also code in regards to what the learner has written rather than what they should have written, which is also common practice in the ICLE learner corpus especially prioritising the identification of the 'grammatical' errors which is, after all, the *raison d'être* of the project, always keeping in mind the use of the different grammatical forms made by Spanish university students.

For the inter-coder study, rather than comparing each coder to each other coder (requiring 120 distinct comparisons), we derived a 'consensus model' (or 'golden standard'), this being the subset of all the error codings on which at least 51% of the coders agreed. We could then compare each individual coder to this consensus project to see how the individual coder differed from the consensus. A report was automatically generated (see figures 5 and 6 below), showing each identified error, how many participants coded it, which participants provided alternative codings of it, cases where the text correction differed, and so on. This document provided the basis of our discussion of the various differences in coding practice, and helped us move towards a consensus.

Reliability scores were calculated by comparing each individual's coding to the consensus model.

As regards consensus rates, some codes were easier to pinpoint by all coders, such as

190 Our scheme, however, includes errors other than grammatical ones, as shown in Figure 2.

those involving spelling, punctuation, article errors, verb tense errors, etc. On the other hand, at least in the initial stages, certain errors involving larger chunks of text, for instance, phrasing errors, proved to be more difficult to assign codes to in a unanimous way. The following examples (extracted from an essay on education in Spain) illustrate some of the issues that have interfered with achieving a higher score in the ICRS.

Example 1

'The education in Spain is a subject that given a lot of play because for one *people this system of education is great, but for other people is awful.'*

In Figure 5 below, it can be seen that R6 did not highlight the minimum segment (***one***) although the code and the correction coincide with the rest of the coders. Technically speaking, consensus was reached, although this is not reflected in the results.

ven a lot of play because for **one** people this system of educati

	Consensus:	grammar-error: np-error: determiner-error: determiner-choice-error	some
✓	R7		✓
✓	R1		✓
✓	R2		✓
✓	R5		✓
✗	R4	grammar-error: np-error: premodifier-error: incorrect-premodifier-category	✓
✗	R3	lexical-error: wordchoice-error: other-wordchoice-error	✓

ven a lot of play because for **one people** this system of education is g

✗	R6	grammar-error: np-error: determiner-error: determiner-choice-error	some people

Figure 5. Example 1 of ICRS

Example 2

There are other cases where the coders did not agree at first with the codes chosen for a certain error. However, after the consensus meetings, acceptable solutions were found, such as in the example below – the error was due to L1 transfer, and should be coded as *Lexical error – lexical transfer error – false friend*.

'there are a lot of players who have an important paper'.

players who have an important **paper**. In this aspect I have to men

	Consensus:	lexical-error: wordchoice-error: other-wordchoice-error: noun-vocab-error	role
✓	R4		✓
✓	R1		✓
✗	R7	phrasing-error: transferred-phrasing	✓
✗	R6	lexical-error: wordchoice-error: transferred-word: borrowing	✓
✗	R3	lexical-error: wordchoice-error: false-friend	✓
✗	R2	lexical-error: wordchoice-error	✓

Figure 6. Example 2 of ICRS

False friends groups together those words that have similar ancestors, such as is the case with cognates, but whose meanings (or some of their meanings) have diverged over time e.g. Spanish *éxito* and English *exit* or Spanish *remover* and English *remove*. At times there may only be partial semantic identity, as Odlin (1989:79) explains – as in Spanish *suceder*, and English *succeed*, which is the case of the word *paper* as the writer took for granted s/he could transfer the metaphorical meaning of Spanish *papel* (*role*) for the literal translation into English *paper*.

The Error Criteria Manual has undergone several changes as a result of the ICRS, and indeed, the coding tool is designed so as to expand on any of the branches already established and in this way, include a wider variety of choices which can pinpoint certain errors with greater accuracy.

5. *Results*

As we mentioned above, so far, a total of 128 texts (57,000 words) have been coded, and 7,428 errors have been identified. The coding done so far reveals some interesting results.

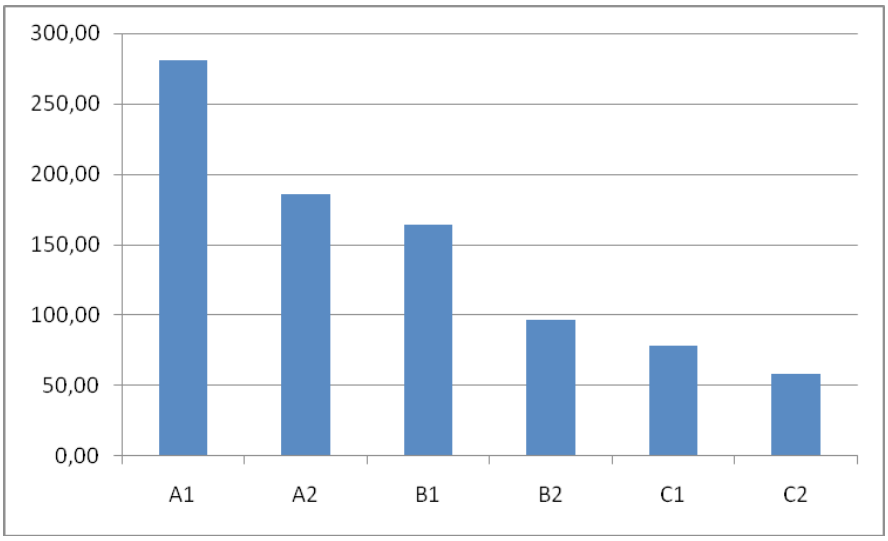


Figure 7. Number of errors per 1,000 words

Figure 7 shows that the number of errors decreases as the proficiency level increases. At the same time, these results provide evidence that the Oxford Quick Placement Test is a valid instrument for identifying students' proficiency level

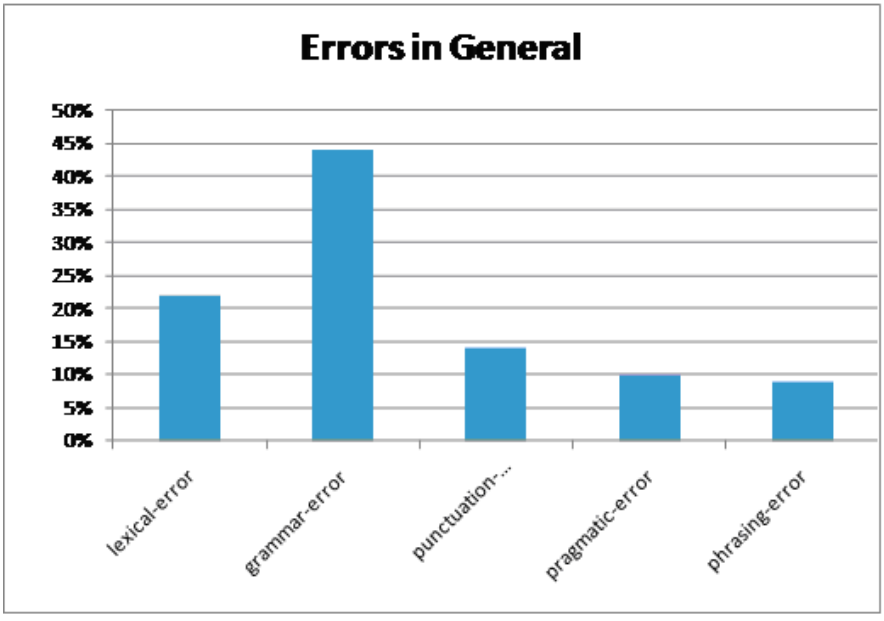


Figure 8. Results for general error categories

Figure 8 shows the percentage of errors in relation to all the errors made by the learners. Of the five main categories of errors, we can observe that grammar errors account for 44% of all errors.

The object of this study is to make profiles of what learners know and do not know (or do not use) according to the different proficiency levels and with particular reference to grammar. Figure 9 indicates how the degree of error occurrence varies according to the CEFR level. Thus, although grammar errors are still the highest, there are some notable differences within this group depending on the level. For example, it is the B1 Group which makes most grammar errors – nearly 50%, and this is 15% more than the C2s. This group, with the A1 group, also makes more errors as regards lexis. On the other hand, the C2 group makes the most errors as regards punctuation and what we have called 'pragmatic' errors. In the case of punctuation errors, the high rate among the C2 proficiency level may be due to the fact that the texts tend to be written with longer sentences, involving more subordination and linking devices, which in turn, require more punctuation. Also, we understand that when analysing the errors of the lower level groups, the coders may not actually be able to concentrate on finding the punctuation errors since first they have to make sense of, and code the other types of errors (grammar, lexical) present in the texts.

As regards these preliminary results, there are some levels of errors which do not seem to follow the expected pattern, for example, the unexplained drop in A2 as regards lexical errors, or the sharp rise in punctuation errors within the C2 level. This is most likely to be a sampling error, and will probably smooth out as more data is obtained.

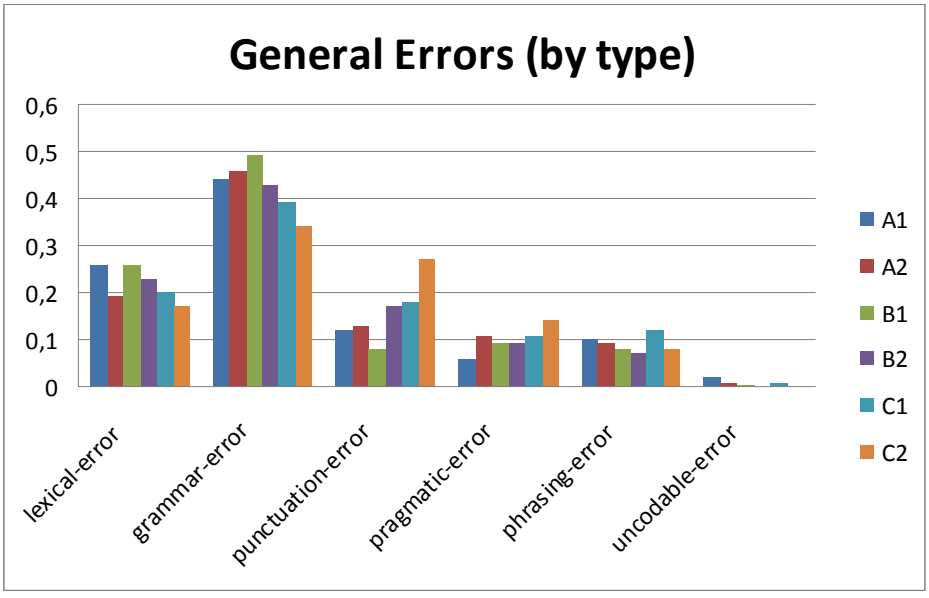


Figure 9. General errors produced at each CEFR level

Figure 10 looks at the distribution of errors within Grammar Error category. Around 40% of grammar errors are within the NP, which suggests that more emphasis in the teaching of NP syntax is needed. The fact that this percent falls in general as proficiency rises suggests that more attention to NPs is needed for the lower proficiency learners. Errors in prepositional phrases are also significant (around 20% of grammar errors), with most of these consisting of wrong selection of the preposition itself. This is another area where increased teaching emphasis is indicated.

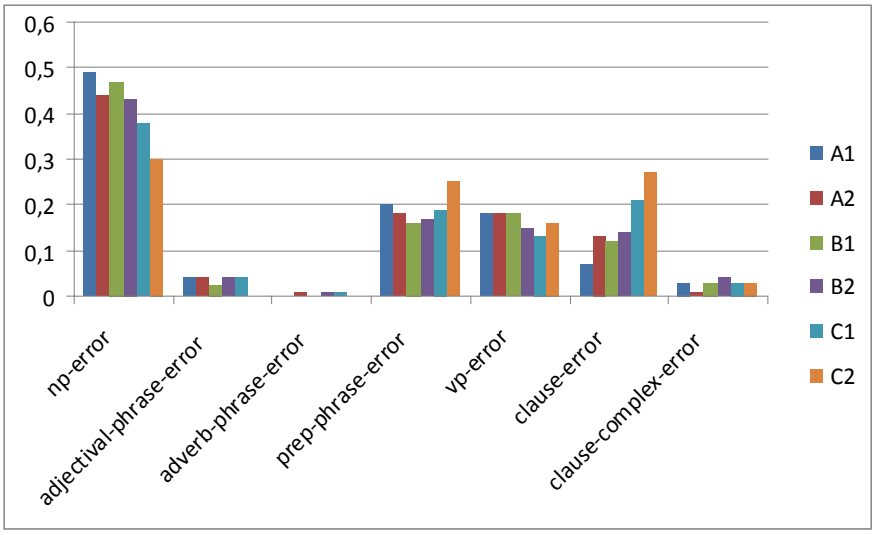


Figure 10. Grammar errors produced at each CEFR level

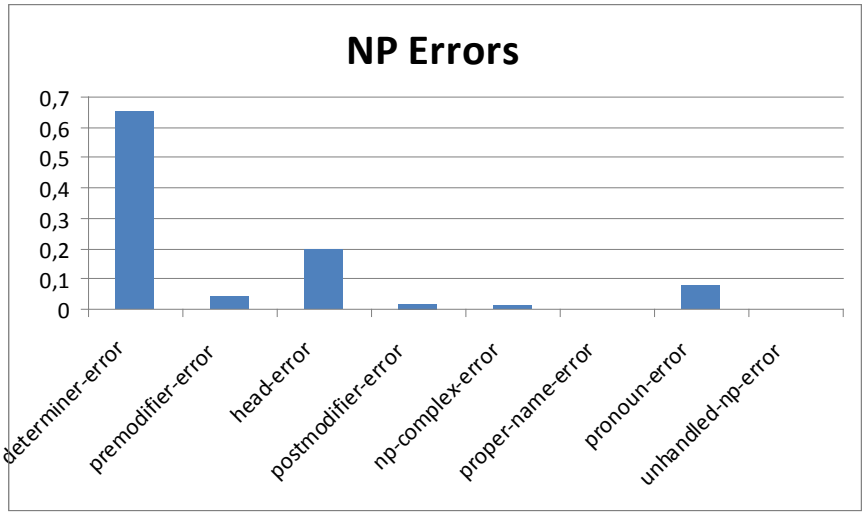


Figure 11. Results for NP error subclasses

In particular, within the noun phrase category (Figure 11), the most frequent errors by far were those related to determiner use (30% of grammar-related errors), and 75% of these involved the wrongful presence or absence of the determiner. The students' mother tongue syntactic structures may have a certain influence here. As Swan & Smith (1987: 83) point out, in Spanish, as regards one determiner, the definite article, this goes with mass nouns and plural count nouns when used with a general meaning, whereas in English this is not the case. Likewise, there are certain contexts in English (i.e. with single count nouns) where articles are needed, and are not required in Spanish e.g. **Do you have bicycle* (*¿Tienes bicicleta?*) or in the following case: **My sister is teacher* (*Mi hermana es profesora*).

In addition to helping with the process of error coding, the *UAM Corpus Tool* allows one to search for particular structures and phrases containing specific examples of errors in context.

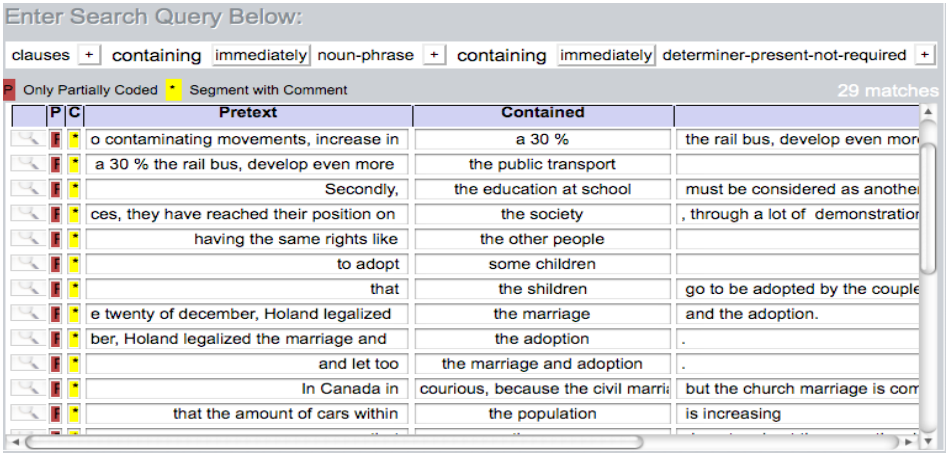


Figure 12. Corpus search of article errors

As figure 12 shows, the researcher wishes to find all NPs which contain an error of type *determiner-present-not-required*. Such searches can be used to recover examples which can be then be used for the teaching of particular topics.

6. Conclusions

From the study carried out so far in the TREACLE project, we can draw conclusions as regards the methodology used, the results obtained from the error coding, and the potential applications of studies such as this.

As regards methodology, the error-coding system we are using is viable and seems to provide codes for almost all of the error types we have encountered, since the main design principle underlying the project has been to ensure that the error scheme should relate

directly to the organisation of grammar topics which are taught within EFL courses at university level.

Our coding scheme differs from previous work in two aspects:

- It is oriented towards the pedagogical organisation of grammar, rather than being based on part of speech (e.g. the ICLE taxonomy of errors)
- As a result, we have a large number of codes, necessary to align the error-analysis with a fine-grained grammar syllabus.

The results from the error coding show that the number of errors decreases as students gain proficiency, although some of the results show trends that must be taken in a tentative way until we have more texts coded.

So far, the numbers and types of errors point towards potential improvement, or re-designing, of the teaching curriculum, and considering other teaching methods:

- The types of errors produced by students change at different levels of proficiency. The teaching curriculum, therefore, needs to adapt to the students' needs at each level.
- In Spanish university English language programmes, whether dedicated English Studies degrees or other more technically-oriented degree programmes, more emphasis could be given to certain grammatical structures, such as, for example, the use of the English determiner system (which does not receive much attention in the Spanish university EFL study programmes), or the noun phrase in general, since the results show that from A1 to B2 levels, there are twice as many noun phrase errors as verb phrase errors. Also, the amount of teaching time devoted to prepositions is minimal, but the results from the coding show that students struggle with prepositions at all levels.

These preliminary results, although tentative, lead us to think about how we can help our students. We propose to reinforce those grammar areas where problems are made. We need, however, to distinguish between explicit teaching of concepts (which is often time-constrained and classroom-centred) versus out of class online drilling and self-study. This latter is another area of interest within TREACLE.

Combining the data obtained from the error analysis (on which this paper has focused) with the automatic syntactic analysis (which is another part of the TREACLE project) will help to create learner profiles which will determine which grammatical features need to be taught, in what order, and with what degree of emphasis and attention in the different Spanish university English language study programmes. Although error analysis is a highly time-consuming activity, it does provide valuable information concerning the learners' interlanguage, partially indicating which direction the latter stages of the project will be taking in the near future. Error analysis, for example, cannot be considered in isolation. On the contrary, it needs to be seen in the context of what the students are attempting. Later work, therefore, will explore learner types based on the experimental versus cautious learners.

7. Bibliography

Andreu, M., Astor, A., Boquera, M., Macdonald, P., Montero, B. and Pérez, C. (2010) Analysing EFL learner output in the MiLC project: An error it's*, but which tag?. In Campoy, M.C., B. Belles-Fortuno and M.L. Gea-Valor (Eds.), *Corpus-Based Approaches to English Language Teaching*. London: Continuum.

Capel, A (2010) Insights and issues arising from the English Profile Wordlists project, *Research Notes,* 41, 2-7.

Dagneaux, E., Denness, S., Granger, S., & Meunier, F. (1996) *Error Tagging Manual Version 1.1*. Centre for English Corpus Linguistics, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Faerch, C., Haastrup, K., Phillipson, R. (1984) *Learner language and language learning*. Clevedon: Multilingual Matters.

Granger, S. (1993) 'International Corpus of Learner English.' in J. Arts, P. de Haan & N. Oostdijk (Eds.), *English Language Corpora: Design, Analysis and Exploitation* (pp.57-71). Amsterdam: Rodopi.

Granger, S. (Ed.) (1998) *Learner English on Computer*. London: Longman.

Granger, S., Hung, J. & Petch-Tyson, S. (2002) *Computer Learner Corpora, Second Language Acquistion and Foreign Language Teaching*. Amsterdam: John Benjamins.

Granger, S. and J. Thewissen (2005a) 'Towards a reconciliation of a "Can Do" and "Can't Do" approach to language assessment'. Paper presented at the Second Annual Conference of EALTA (European Association of Language Testing and Assessment), Voss, Norway, 2-5 June 2005.

Granger, S.and Thewissen, J. (2005b) "The contribution of error-tagged learner corpora to the assessment of language proficiency. Evidence from the International Corpus of Learner English". Paper presented at the 27th Language Testing Research Colloquium, Ottawa (Canada), 18-22 July 2005.

James, C. (1994) Don't shoot my dodo: On the resilience of contrastive and error analysis. *International Review of Applied Linguistics,* 32 (3), 179-200.

Leech, G. (1998) 'Preface' in S. Granger (Ed.) *Learner English on Computer* (pp. xiv-xx). London: Longman.

Hawkins, J. A., & Buttery, P. (2009) Using learner language from corpora to profile levels of proficiency: Insights from the English Profile Programme. In Proceedings of the 3rd ALTE Conference 2008. Cambridge University Press.

O'Donnell, M., Murcia, S., García, R., Molina, C., Rollinson, P., MacDonald, P., Stuart, K. and Boquera, M. 2009. Exploring the proficiency of English learners: The TREACLE project. *Proceedings of the Fifth Corpus Linguistics Conference, Liverpool*.

Odlin, T. (1989) *Language Transfer*. Cambridge: Cambridge University Press.

O'Donnell, M. (2008) Demonstration of the UAM CorpusTool for text and image annotation. *Proceedings of the ACL-08:HLT Demo Session (Companion Volume), Columbus, Ohio, June 2008* (pp. 13-16). Association for Computational Linguistics.

Polio, C. (1997) Measures of linguistic accuracy in second language writing research. *Language Learning*, 47, 101-143.

Rollinson, P. and Mendikoetxea, A. (2010) Learner corpora and second language acquisition: Introducing WriCLE In: J. L. Bueno Alonso, D. Gonzáliz Álvarez, U. Kirsten Torrado, A. E. Martínez Insua, J. Pérez-Guerra, E. Rama Martínez & R. Rodríguez Vázquez (Eds.), *Analizar datos>Describir variación/Analysing data>Describing variation* (pp. 1-12).Vigo: Universidade de Vigo (Servizo de Publicacións).

Programación didáctica mediante el uso de corpus

Montserrat Mola

Jordi Cicres
Universitat de Girona

El propósito de este artículo es analizar algunas de las programaciones de enseñanza del catalán como L2 para adultos desde la óptica de la lingüística de corpus. Parece lógico que si el objetivo es que los alumnos sean competentes en el uso de la nueva lengua en un entorno real, entre los criterios utilizados para la programación de los contenidos lingüísticos de una L2 debería encontrarse la frecuencia de utilización de los distintos elementos morfológicos, sintácticos o léxicos que se pretenden enseñar. Sin embargo, en ocasiones la secuenciación de los contenidos lingüísticos no tiene relación con su frecuencia de uso real en la lengua. Así, proponemos utilizar los corpus lingüísticos como una herramienta útil para asistir a los programadores, y que, de este modo, puedan organizar los materiales en función de criterios más realistas y más acordes con un enfoque comunicativo.

programaciones didácticas, catalán como L2, corpus de referencia, secuenciación

The purpose of this paper is to analyze some syllabuses regarding teaching Catalan-L2 to adults from the perspective of corpus linguistics. It seems logical that if the goal is for students to become competent in the use of the new language in a real environment, the criteria used for programming the linguistic content of L2 course should take into account the frequency of use of different morphological, syntactic or lexical elements. However, sometimes the sequencing of the linguistic content in the syllabus is not related to the frequency of actual use in the language. Therefore, we propose to use linguistic corpora as a useful tool to assist syllabus designers, and thus, to organize their content according to more realistic and consistent criteria, and with a more communicative approach.

syllabus, Catalan as L2, reference corpora, sequencing

1. Introducción

El propósito de este artículo es analizar algunas de las programaciones de enseñanza del catalán como L2 para adultos desde la óptica de la lingüística de corpus. Parece lógico que si el objetivo es que los alumnos sean competentes en el uso de la nueva lengua en un entorno real, entre los criterios utilizados para la programación de los contenidos lingüísticos de una L2 debería encontrarse la frecuencia de utilización de los distintos elementos morfológicos, sintácticos o léxicos que se pretenden enseñar. Sin embargo, la realidad de las programaciones didácticas analizada desde la óptica de la lingüística de corpus muestra que, a menudo, la secuenciación de los contenidos lingüísticos no tiene relación con su frecuencia de uso real en la lengua.

El contexto actual de enseñanza de idiomas, que parte del concepto de competencia comunicativa (Hymes, 1971), pone el foco del proceso de enseñanza-aprendizaje en la comunicación en contextos reales de comunicación. El objetivo del aprendizaje de lenguas no se limita a conocer la estructura interna de la lengua meta (su fonética y fonología, su morfología, su sintaxis, su léxico, etc.) sino que se amplía al conocimiento del uso en la sociedad (competencia sociolingüística) y de sus funciones pragmáticas (competencia pragmática).

Así, el *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación* (MCER, en adelante) establece que el proceso de aprendizaje "en sentido general, se centra en la acción en la medida en que considera a los usuarios y alumnos que aprenden una lengua principalmente como agentes sociales, es decir, como miembros de una sociedad que tiene tareas (no sólo relacionadas con la lengua) que llevar a cabo en una serie determinada de circunstancias, en un entorno específico y dentro de un campo de acción concreto" (MCER: 9). Es decir, pone el acento en el uso social de la lengua.

Eso se traduce en la práctica en el aula en varios aspectos: trabajo con textos reales y completos; uso de variantes dialectales y diastráticas; preponderancia de la lengua oral; etc. Así, los conceptos de enseñanza y aprendizajes basados en la traducción y en el estructuralismo lingüístico ceden paso a otros basados en las funciones del lenguaje. Partiendo de este enfoque, los materiales de apoyo al aprendizaje (ejercicios, libros de texto, actividades) deben de tener en cuenta el uso real del lenguaje en la sociedad.

La lingüística de corpus es la rama de la lingüística que utiliza el análisis de corpus (y por tanto, que reflejan ciertos usos reales de una lengua) con fines varios: extracción de léxico para la elaboración de diccionarios, obtención de datos para hacer una descripción de unos usos determinados de la lengua (lenguaje oral espontáneo, entrevistas, interlengua, etc.), descripción de las estructuras lingüísticas más habituales para mejorar sistemas computacionales, etc.

En este contexto proponemos utilizar los corpus lingüísticos como una herramienta útil para asistir a los programadores, y que, de este modo, puedan organizar los materiales en función de criterios más realistas y más acordes con un enfoque comunicativo (Brumfit & Johnson, 1979; Canale & Swain, 1980) o combinado al lado de la instrucción formal (Long, 1991; Muranoi, 2000). La lingüística de corpus, por su carácter empírico, es una aproximación necesaria para elaborar las programaciones didácticas en los cursos de lenguas extranjeras.

En este artículo tratamos tres ejemplos concretos: en primer lugar, comparamos el uso de los tiempos de pasado en los manuales con respecto al uso real reflejado en un corpus de referencia. En segundo lugar, analizamos algunos elementos léxicos (sustantivos, verbos y adjetivos) introducido en los manuales utilizados y lo comparamos con su uso en el corpus. Y finalmente, examinamos la introducción de los pronombres de relativo en las programaciones didácticas con el fin de comprobar si los más frecuentes son los que se introducen antes.

2. Metodología

Para identificar la problemática relacionada con la secuenciación de los contenidos gramaticales en las programaciones didácticas de los manuales de catalán para extranjeros, se ha realizado un análisis pormenorizado de los tres temas en que se centra este artículo (los tiempos de pasado, el léxico habitual y los pronombres de relativo) en tres manuales de catalán como L2, cuyos rasgos más importantes se detallan en el apartado 2.1. En el siguiente apartado se justifica la elección del corpus de referencia y se señalan sus características.

2.1. Programaciones didácticas

2.1.1. Digui, digui

Se trata de uno de los primeros cursos populares de catalán como L2 (editado entre 1984 y 1985), con un enfoque nocional-funcional (en cada unidad se busca que el alumno trabaje un acto comunicativo concreto), aunque combina aspectos tradicionales (de repetición de estructuras lingüísticas mediante ejercicios descontextualizados). Cada lección presenta una situación comunicativa y una serie de ejercicios relacionados con ella, y al final concluye con un recuadro con el resumen de los aspectos estudiados en la lección (objetivos comunicativos, estructuras lingüísticas y léxico). El manual está pensado para lograr el nivel básico (primer volumen) o elemental (segundo).

2.1.2. Veus

Este manual ya está programado siguiendo las directrices del MCER, y su enfoque se basa en el trabajo por tareas, aunque para los aspectos de gramática se recurre a los inventarios nocionales-funcionales. Los ejercicios propuestos para cada unidad tienen por objetivo tratar una función del lenguaje con un enfoque comunicativo (salutaciones, presentaciones, etc.) mediante recreaciones de situaciones verosímiles. El curso consta de tres volúmenes, que conjuntamente permiten alcanzar un nivel similar al B1.

2.1.3. Passos

Se presentan dos volúmenes como un curso de catalán basado en el enfoque comunicativo por tareas. El primer volumen se corresponde a un nivel A2 del MCER, mientras que con

el segundo se obtiene un nivel equiparable al B1. El enfoque metodológico se traduce en el trabajo por tareas, la gramática y tipología textuales, y el trabajo conjunto de las cuatro habilidades lingüísticas básicas (leer, escribir, hablar y escuchar).

2.2. Corpus

Las nuevas tecnologías han facilitado enormemente el trabajo de creación de distintos tipos de corpus. Estos pueden ser específicos según varios criterios (modalidad de lengua, géneros textuales, periodo histórico, etc.). Según la modalidad, los corpus pueden ser escritos (conformados únicamente por textos escritos), orales (a partir de muestras de lengua oral, ya sea mediante transcripción o con acceso a los archivos de audio) o mixtos.

El corpus utilizado en este estudio es la fracción no literaria del Corpus Textual Informatitzat de la Llengua Catalana (CTILC) elaborado por el Institut d'Estudis Catalans. Las principales características de este corpus son las siguientes:

- Se trata de un corpus extenso: en su parte no literaria, suma más de 29 millones de ocurrencias, a las que hay que añadir otros 13 millones de la parte literaria.
- Está lematizado (con revisión manual de las lematizaciones).
- El acceso a la información del corpus se puede realizar desde el Diccionari de Freqüències (que consta de tres volúmenes: uno dedicado a la parte no literaria; otro a la parte literaria; y el último que ofrece los datos globales), o bien desde la plataforma digital en CD-Rom o vía web (http://ctilc.iec.cat).
- Aporta distintos niveles de información cuantitativa (como se explicita más adelante).
- Es reflejo de la lengua (escrita) real, e incluye también, por tanto, formas no normativas.

Los textos que constituyen este corpus abarcan un periodo cronológico que abarca desde mediados del s. XIX hasta 1988, y se agrupan según las siguientes temáticas: ensayo, narrativa, poesía y teatro (subcorpus literario) y filosofía, religión, teología, ciencias sociales, prensa, ciencias puras y naturales, ciencias aplicadas, bellas artes, ocio, deportes, lengua y literatura, historia, geografía y correspondencia (subcorpus no literario).

Los datos cuantitativos relevantes para este estudio y que están disponibles en el corpus son los siguientes (Rafel, 1996):

- *Frecuencia absoluta* y *frecuencia relativa*. Se trata del número de ocurrencias de un lema (en cualquiera de sus formas), la primera, y el porcentaje que representa un lema en relación al conjunto del cropus.
- *Índice de dispersión*. Indica el grado de uniformidad de la aparición de los distintos lemas en el conjunto de temáticas en que se divide el corpus. Así, un índice cercano a 1 indica que el lema aparece en un porcentaje de casos proporcional en todos los textos que conforman cada temática del corpus.

- *Uso*. Es el resultado de multiplicar la frecuencia absoluta por la dispersión. Así, se corrige el valor de la frecuencia absoluta al disminuir el valor del uso si el lema no aparece homogéneamente en los distintos tipos de texto del corpus.

En este estudio hemos tomado como parámetro de referencia el uso, con el fin de evitar lemas con una alta frecuencia de aparición pero con un grado de especialización. Sin embargo, a pesar de que el CTILC es el corpus en catalán lematizado más extenso, presenta ciertas limitaciones por lo que respecta a los objetivos de este estudio.

- Se trata de un corpus únicamente escrito.
- El periodo histórico en que se hallan comprendidos los textos, que se inicia a mediados del s. XIX y finaliza a finales del XX.
- La temática de los textos del subcorpus no literario (ensayo, historia, ciencias, etc.) no son representativos del uso coloquial de la lengua.

3. Resultados

3.1. Tiempos verbales

Si se analizan los tiempos verbales que aparecen en los niveles iniciales de *Digui, digui*, *Passos* y *Veus* —que corresponden a diez unidades en el caso de *Digui, digui* y *Passos*, y a seis, en el caso de *Veus*—, se puede observar que en los dos primeros casos tan solo se presentan al estudiante las formas de presente de indicativo, gerundio e infinitivo; en cambio, en *Veus* sí que se introducen, además de las anteriores, dos formas de pasado —el pretérito perfecto simple, en su forma perifrástica, y el imperfecto de indicativo, ambos en la unidad 5 del manual— y el imperativo, en la unidad 6.

A raíz del análisis del uso de los tiempos verbales en el CTILC, observamos que efectivamente aparece en mayor proporción el presente de indicativo que las formas de pasado (pretérito perfecto simple e imperfecto de indicativo), casi en una proporción de 2 a 1. Asimismo, se observa que la frecuencia de aparición del imperfecto de indicativo es mucho mayor que el pretérito perfecto simple. Y finalmente, choca la baja frecuencia absoluta de las formas de gerundio, por lo general, lo que no se corresponde con la preponderancia otorgada en los libros de texto. A modo de ejemplo, la Tabla 1 muestra los resultados de 3 verbos:

Tabla 1. Resumen del número de apariciones de distintas formas verbales en el CTILC.

Verbo	Infinitivo	Gerundio	Presente de indicativo	Imperferfecto	Perfecto
Anar	22891	2035	35427	17816	Aprox. 3900
Cantar	2622	1449	3073	1392	Aprox. 190
Dormir	3221	446	1892	1392	Aprox. 120

3.2. Léxico

En este apartado se analizan los substantivos y adjetivos calificativos que se presentan al estudiante en los primeros pasos del aprendizaje de catalán. La secuenciación de introducción del vocabulario es acorde al planteamiento nocional-funcional o por tareas de los diferentes manuales, es decir, se presenta el vocabulario de acuerdo con los campos semánticos relacionados con la noción-función o tarea que se trabaja. De este modo el estudiante puede establecer relaciones entre las diferentes unidades léxicas que va incorporando, lo que facilita su aprendizaje y su uso posterior. Los campos semánticos escogidos se basan en la clasificación que propone el MCER (2002: 55-56): "identificación personal; vivienda, hogar y entorno; vida cotidiana; tiempo libre y ocio; viajes; relaciones con otras personas; salud y cuidado corporal; educación; compras; comidas y bebidas; servicios públicos; lugares; lengua extranjera; y condiciones atmosféricas". Concretamente, los manuales analizados presentan en primer lugar el léxico relacionado con la familia y las relaciones personales y, a partir de ahí, y siguiendo un orden similar en los tres cursos estudiados, introducen términos vinculados a diferentes elementos de uso cotidiano (el tiempo, la vivienda o las localizaciones, la comida, los comercios…).

Sin embargo, partiendo de este planteamiento, los cursos analizados agrupan vocabulario con un uso muy distinto. Vemos que varios términos se presentan en unidades muy tempranas a pesar de tener un uso muy bajo, por lo que extrañamente van a ser utilizadas en un acto comunicativo real. Se trata, por ejemplo, de términos como *vidu* (adjetivo), con un uso de 61,89, *vidu* (sustantivo), con 18,82 —que aparecen en la unidad 1 de *Veus*—, los signos del zodíaco —todos ellos con una presencia inferior a 10 pero que en *Veus* se presentan en la unidad 5—, algunos de los términos que se muestran en relación con la comida en *Digui, digui* como *sípia* (frecuencia: 26), *espàrrec* (153) o *formatgeria* (7), o algunos de los adjetivos que encontramos en *Passos* como *arrissat* (39), *xerraire* (30), *tafaner* (26) o *gandul* (adjetivo, 36).

En contraste, varios estudios muestran como las repeticiones del léxico distribuidas en el tiempo favorecen su consolidación, frente a las realizadas en una sola clase (Gómez Molina, 1997: 69-93). Y sin embargo, palabras como las que hemos ejemplificado, difícilmente volverán a aparecer en lecciones posteriores o serán necesarias para avanzar en el aprendizaje del idioma. Vemos, pues, que la introducción de nuevo léxico por parte del profesor requiere un proceso de selección que debe considerar, entre otros, la frecuencia de uso de los nuevos términos presentados, sin olvidar que deben primar las necesidades concretas de cada alumno o de cada grupo.

3.3. Pronombres de relativo

Por lo que respecta a elementos sintácticos, este trabajo focaliza la atención en el uso y la enseñanza-aprendizaje de los pronombres relativos. Una vez analizados los tres manuales y comparados con el uso de estos elementos que refleja el corpus (*que*: 477.893,54; *qual*: 58.830,92; *quan*: 43.299,76; *què*: 35.526,97; *on*: 29.407,17; *qui*: 24.840,27; *quant*: 5.003,31), se puede observar que se dedica poca atención a los pronombres relativos y que estas estructuras se introducen en niveles medios o superiores (solamente *Passos* introduce las relativas especificativas en el primer volumen). Por otro lado, las tres programaciones estudiadas coinciden en la manera de empezar a trabajar los relativos: a través de las oraciones adjetivas especificativas —relacionándolas con los textos descriptivos—. Por último cabe comentar que en ninguno de estos manuales se explican las relativas con antecedente explícito.

Como se muestra en la Figura 1, los tres pronombres de relativo que se presentan en las primeras unidades son *que*, *on* y *qui*, aunque estos dos últimos se presentan antes que otros más frecuentes, y en principio con un nivel de dificultad similar (como *quan*). Los demás relativos que no se introducen en los niveles iniciales son *qual*, *què* y *quant*.

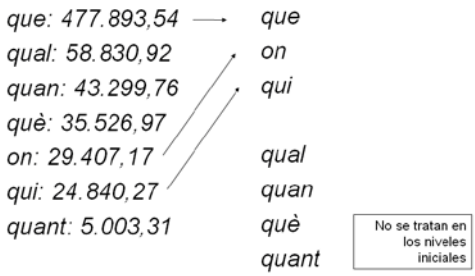


Figura 1. Esquema del orden de aparición de los pronombres relativos.

4. Discusión y conclusiones

Con el fin de abordar los criterios para seleccionar y secuenciar los elementos léxicos y gramaticales en los procesos de enseñanza-aprendizaje de lenguas extranjeras, compartimos la visión de Hinkel (2006: 112) que "many L2 methodologists believe, however, that corpus findings can make L2 teaching far more effective and efficient by identifying the language features that learners must know to achieve their learning goals (e.g., Byrd, 2005; Byrd & Reid, 1998; Conrad, 2000)". Nuestra propuesta es, sin embargo, usar los corpus lingüísticos como herramienta orientativa: ante las dudas acerca de qué estructuras enseñar planificadamente y cuáles no, o en el momento de secuenciar los contenidos, defendemos que deben de primar los aspectos más frecuentes, siempre que cognitivamente no impliquen una dificultad mayor para los aprendices.

Esto es especialmente válido para la enseñanza del léxico. Si bien no hay que intervenir en el aprendizaje del léxico ocasional (el que aparece de forma no planificada a partir de lecturas u otras actividades con otros objetivos), es importante establecer criterios

claros en el aprendizaje del léxico planificado. Mas si cabe teniendo en cuenta que "según los resultados de varias investigaciones en inglés, italiano y español, los recuentos de frecuencias señalan que el vocabulario que utilizan los adultos en la vida diaria no suele exceder de las dos mil palabras" (Gómez Molina, 2000: 26). En este sentido, compartimos los cuatro criterios propuestos por Gómez Molina (2000): "frecuencia de uso junto con la eficacia de la unidad léxica"; "frecuencia de uso y la dispersión o rango en corpus reales"; "uso de textos reales"; y "no seleccionar sino adaptarse a las necesidades del alumno en cada momento". Todo ello sin importar si se fragmentan los campos semánticos (algo que es, siempre, inevitable).

Referencias

Areizaga, E. (2002). El enfoque comunicativo. Propuestas didácticas. En U. Ruiz Bikandi *et al.* (Eds.). *Didáctica de la segunda lengua en Educación Infantil y Primaria* (pp. 137-162). Madrid: Síntesis.

Brumfit, C. J. y Johnson, K. (Eds.). (1979). *The Communicative Approach to Language Teaching*. Oxford: Oxford University Press.

Byrd, P. (2005). Instructed grammar. En E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 545–562). Mahwah, NJ: Lawrence Erlbaum.

Byrd, P. y Reid, J. (1998). *Grammar in the composition classroo*m. Boston: Heinle & Heinle.

Canale, M. y Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1-41.

Conrad, S. (2000). Will corpus linguistics revolutionize grammar teaching in the 21st century? *TESOL Quarterly, 34*, 548–560.

Consejo de Europa (2002). *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación* (Instituto Cervantes, Trad.). Madrid: Instituto Cervantes-Ministerio de Educación, Cultura y Deporte, Anaya (Trabajo original publicado en 2001).

Gómez Molina, J.R. (1997). El léxico y su didáctica: una propuesta metodológica. *Reale, 7*, 69-93.

Gómez Molina, J.R. (2000). La competencia léxica en la enseñanza del español como L2 y LE. *Mosaico, 5*, 23-29.

Hinkel, E. (2006). Current Perspectives on Teaching the Four Skills. *TESOL QUARTERLY, 40* (1).

Hymes, D. (1971). *On Communicative Competence.* Philadelphia: University of Philadelphia.

Long, M. (1991). Focus on form: A design feature in language teaching methodology. En K. De Bot, R. Ginsberg, y C. Kramsch (Eds.). *Foreing language research in cross-cultural perspective* (pp. 39–52). Philadelphia: John Benjamins.

Mas, M., Melcion, J., Rosanas, R. y Vergés, M.H. (1984). *Digui, digui: curs de català per a no-catalanoparlants adults* (Vols. 1-6). Barcelona: Publicacions de l'Abadia de Montserrat.

Mas, M., Vilagrasa, A. (Coord.) (2008). *Veus. Curs de català* (Vols. 1-9). Barcelona: Publicacions de l'Abadia de Montserrat.

Muranoi, H. (2000). Focus on Form Through Interaction Enhancement: Integrating Formal Instruction into a Communicative Task in EFL Classrooms. *Language Learning, 50*(4), 617-673.

Rafel, J. (Dir.) (1996). *Diccionari de freqüències, 1. Llengua no literària.* Barcelona: Institut d'Estudis Catalans.

Roig, N. (Coord.) (2007). *Passos. Curs de català per a no catalanoparlants* (Vols 1-10). Barcelona: Octaedro.

Corpora as tools and resources for the teaching of English vocabulary[191]

María Luisa Roca Varela

University of Santiago de Compostela

Abstract

Corpora are increasingly being used in both theoretical and applied linguistics (Granger, 1994; Palacios, 2005) with satisfactory results. In spite of this, the explicit use of corpora is relatively scarce within the field of applied linguistics and in particular within the area of language teaching. In this paper, I will make reference to how teachers can take advantage of these language databases in EFL settings (Oghigian & Chujo, 2010) and why corpora are valuable resources for the study of vocabulary. After examining some of the most outstanding corpus tools at our disposal, this paper will present different corpus-based tasks which can be used in the classroom for the teaching and learning of difficult vocabulary items. These activities will allow us to reflect on the information provided by native corpora regarding the meaning and use of different lexical items (collocations, colligations, semantic prosody), and we will see how useful it may be to compare and contrast native and learner corpus data to draw conclusions on what learners "know" about a L2 item and what they really "need to know." The ultimate goal of this study is to discuss the pedagogical usefulness of corpora for the teaching of vocabulary.

Key words: corpus linguistics, corpora, pedagogical applications, teaching of vocabulary

Resumen

Estamos asistiendo a un incremento del uso de los córpora en el campo de la lingüística teórica y también aplicada (Granger, 1994; Palacios, 2005). A pesar de ello, el uso explícito de los córpora en el ámbito de la enseñanza de lenguas es casi inexistente. En este trabajo, veremos cómo se pueden aprovechar estas bases de datos para la enseñanza del inglés (Oghigian & Chujo, 2010) y valoraremos más en concreto la importancia de estas herramientas para el estudio de vocabulario. Después de examinar algunos de los corpus más destacados, en este trabajo se presentan diferentes tareas basadas en corpus que podrían ser utilizadas para la enseñanza y el aprendizaje de palabras que son difíciles para los estudiantes españoles de inglés. Asimismo, estas actividades nos permitirán reflexionar sobre la información proporcionada por los córpora en cuanto al significado y el uso de diferentes elementos léxicos (colocaciones, "colligations", la prosodia semántica), y veremos la utilidad que tiene comparar y contrastar los datos en los corpus nativos y los corpus de estudiantes con el fin de extraer conclusiones sobre aquello que los estudiantes "saben" y lo que realmente "necesitan saber. " El objetivo último de este artículo consiste en demostrar la utilidad pedagógica de los córpora para la enseñanza del vocabulario.

Palabras clave: lingüística de corpus, córpora, aplicaciones pedagógicas, enseñanza del vocabulario

191 The research reported in this paper was supported by the Spanish Ministry of Education (grant reference number AP2007-04477) and by the Galician Ministry of Innovation and Industry (INCITE grant number 08PXIB204033PRC-TT-206). These grants are hereby gratefully acknowledged.

1. The contribution of corpora to language teaching

Corpus linguistics is a branch of linguistics, or an approach to the study of language (Gries, 2009), which focuses on the analysis of real samples of language use. Corpus linguistics was born with John Sinclair and the Cobuild project at the University of Birmingham (UK). From its emergence in the 1960s, the popularity of this discipline has grown so much as to influence language teaching. In fact, nowadays, the influence of corpus research is felt on syllabus design, teaching materials (dictionaries and books) and classroom activities (Barlow, 2002; Krieger, 2003). In the case of classroom tasks, the explicit use of corpora has favoured the introduction of new technologies in the classroom. In addition to this, the analysis of corpus-based data has encouraged the learners' involvement and participation and has promoted the development of some basic skills for lifelong learning (*European framework for key competences*, 2007). In this sense, corpora have largely contributed to a shift in language teaching by emphasising autonomous learning and the importance of analysing real language use.

After examining some of the most well-known corpus tools which can be used for vocabulary teaching and referring to how teachers may take advantage of these language databases in EFL settings (Oghigian & Chujo, 2010), this paper presents some corpus-based tasks which show different ways of exploiting corpora for vocabulary teaching and learning. The ultimate goal of this study is to discuss the advantages and disadvantages of corpus work in the acquisition and learning of vocabulary.

2. Some corpus tools for the study of english vocabulary

Fortunately, the growing interest in the area of corpus linguistics has paved the way for the creation of a wide range of different corpora with different characteristics (e.g. ICLE: learner corpus vs. BROWN: native corpus) representing different varieties of English (BNC: British English, COCA: American English, ICE: different varieties of English), registers (COLT: Teenage Language vs MICASE: Academic Spoken English) and modes of communication (BAWE: written English vs LLC: spoken English). These corpora are the result of either individual or collective efforts which gave way to valuable resources for vocabulary analysis.[192] In addition to these large language databases, there are other tools, like free concordances, online versions of commercial corpora and websites which can be helpful for the teacher who is willing to introduce corpus work in the classroom. The so-called *Corpus Concordance* http://www.lextutor.ca/concordancers/concord_e.html, for instance, allows users to obtain concordances from different corpora and work with different languages,[193] and websites such as *Corpora4learning* at http://www.corpora4learning.net/, led by Sabine Brown at the University of Surrey, provide good links and references for the use of corpora in the context of language teaching. Corpus

192 Some practical information which should be mentioned is that most corpora have two options: a reduced online free version and a expanded commercial version in CD-Rom. Free online versions of corpora are sufficient to introduce corpus work in the classroom.

193 This concordancer is part of a bigger website called the *Compleat Lexical Tutor* designed for "data-driven language learning on the web". This tool does not need a concordance program to process corpus information and allows us to make searches and find out about the particular features and usage of different English items using data from different corpora (Brown, BNC, etc).

tools for the English language are numerous and diverse in nature so English teachers should make use of them and resort to corpus-based techniques as a new way of dealing with some areas of the language, for instance, they provide good evidence of vocabulary use in context.

The expansion of a student-centered methodology and the importance given to discovery learning seems to have given way to a proliferation in the use of corpus in the classroom. In addition to this, the fact that corpora provide samples of authentic language use has attracted many teachers to this methodology which allows students to be exposed to the real use of language. We can distinguish two main approaches to the use of corpora for the study of vocabulary: explicit or implicit. Teachers can either encourage students to make use of the relevant software through hands-on, practical activities (explicit) or base their tasks on corpus data (implicit). In any case, corpus-based activities should be motivating, enlightening and make sense to learners. In the following section, I will propose some activities with corpora to work on particular features of vocabulary, aspects such as: a) the meaning and use of English items; b) the pragmatics behind stylistically different doublets; or c) the frequency of occurrence, common patterns and specific collocations of particular lexical items. The effectiveness of corpus work for the teaching of English has been proved by researchers, such as Vannestal and Lindquist (2007) or Kim (2009). Thus, I propose the application of some corpus-based activities for the teaching and learning of difficult English words, such as: *carpet, suburbs* and *career*. An informal introduction of these activities in the classroom proves to be effective and help students realize about the semantic and lexical properties of these words.

3. Activities with corpora for vocabulary learning

The activities proposed in this section are varied as regards format and contents. The main aim of these activities is to help students encounter words in real samples of language use and its lexical connections. With the help of these activities, students become aware of the semantic nuances and the syntactic traits of a number of vocabulary items. The target words in this case are lexical items which appear to be difficult to grasp and use by Spanish learners of English.

The first activity focuses on the analysis of the noun "carpet". This noun is included in the *Longman Communication 3000 frequency list* so it is a high-frequency word in English. However, this item is considered to be problematic for Spanish students of English. The formal resemblance between English *carpet* and Spanish *carpeta* leads students to think that both words mean the same. However, this is not the case. One of the functions that corpora may have for the study of this word is to make the meaning and use of this word clear to students. Students may check the lack of correspondence between both items by searching in a corpus, such as the BNC. So an activity that must be useful for the analysis of this word in context would be the following:

ACTIVITY 1

- Lexical item under analysis: "CARPET". This noun may be misleading for Spanish learners of English.

- Level: Lower-intermediate
- Learning objective: To determine the meaning and lexical context of use that surrounds this lemma.
- Materials: BNC online sampler. Free online at http://www.natcorp.ox.ac.uk/.
- Description of the activity: This activity can be done in two different ways; on the one hand, we can provide students with the printed concordance lines provided by the BNC online sampler and ask them to search for meaningful examples which shed light on the use and meaning of this word; or on the other hand, we can ask them to do a hands-on activity by going to the BNC online sampler website at http://www.natcorp.ox.ac.uk/. Once they click on this link, they must enter the word "carpet" in the "type box" and look for the common patterns and any revealing example that clarifies the meaning of this word form in the examples provided. Below I show the screen shot with the results.

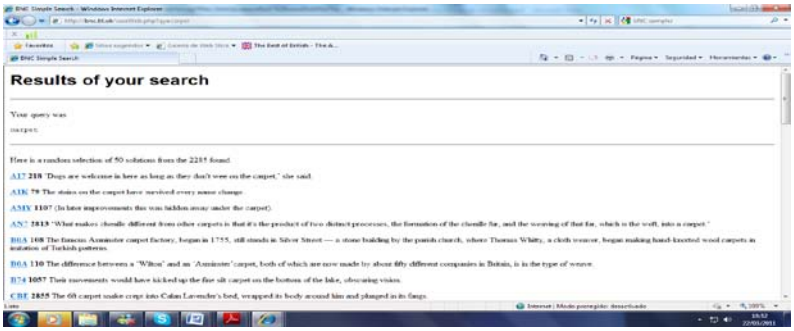


Figure 1: Results for carpet from the BNC online sampler

If we analyse the concordances provided by the BNC online sampler, firstly, we should disregard those examples that are not relevant for our analysis. It is the case of an example like this:

(1) *When you're buying a* carpet, *look out for the BCMA logo on the underside (Code: CCX 1190)*

Here what we know is that carpets can be bought and that they have two sides, but it is quite ambiguous as regards the meaning of the item. We continue with our analysis and we can observe several patterns that occur over and over again, such as the prepositional phrase: *on the carpet.* Furthermore, we can see that this noun is normally associated with negative things related to dirtiness: *wee, stains, mess,* and that it is associated with a particular place "on the floor".

(2) *'Dogs are welcome in here as long as they don't* wee on the carpet,*' she said. (Code: A17 218)*

(3) *The* stains on the carpet *have survived every name change. (Code: A1K 79)*

(4) *There were no pictures on the walls,* no carpet on the floor*, only rough unpolished wooden planks, and there were gaps between the planks where dust and bits of grime had gathered. (Code: CH4 2687)*

(5) *The* mess on that carpet *wants cleaning. (Code: KR0 2389)*

With this information we can draw several conclusions on the use of this noun. Some of them are: the noun tends to occur in a prepositional preceded by "on", and due to its association, it must be something that can get dirty easily. Apart from that, there is an example which is enlightening as regards the meaning and use of this word. It likens the words *rug* and *carpet.* These data give us clues concerning the semantic characteristics of a carpet.

(6) *The terms "rug" and "*carpet*" are normally used to denote size — a c*arpet *being any rug with a surface area in excess of 4.4 m 2 , and whose length is not more than 1½ times its width, i.e., 9′ × 6′ (2.74 × 1.83 m) or 12′ × 8′ (3.66 × 2.44 m). (Code: EX0 82)*

Apart from these sample sentences, students may have focused their attention on other examples. Teachers will ask students what examples have called their attention and why; then, I will ask them to translate this word into Spanish and to build a sentence which would be meaningful to them in order to learn this word correctly. Apart from that, corpora can be exploited to become aware of the syntactic differences between two confusing pair of verbs, such as: *signify/mean*. Apparently, the semantic differences between those verbs are not really outstanding; however, there is a stylistic difference that must be perceived in the analysis of the collocates of these verbs and in the frequency of use of both lexical items. A good tools which allows us to compare and contrast words is the Collins Wordbanks *Online* based on the Bank of English which students can sign up for a trial and use it freely for a month (http://www.collinslanguage.com/wordbanks/). In this case, we can propose a hands-on activity in which students make use of this online resource.

ACTIVITY 2

- Lexical items under comparison: "SIGNIGY vs. MEAN" (two quasi-synonymous terms).
- Level: Advanced
- Learning objective: To analyze the similarities and differences in the lexical collocates of this pair of words.
- Materials: Computer/ Internet connection. Tool: Collins Wordbanks *Online.* Collins Wordbanks *Online* is an online corpus service which contains 57 million words of written and spoken English, from both American and British sources, from the Bank of English http://www.collinslanguage.com/wordbanks/ Login for the trial version (one month).
- Description of the activity: Go to http://wordbanks.harpercollins.co.uk/auth/. Click on WordBanks Online: English. Click on "Word Sketch Differences" and

enter the lemmas to be compared. Below I show the screen shot of word sketch differences.

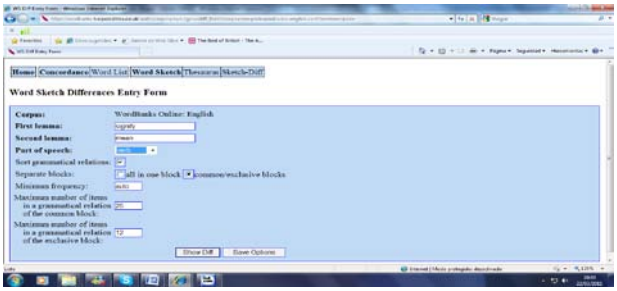


Figure 2: Word Sketch from Collins Wordbanks Online

When we click on "show differences", we have access to another window where we are given information on the frequency of these items; and we see that *mean* (with 256632 occurrences) is far more common than *signify* (2382 examples). In addition to these data, we are presented with common patterns and divergent patterns. It is then when we find that both verbs share some patterns, for instance, it is very common to find *mean* followed by *nothing, anything* and *something,* and *signify* can also be followed by these indefinite pronouns. On the other hand, adverbs, such as *necessarily* and *usually,* are frequent modifiers of both verbs. Concerning "the only patterns of these verbs," the patterns provided in the case of *mean* are clear and understood from the concordances provided.

"mean" only patterns	
Wh_comp 5252	**Adverb modifier 31110**
That , when, whatever	*Literally, inevitably, really*

Figure 3: Recurrent patterns for mean

However, the collocates provided for the "signify only patterns" section are quite ambiguous and of little help for students since if we click on them to see the concordances we see that the same example is repeated several times. The only collocate which could be of a little help could be the object pattern in which we find the noun "sincerity", which can give us an idea of the associations and use this verb has ("to let somebody see, to exhibit"). If we click on sincerity, we get these concordances:

(7) *practised so effortlessly by Bill Clinton to* signify *sincerity and trust. But, hey, this is (Code: brregnews)*

(8) *boss (Claire King). He widens his eyes to* signify *sincerity and amazement. Ah, but will (Source: times)*

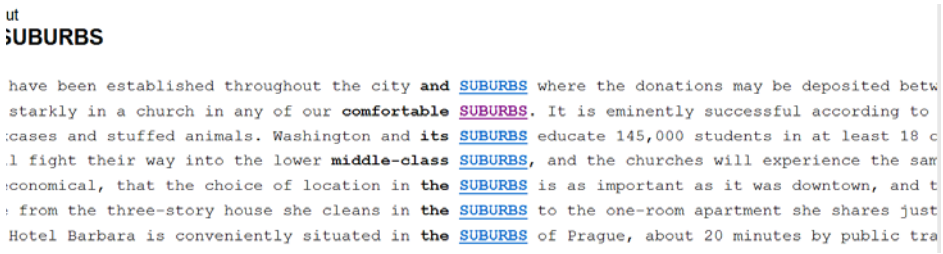
(9) *he right hand over his heart, as Iraqis do, to* signify *sincerity. Fifty dollars must have been (Source: times)*

In any case, when the meaning is not clear enough from the data provided by this corpus, we can look up the term in other corpora or even in a monolingual dictionary when our interest is in the meaning, more than in the context of use of the item (LODCE).

We can also use corpora to look into the polysemy of words or in the associations and semantic prosody of nouns like "suburb."

ACTIVITY 3

- Lexical item under analysis: The noun "SUBURBS"
- Learning objective: To look for the most striking collocates of this word. Students are expected to come to the conclusion that *suburbs* is associated with positive adjectives.
- Materials: Printed version of the results provided by the 'corpus-based concordance' section in the *Compleat Lexical Tutor.*
- Description of the activity: Students will be presented with the results obtained from an advanced search of this item in the BNC and Brown Corpus.


ut
SUBURBS

have been established throughout the city **and** SUBURBS where the donations may be deposited betw
starkly in a church in any of our **comfortable** SUBURBS. It is eminently successful according to
cases and stuffed animals. Washington and **its** SUBURBS educate 145,000 students in at least 18 c
l fight their way into the lower **middle-class** SUBURBS, and the churches will experience the sam
conomical, that the choice of location in **the** SUBURBS is as important as it was downtown, and t
from the three-story house she cleans in **the** SUBURBS to the one-room apartment she shares just
Hotel Barbara is conveniently situated in **the** SUBURBS of Prague, about 20 minutes by public tra

Figure 4: Collocations for suburbs from the Corpus Concordance

When analyzing these concordance lines, we observe that positive words collocate with the word *suburbs*: *comfortable*, *Washington and its suburbs educate*,…

Surprisingly the last concordance line provided by this screenshot reads *Hotel Barbara is conveniently situated in the SUBURBS of Prague;* this is revealing of the positive connotations of this term. The conclusion to draw is that the noun *suburb* does not have a negative connotation. Do you think that advertisers would be emphasising the situation of the hotel if it was a negative term denoting a poor area?

ACTIVITY 4:

A last thing we can try out is to compare real data from native and learner corpora and examine how they differ. I will propose to analyse the noun "career" both in SULEC (Santiago University Learner of English Corpus) and in the BNC in order to determine

the differences and different associations that native and learner speakers assign to this word.

- Lexical item under analysis: "CAREER"
- Learning objective: To look for the differences in the use of this word by native speakers and Spanish learners of English
- Materials: Printed version of two representative examples (one taken from SULEC and one from BNC) which summarize the use of this item by both native speakers of English and Spanish learners of English.
- Description of the activity: Students will be presented with these two examples of *career*. One from SULEC, the other from the BNC and they have to work out in what ways the use by native speakers differ from the use learners give to this word.

SULEC	**BNC**
Most of the students that finish their *career* are not prepared to assume that they probably do not get a job related to the university degree that they have been studying for 3-5 years of their live. (code: SULEC-AE-19)	He became a trader in Nigeria and when this *career* failed, worked for a time as a clerk to Richard Beale Blaize, publisher of the short-lived Lagos Times (code: CDU 52)

Figure 5: Examples of career from SULEC and from the BNC

This activity allows students to reflect on the different connotations and uses of this word by native and non-native speakers allowing us to compare and contrast the results. Students will come to the conclusion that learners use this word with the meaning of "university course or degree" while native speakers use the noun *career* to refer to a "position or job." According to Spanish learners of English, the noun *career* is used in connection with students; thus, s*tudents can finish a career* while in the BNC, *career* is associated with job; for instance, in the example above we see that being a *trader i*s a *career*.

These four activities have shown the potential that corpora have for vocabulary teaching and learning. As explained above, several learning objectives can be accomplished with the use of corpora. We can then conclude that corpora are very useful tools and resources for vocabulary teaching and learning.

4. Discussion and conclusions

As shown in the activities proposed, corpora are useful tools and resources in vocabulary teaching. They allow both students and teachers to give an answer to many different problems and help them understand the mechanics of vocabulary use by discovering powerful attractions between words, the importance of natural and recurrent patterns in

language use. With the introduction of corpora, students are exposed to real language in a different way and they get to know and have evidence that not everything in language is so tidily organized as we sometimes learn in the classroom. The rules of language use change according to the nature of communication (oral or written) or to the social status of speakers (students vs adults). Students get ideas on how language really works by analyzing the language of corpora. In this sense, students are confronted with the practical side of language; they experience language as it really is. Corpus-based language learning also contributes to the development some basic skills and, promotes learners' autonomy through data-driven learning (Johns, 1991; Leech, 1997). Moreover, corpus analysis can be said to be ideal to prepare students for real life. They play an active role in these activities and are encouraged to discover things by themselves and come to conclusions on their own. This discovery idea is motivating for learners who feel to be more involved in the learning process and feel that they can do things by themselves.

5. Pedagogical needs and usefulness of corpora for vocabulary teaching

Corpora are, therefore, suitable for vocabulary study and they can be fruitful if we design motivating activities relevant to the learners' interests. The contribution of corpora to the study of vocabulary is remarkable; several are the advantages of using these language databases. On the one hand, corpora bring real English into the classroom and together with it, the importance of learning autonomously. Apart from that, "corpora allow access to detailed and quantifiable syntactic, semantic and pragmatic information about the behaviour of lexical items" (Carter, 1998: 233), they allow students to analyze the meaning, cotext and situational contexts in which certain structures typically occur. This gives students a more realistic picture of how language works. With the use of a corpus methodology for the study of vocabulary, students become aware of the importance of context in order to carry out an analysis of communication patterns in real language use; they also learn to develop an analytic and critical approach to data. On the other hand, students can feel that they are in contact with language use in real contexts. They practice their deductive skills and notice that corpora may also provide typical and atypical collocations that can be relevant for an accurate use of the target language. However, the introduction of hands-on activities based on corpora could also have a number of disadvantages. As we need computers and the corpus software, students should have a computer at their disposal; furthermore, computers may have either Internet connection for the use of an online corpus or the corpus software installed. In case the corpus does not have a concordancer, there is a need to look for a suitable concordancer to process and analyse data quantitatively and qualitatively. Apart from technological problems that may arise, students need to be familiar with key aspects of corpus work (Cobb, 1997): e.g. background information on corpus representativeness (register, type of data: written spoken, genres represented, etc) and use and interpretation of data provided by corpora. The other important issue about corpus-based discovery activities is that they primarily focus on receptive processes (inferences through exposure to the language) rather than on production, on the productive use of language (speaking and

writing). That is why it is important to add "a communicative bit" after these activities in which students need to put the acquired knowledge into practice (Oghigian & Chujo, 202:2010). Anyway, the introduction of corpus data in language learning has somehow revolutionized and provided a new approach to vocabulary teaching and has helped students to experience real English. The advantages of the use of corpora for vocabulary teaching outnumbered the disadvantages. However, the teaching of vocabulary should not be entirely based on the use of corpus data because it may become redundant and boring as everything else, we should have a good combination of different techniques for the presentation, practice and consolidation of new vocabulary.

6. References cited

Cobb, T. (1997). Is there any measurable learning from hand-on concordancing? *System, 25*(3), 301-315.

Granger, S. (1994). "The learner corpus: A revolution in applied linguistics." *English Today* 39(10/3): 25-29.

Hunston, S. (2002). *Corpora in applied linguistics.* Cambridge: Cambridge University Press.

Johns, T. (1991) "Should you be persuaded: two samples of data-driven learning materials". *English Language Research Journal 4: 1-16*.

Kim, YJ. (2009). *Effectiveness of on-line corpus research in L2 writing: Investigation of proficiency in English writing through independent error correction*. Master of Arts (English as a Second Language). Available at http://digital.library.unt.edu/ark:/67531/metadc12140/m1/1/high_res_d/thesis.pdf

Krieger, D. (2003). "Corpus linguistics: What it is and how it can be applied to Teaching". *The Internet TESL Journal 9.3*. Available at http://iteslj.org/Articles/Krieger-Corpus.html

Leech, G. (1997). Teaching and language corpora: A convergence. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (eds), *Teaching and language corpora,* (pp. 1-23). Edinburgh: Addison Wesley Longman Limited.

Oghigian, K. & K. Chujo. (2010). "An effective way to use corpus exercises to learn grammar basics in English". *Language Education in Asia* 1: 200-214.

Palacios Martínez, I. M. (2005). "Las nuevas technologías y la investigación en el campo de la adquisición de segundas lenguas." In M. Cal, P. Núñez & I.M. Palacios (eds), *Nuevas Technologías en Lingüística, Traducción y Enseñansa de Lenguas,* (pp. 203-223). Santiago: Universidad de Santiago.

The Key Competences for Lifelong Learning – A European Framework. 2007. Luxembourg: Official Publications of the European Communities, Available at http://ec.europa.eu/dgs/education_culture/publ/pdf/ll-learning/keycomp_en.pdf

VANNESTÅL, M.E. & H. LINDQUIST. (2007). "Learning English grammar with a corpus: Experimenting with concordancing in a university grammar course". *Cambridge University Press, 19*(3), 329-350.

CORPORA CITED

BNC ONLINE SAMPLER: XLM EDITION. University of Oxford. Available at http://www.natcorp.ox.ac.uk/

DAVIES, M. (2008-). *The Corpus of Contemporary American English* (COCA). Brigham Young University. Available at http://www.americancorpus.org.

GRANGER, S. (2002). *The International Corpus of Learner English* (ICLE). Université Catholique de Louvain. Available at http://www.uclouvain.be/en-cecl-icle.html

GREENBAUM, S. (1988). *International Corpus of English* (ICE). University College London. Available at http://www.ucl.ac.uk/english-usage/projects/ice.htm

KUCERA, H. & N. FRANCIS. (1960s). *The Brown Corpus*. Brown University. Available at http://www.helsinki.fi/varieng/CoRD/corpora/BROWN/

NESI, H. (2008). *British Academic Written English (BAWE) corpus*. Universities of Oxford Brookes, Reading and Warwick. Available at http://www2.warwick.ac.uk/fac/soc/al/research/collect/bawe/

PALACIOS MARTÍNEZ, I.M. *The Santiago University Corpus of Learner English*. Santiago: University of Santiago de Compostela. Available at < http://sulec.cesga.es/ >

SIMPSON, R. C., S. L. BRIGGS, J. OVENS, & J. M. SWALES. (2002). *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: The Regents of the University of Michigan. Available at http:// quod.lib.umich.edu/m/micase

STENSTRÖM, A.B. (1993). *The Bergen Corpus of London Teenage Language (*COLT). University of Bergen. Available at http://www.hd.uib.no/colt/

Estudio estadístico del uso de la puntuación en estudiantes de educación secundaria

Jorge Roselló
UNED-Valencia

La puntuación es un tema que ha recibido poca atención por parte de la lingüística y que tampoco ha sido muy estudiado en el mundo educativo. Muchos profesores, a menudo, se encuentran con dificultades para enseñar a puntuar y los libros de texto pocas veces plantean ejercicios eficaces que ayuden a los estudiantes a mejorar su conocimiento. En este artículo se analiza el uso de los signos de puntuación en un corpus de textos escritos por estudiantes de secundaria y se realiza un estudio estadístico para observar las variables que más influyen en su comportamiento y extrapolar, en la medida de lo posible, los datos con el fin de extraer conclusiones que sean de utilidad para plantear un aprendizaje más eficaz de los signos de puntuación.

Palabras clave: puntuación, aprendizaje, estudio estadístico.

Punctuation as a topic has received little attention from linguistics and has also been neglected by the educational world. Teachers frequently find difficulties teaching proper punctuation and textbooks rarely offer efficient exercises to help students improve their knowledge in this area. This paper analyzes the use of punctuation marks in a corpus of texts written by students of secondary education and a statistical study is carried out to observe the variables that have a greater influence on this behaviour and to extrapolate, as far as possible, the information in order to extract useful conclusions to promote a more effective learning of punctuation marks.

Keywords: punctuation, learning, statistical study.

1. Introducción

Los signos de puntuación no han tenido un trato de favor en la enseñanza. Como señala Cassany (1995: 175), pocos han tenido la suerte de que se les enseñara a puntuar en la escuela, y si se ha hecho, ha provocado confusiones perniciosas, como la de relacionar en exceso la puntuación con la entonación. El escritor Eduardo Mendoza, no sin cierta ironía, refería de este modo su descubrimiento de la puntuación:

Durante mis largos y aburridísimos años de enseñanza primaria y media, solo se dedicó una hora a explicar superficialmente qué cosa eran los signos de puntuación y su uso. No digo esto como exageración. Por el contrario, me consta que así fue, porque aquella hora única provocó en mí una impresión tan profunda que hoy puedo decir que condicionó mi vida: había descubierto los signos de puntuación y, al mismo tiempo, mi vocación de escritor (Mendoza, 1990: 191)

Por otra parte, el espacio dedicado a estos elementos en los libros de texto que utilizan los estudiantes de enseñanza secundaria es, en líneas generales, muy breve y descontextualizado. No hay ninguna unidad didáctica dedicada a los signos de puntuación, y la información relativa a este apartado se va repartiendo a lo largo de las lecciones como si de compartimentos estancos se tratara y sin establecer una mínima relación entre cada una de las informaciones (Catach, 1994). Esto supone que, cuando se explica, por ejemplo, el punto y coma, no se proponen ejercicios que incluyan otros signos ya estudiados, por lo que se crea así la sensación de que no existe excesiva vinculación entre ellos.

Los signos de puntuación forman parte de la escritura, porque su función principal es la de articular y distribuir la información en el texto. Como señala Figueras (2001: 27), la puntuación debe concebirse como un mecanismo más de organización del texto (al igual, por ejemplo, que los marcadores del discurso), puesto que permite delimitar distintas unidades textuales y señalar de qué modo deben ser interpretadas por el lector.

2. Estudio estadístico

Con el fin de conocer cuál era el grado de conocimiento real de los signos de puntuación, se les pidió a un grupo de estudiantes de enseñanza secundaria (ESO y bachillerato) que escribieran una serie de textos. Nuestro propósito era llevar a cabo con un grupo de ellos (grupo experimental) una actividad didáctica en donde se tratara con cierta profundidad el tema de la puntuación desde un punto de vista discursivo, esto es, considerando los signos de puntuación como elementos de cohesión textual. Una vez terminada la experiencia, se compararían los resultados con otro grupo no sometido a tal actividad (grupo de control) y se diseñarían algunas actividades didácticas encaminadas a mejorar el uso de la puntuación en los textos escritos por los estudiantes.

El corpus lingüístico que sirvió de base para realizar el análisis estaba constituido por un conjunto de producciones escritas por alumnos de ESO (20 alumnos en el grupo experimental y 19 en el grupo de control) y de bachillerato (20 alumnos, tanto en el grupo experimental como en el de control) matriculados en el IES Camp de Túria (Llíria, Valencia) a lo largo de los dos años académicos que duró la experiencia. Sobre estos

textos, escritos al principio y al final de cada uno de los dos cursos, se realizó un estudio estadístico de los signos de puntuación elegidos[194] para el trabajo:

1. Punto y aparte.
2. Punto y seguido.
3. Punto y coma.
4. Dos puntos:
   a) Para anunciar una enumeración.
   b) Para anunciar una cita textual.
   c) Para establecer relaciones.
5. Coma:
   a) En enumeraciones.
   b) En incisos.
   c) Alteración normal de la frase, tematizaciones, inversiones de orden…
   d) Para separar conectores o modalizadores.

Estas variables lingüísticas o dependientes se pusieron en relación con las variables de tipo social, temporal o estilístico que a continuación se detallan:

a) Tipología textual: textos descriptivos, narrativos y expositivo-argumentativos.
b) Grupo: grupo experimental, con el que se realizaron actividades relacionadas con la escritura y la puntuación, y el grupo de control, que nos permitió comparar los resultados.
c) Grado: se trabajó con dos grupos de secundaria y dos de bachillerato.
d) Tiempo de realización: dos cursos académicos (2005-2006 y 2006-2007).
e) Por último, las variables de tipo social: sexo, nivel sociocultural (alto, medio y bajo, a partir del nivel de estudios y ocupación de los padres) y lengua habitual (castellano, valenciano).

Del corpus se extrajeron 15.517 signos de puntuación, que fueron sometidos a tratamiento informático con diversos programas informáticos. En concreto, se utilizó el programa SPSS 11.5 para *Windows*, con el que se realizaron análisis de frecuencias (tanto absolutas como relativas), tablas de contingencia y análisis factoriales. Para el tratamiento de la variación lingüística y la realización de análisis de regresión logística se utilizó el programa *Goldvarb* 2001.

En la tabla 1 se ofrece la frecuencia de uso en la utilización de los signos y el porcentaje de error cometido en cada uno de ellos.

194 Siguiendo el criterio de algunos autores (Catach, 1994; Figueras, 2001), consideramos estos signos básicos o de primer orden.

Tabla 1. Frecuencia de uso y porcentaje de error en cada signo.

Signo	Frecuencia absoluta	Frecuencia relativa	Porcentaje de error
Punto y aparte	1693	11%	11%
Punto y seguido	3853	25%	9%
Punto y coma	142	1%	21%
Dos puntos (global)	248	2%	21%
-Enumeración	117	47%	12%
-Cita	94	38%	7%
-Relación	37	15%	2%
Coma (global)	9581	62%	34%
-Enumeración	1817	19%	1%
-Inciso	4368	46%	20%
-Alteración orden	683	7%	3%
-Conector	2713	28%	10%

Pero, más que contabilizar los errores, nuestro interés residía en saber en qué medida se podían extrapolar los datos de ese corpus a entidades más amplias. Como se sabe, existen dos tipos de estadística: la estadística descriptiva y la estadística de inferencias. La primera busca solamente contabilizar y ordenar de manera cuantitativa los datos, mientras que la segunda nos permite aplicar de forma válida esos datos a entidades mayores que, en realidad, no han sido investigadas en su totalidad.

Realizamos, en primer lugar, un análisis bivariante, que nos permitió conocer la fiabilidad de los resultados obtenidos para poder refutar o no la hipótesis nula o de independencia. Esta hipótesis supone que los resultados obtenidos dependen del azar o de causas aleatorias; es decir, que no son estadísticamente significativas. Mediante la prueba del *chi* cuadrado podemos comprobar si se ha obtenido un resultado estadísticamente significativo. Solo los valores que estén por debajo de 0,05 ($p<0,05$)[195] señalan un rechazo de la hipótesis nula y, por tanto, serán estadísticamente significativos.

En la tabla de contingencia que, a modo de ejemplo, se ofrece a continuación (tabla 2), obtenida con la aplicación informática del programa SPSS, se ha calculado el *chi* cuadrado del signo punto y seguido, y en ella se puede ver los valores estadísticamente significativos ($p<0,05$) relacionados las variables independientes.

Tabla 2. Análisis del chi cuadrado del signo de punto y seguido.

Variables explicativas	Valores chi^2	gl	Significación (p)
Tipología textual	6,847	2	**,033**
Grupo de trabajo	79,576	1	**,000**
Grado	12,592	1	**,000**
Tiempo de realización	10,743	3	**,013**
Sexo	20,895	1	**,000**
Nivel sociocultural	25,240	2	**,000**
Lengua habitual	4,856	1	**0,28**

195 El nivel 0,05 quiere decir que un resultado obtenido que es significativo a dicho nivel, podría ocurrir por azar sólo cinco veces en 100 intentos o experimentos. Este nivel se escogió originalmente porque se consideró que representaba un riesgo razonablemente satisfactorio.

Con el fin de conocer mejor las relaciones entre las distintas variables se realizó a continuación, también con el mismo programa informático, un análisis factorial de componentes principales, cuyo objetivo fundamental es reducir las dimensiones de los factores analizados para hacerlos más fácilmente comparables y descubrir entre qué variables se da una mayor interacción. Se trata de una técnica para representar las variables en un espacio de pequeña dimensión, denominado espacio factorial, que permite interpretar las relaciones entre ellas.

Vamos a ejemplificar con el signo de punto y aparte las operaciones realizadas. Como se ve en la tabla 3, tres factores nos explican casi el 61% de la varianza total en la producción de errores. Las proyecciones de cada una de las variables sobre cada uno de los primeros factores, denominadas saturaciones, se disponen en la llamada matriz factorial (matriz de componentes), tal y como muestra la tabla 4. Observamos que las variables aparecen dispuestas en tres bloques asociados con cada uno de los tres factores. Cada bloque contiene aquel conjunto de variables que presentan máxima saturación en valor absoluto sobre un mismo factor. Si un conjunto de variables presenta saturaciones muy próximas a 1 en un mismo factor, dichas variables estarán correladas entre sí.

Tabla 3. Análisis factorial sobre el punto y aparte para los casos de error en su uso.

Componentes	Autovalores iniciales			Suma de las saturaciones al cuadrado de la extracción			Suma de las saturaciones al cuadrado de la rotación		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	1,705	24,364	24,364	1,705	24,364	24,364	1,555	22,216	22,216
2	1,431	20,440	44,805	1,431	20,440	44,805	1,428	20,400	42,616
3	1,129	16,128	60,933	1,129	16,128	60,933	1,282	18,317	60,933
4	0,921	13,158	74,091						
5	0,772	11,030	85,121						
6	0,611	8,735	93,856						
7	0,430	6,144	100,000						

Tabla 4. Matriz de componentes rotados sobre el punto y aparte en los casos de error.

	Componentes		
	1	2	3
Tipología textual	0,823	0,040	0,105
Grupo de trabajo	0,147	0,580	0,522
Grado	-0,226	0,614	-0,054
Tiempo realización	0,858	-0,023	-0,004
Sexo	-0,152	0,502	-0,635
Nivel sociocultural	-0,016	0,049	0,770
Lengua habitual	-0,215	-0,676	0,021

Por último realizamos un análisis de regresión logística. Hasta ahora habíamos averiguado qué variables o grupos de factores eran estadísticamente significativos en cada uno de los signos de puntuación y cuáles mostraban mayor incidencia en la varianza total,

pero todavía no sabíamos la importancia o peso específico de cada una de las variantes que la componen. Es decir, podíamos saber, por ejemplo, que el grado era una variable estadísticamente significativa a la hora de analizar los errores que se habían producido en el punto y aparte, pero no sabíamos el peso específico de cada uno de los grupos (ESO y Bachillerato) en el comportamiento de la variable dependiente. Hoy en día existen técnicas y aplicaciones estadísticas que nos permiten averiguan estas y otras informaciones sin necesidad de realizar complicadas operaciones y sin necesidad de tener conocimientos muy profundos de estadística. Los programas *Varbrul* permiten trabajar con la covariación sociolingüística y presentan la información de una forma adecuada a los intereses de los investigadores. Uno de los programas de la familia *Varbrul* es el *Goldvarb* 2001, que hemos utilizado en este estudio.

La aplicación de estos programas de análisis probabilístico nos da información sobre cuándo una serie de factores explicativos aparecen conjuntamente, y si estos factores contribuyen significativamente a explicar los datos o si deben ser desestimados. Así pues, después de este análisis, estamos en condiciones de saber qué grupos de factores (variables) y qué factores (variantes) son realmente importantes para nuestro trabajo porque están explicando el comportamiento lingüístico en la producción de errores[196] y qué factores pueden ser desestimados porque no están influyendo para nada en el uso de los signos. Por tanto, la aplicación de estos programas tiene como uno de sus principales objetivos determinar la fiabilidad de los análisis.

Ejemplificamos aquí el estudio realizado para el signo de punto y aparte.

```
Input 0,106
Run  #  17, 24 cells:  Up
Convergence at Iteraction  5
Group  #   3 - - 2: 0,453, 1: 0,547
Group  #   4 - - 1: 0,606,  2: 0,551,  3: 0,441,  4: 0,401
Group #    6 - - 2: 0,514,  1: 0,390,  3: 0,597
Log likelihood = -558,315  Maximum posible likelihood = -534,802
Fit:   X-square(18) = 47,025, rejected, p = 0,0105

  Run  #  43, 24 cells:  Down
Convergence at Iteraction  5
Group  #   3 - - 2: 0,453, 1: 0,547
Group  #   4 - - 1: 0,606,  2: 0,551,  3: 0,441,  4: 0,401
Group #    6 - - 2: 0,514,  1: 0,390,  3: 0,597
Log likelihood = -558,315  Maximum posible likelihood = -534,802
Fit:   X-square(18) = 47,025, rejected, p = 0,0105

    All remaining groups significant
Groups eliminated while stepping down:   7  2  5  1
Best stepping up run:  # 17
Best stepping down run:  # 43
```

Cuadro 1. Análisis binomial de subida y bajada en el punto y aparte.

196 Por ejemplo, si el uso erróneo del punto y aparte tiene una relación directa con la tipología textual y si la mayor incidencia en el fallo está en los textos narrativos.

La interpretación (resumida) de los datos es la siguiente:

- Los grupos de factores eliminados en este análisis son: 7 (lengua habitual), 2 (grupo de trabajo), 5 (sexo) y 1 (tipo de texto).
- La mayor incidencia en el error viene señalada por los siguientes grupos de factores:
  - 3 (grado), donde la variante ESO (0,547) presenta mayor peso específico, aunque con escasa diferencia sobre el Bachillerato (0,453).
  - 4 (tiempo de realización), donde el mayor peso específico (0,606) se encuentra en la variante 1 (fase inicial), y va descendiendo en la siguientes fases: 2 (0,551), 3 (0,441) y 4 (0,401).
  - 6 (nivel sociocultural), donde la variante 3, correspondiente al nivel sociocultural bajo, presenta mayor peso específico (0,597), con escasa diferencia con el nivel medio (0,514), pero a una distancia considerable del el nivel sociocultural alto (0,390).

3. CONCLUSIONES

El análisis estadístico que hemos realizado nos permite ver, a partir del corpus utilizado, cuáles han sido los errores más habituales en los estudiantes y, sobre todo, nos permite también extrapolar los datos a entidades mayores, a una población de similares características, por lo que podemos idear propuestas didácticas que sean válidas para estudiantes de enseñanza secundaria obligatoria y de bachillerato.

Si nos fijamos en las variables tenidas en cuenta para realizar el estudio, podemos señalar algunas conclusiones. Observamos, en primer lugar, que en los textos narrativos es donde se produce un mayor número de errores (sobre todo en el signo de la coma), por lo que convendría enfocar la didáctica de este signo teniendo en cuenta la diversidad textual y, especialmente, la especificidad del texto narrativo (Silva y Morais, 2007).

Los alumnos de Bachillerato obtienen mejores resultados que los de la ESO, lo cual corrobora la idea defendida por algunos autores (Rocha, 1995), que señalan que la puntuación es una adquisición tardía, relacionada con la sintaxis y la planificación textual, lo que supone cierta madurez en el alumnado.

Otra conclusión importante tiene relación con la variable tiempo. Como hemos señalado, la experiencia se prolongó durante dos años académicos, y los resultados estadísticos muestran que los errores van disminuyendo a lo largo del curso, pero vuelven a aumentar al año siguiente. Quiere ello decir que el periodo vacacional actúa de manera negativa para el aprendizaje, puesto que en el segundo año de la experiencia se vuelve casi a los mismos niveles que se tenía al comienzo. Por consiguiente, hay que contar con que parte de lo aprendido se olvida durante este periodo y conviene reforzar los aprendizajes ya adquiridos a principio de curso.

Otro aspecto importante tiene que ver con el nivel sociocultural del alumnado. Los alumnos pertenecientes al nivel sociocultural alto obtienen mejores resultados, tanto si

pertenecen al grupo experimental y, por tanto, han recibido una información específica sobre el tema, como si pertenecen al grupo de control y no han realizado ejercicios relacionados con los signos de puntuación.

Por último, conviene señalar que los signos de punto y coma y dos puntos fueron muy poco utilizados durante la experiencia en relación con los demás, por lo que habrá que tenerlos muy en cuenta a la hora de diseñar planteamientos didácticos, pues al poco uso que se hace de ellos se suma el bajo dominio de los que sí se atreven a utilizarlos en sus escritos.

Referencias bibliográficas

Cassany, D. (1995). *La cocina de la escritura*, Barcelona: Anagrama.

Catach, N. (1994). *La ponctuation (Historie et système),* Paris: Presses Universitaries de France.

Figueras, C. (2001). *Pragmática de la puntuación*, Barcelona: Octaedro/EUB.

Mendoza, E. (1990). Los signos de puntuación. En M. Mayoral (Coord.), *El oficio de narrar* (pp. 191-198). Madrid: Cátedra/Ministerio de Cultura.

Rocha, I.L.V. (1995). Adquisición de la puntuación: usos y saberes en la escritura de narraciones. *Lectura y vida*, 16: 4, 41-46.

Silva, A. y Morais A.G. (2007). Pontuação e generos textuais: uma análise das produções escritas de alunos da escola pública. *Língua Escrita,* v.1, 61-76.

Uso de corpus orales de aprendientes para la enseñanza del francés como lengua extranjera

Ana Valverde Mateos
Universidad de Castilla- La Mancha

Resumen

Los aprendientes de segundas lenguas suelen tener muchas dificultades para expresarse y comprender mensajes provenientes de hablantes nativos, en parte, porque están acostumbrados a trabajar con registros de habla considerados estándar basados en la norma escrita. Nuestro estudio supone la creación de CORAF, un corpus oral compuesto por entrevistas a hablantes de francés como L2 y LE de diferentes niveles, que pretende analizar los aspectos más problemáticos de asimilar para los estudiantes hispanohablantes y sus errores más frecuentes. Todo ello con el fin de sensibilizar a los docentes en el uso de materiales auténticos en un entorno no nativo y para incentivar la creación de métodos de enseñanza más efectivos, así como herramientas y aplicaciones tecnológicas de carácter pedagógico.

Palabras clave: *corpus orales, corpus de aprendientes, francés lengua extranjera, adquisición de segundas lenguas, análisis de errores*

Abstract

Second language learners usually face several problems to communicate and understand messages from native speakers, partly because they normally work with a standard language based on written norm. Our work involves the development of a multi-level oral corpus from French second language learners' interviews. We try to analyze problematic aspects of language acquisition and usual errors made by Spanish mother tongue learners. All this to make teachers aware of authentic materials use in non-native environment and to motivate the development and implementation of more appropriate teaching methods and language learning software applications.

Keywords: *oral learner corpora, French, second language acquisition, error analysis.*

1. Introducción

Es un lugar común referirse a la necesidad del dominio de lenguas en la comunicación académica, profesional o personal en una sociedad globalizada. En el citado marco, parece oportuno centrarnos en el doble hecho de que al parecer, el dominio de la destreza de la expresión oral requiera un mayor recorrido que la del escrito, y que la comprensión oral de los mensajes espontáneos de nativos presente asimismo un nivel de dificultad superior a la del escrito.

Con la introducción creciente de documentos audiovisuales en las aulas, es indudable que, cada día con mayor frecuencia, los estudiantes están expuestos a muestras variadas de expresión características de diferentes registros, desde los más coloquiales hasta los más formales. No es menos cierto que el habla espontánea de los nativos pertenece al registro coloquial, mientras que en la educación sistemática, -mediante la práctica de intercambios, la realización de ejercicios u otras actividades en el aula-, se suele hacer especial hincapié en los modelos de lengua formal o estándar, la cual se basa normalmente en la norma escrita, que es mucho más ordenada y presenta menos variaciones que el francés oral espontáneo. De modo que el registro que acaban manejando *de facto* los estudiantes de FLE resulta finalmente muy alejado del de la lengua oral espontánea de los nativos. Esa circunstancia podría explicar por sí sola las dificultades de comunicación entre ambos.

Partiendo de estas consideraciones, nos planteamos la posibilidad de realizar un estudio que permita conocer con más precisión los obstáculos a los que los aprendientes de francés se enfrentan en su proceso de adquisición de las competencias orales, dejando al descubierto necesidades formativas específicas para una comunicación efectiva en la lengua meta.

Para la fase de análisis, era conveniente contar con una gran cantidad de datos, -en este caso, producciones orales de aprendientes de Francés LE-, con el fin de ponernos en condiciones de extraer hipótesis válidas para el conjunto. El mejor camino para disponer de datos de este tipo es trabajar sobre la base de un corpus lingüístico. Sin embargo, se daba la circunstancia de que la gran mayoría de corpus de aprendientes existentes eran escritos, y no estaban disponibles para el público en general. Casi todos los corpus nacen en el seno de laboratorios de investigación en el marco de un proyecto que tiene un fin determinado y cuyo acceso es restringido. Por otra parte, aquellos corpus que sí están disponibles y se encuentran publicados en editoriales internacionales, son demasiado costosos, y parecen no satisfacer todos nuestros requisitos. En consecuencia, tomamos la decisión de elaborar un corpus oral con aprendientes adultos de distintos niveles, transcrito ortotipográficamente y con anotación de errores de forma manual.

2. Objetivos iniciales del proyecto[197]

En coherencia con todo lo expuesto anteriormente, además de la creación de un corpus oral de aprendientes de francés, nuestro proyecto abarca los siguientes objetivos:

197 La presente investigación se financia gracias a una beca FPU concedida por el Ministerio de Educación.

a) Efectuar un análisis de los errores más frecuentes en la expresión oral de los estudiantes de francés, teniendo en cuenta los distintos niveles del Marco Común Europeo de Referencia para las Lenguas y, a continuación, deducir sus posibles causas, siguiendo la metodología clásica de Corder (1967);

b) Sensibilizar a los futuros docentes (habitualmente también hablantes no nativos) sobre las características y necesidades especiales de estudiantes cuyo aprendizaje se desarrolla en un entorno donde la lengua de estudio no se habla fuera del aula, y en el que existen escasas oportunidades para la comunicación en L2;

c) Proponer pautas metodológicas de enseñanza adaptada a las necesidades de los aprendientes hispanohablantes para favorecer un aprendizaje más rápido y eficaz de la comunicación oral.

3. Primera aproximación al Corpus de aprendientes de Francés Lengua Extranjera (CORAF)

CORAF es un corpus oral de aprendientes de francés que pretende servir de apoyo a diversas investigaciones sobre la adquisición de segundas lenguas. En esta sección, nos referiremos a su naturaleza, así como a su proceso de concepción y de implementación.

3.1. Descripción del corpus

CORAF se define como un corpus monolingüe con fines de investigación. Por consiguiente, se trata de un corpus diseñado *ad-hoc*, que no aspira a convertirse en un corpus de referencia (la complejidad de la realización de un corpus de referencia sería prácticamente inabarcable para un solo investigador). Como se verá, su tamaño es relativamente reducido, cerca de 30.000 palabras, pero se prevén ampliaciones en sucesivas etapas con el fin de aumentar su grado de representatividad de la expresión del conjunto de aprendientes con lengua materna española.

Es de sobra conocido que los corpus lingüísticos no se emplean habitualmente en la enseñanza debido a su complejidad. Muchos de los potenciales usuarios se pierden ante el gran volumen de datos, que, por otra parte, estos no son nada explícitos y presentan una apariencia con la que los docentes no están familiarizados debido a la anotación y la transcripción. Es por ello que siguiendo a Aston (2002), quien recomienda para ámbitos como el aprendizaje la creación de corpus más pequeños -de 20.000 a 200.000 palabras-, a los que llama *home-made corpora* [Aston (2002:9)], que en muchas ocasiones son además expresamente adaptados para el uso por aprendientes. Parece evidente que un corpus más reducido y transcrito de manera más inteligible resulta más sencillo de manejar tanto por los aprendientes como por los docentes que no se perderán entre la masa de datos, y pueden interpretar sin problemas aquello que pretenden investigar. Ese es precisamente nuestro propósito: realizar un corpus accesible y, sobre todo, que pueda ser reutilizado en futuras aplicaciones tecnológicas y pedagógicas.

En la actualidad, nuestro corpus cuenta con 64 entrevistas de una duración media de 15 minutos cada una, y con un total de 16 horas de grabación. Dichos documentos orales están

basados en un diálogo espontáneo sobre aspectos de la vida cotidiana de los aprendientes, preferencias, planes futuros, experiencias lingüísticas, etcétera.

Las grabaciones del corpus CORAF se organizaron según los seis niveles de referencia estipulados en el Marco Común Europeo de Referencia para las Lenguas (en adelante MCER): A1, A2, B1, B2, C1 y C2.

3.2. Características de los entrevistados

En la compilación de este corpus participan aprendientes de todos los niveles del MCER, asignados en función del curso en el que están escolarizados en sus respectivas entidades educativas[198].

Así, planteamos a los centros colaboradores la necesidad de incluir entrevistados de ambos sexos, de un rango de edad amplio y con características socio-profesionales diversas. Todos esos datos específicos, muy útiles para un análisis sociolingüístico, aparecerán detallados en las transcripciones junto con otros relevantes para el análisis del discurso y el análisis de errores como: los años que lleva cada estudiante aprendiendo francés, el tiempo de estancia en países francófonos o su grado de conocimiento de otras lenguas.

Para que las entrevistas se asemejaran en lo posible a una conversación espontánea entre hablantes de L1, preferimos trabajar con estudiantes con una motivación elevada para el estudio, por lo que, una vez explicado el proyecto, pedimos que se ofrecieran como voluntarios a estudiantes entre los definidos como *non-captifs* (Courtillon, 2003), es decir, aquellos que estudian el francés, no como parte de sus estudios obligatorios, sino que han decidido aprenderlo de *motu* propio. Por consiguiente, nos orientamos con prioridad a las Escuelas Oficiales de Idiomas y, en segundo término, a los estudios de Filología.

Partiendo de este requisito, podemos señalar que nuestro corpus está compuesto, hasta el momento, por 64 entrevistas, con una cohorte de 44 mujeres y 20 hombres, de edades comprendidas entre los 17 y los 66 años, con estudios previos (en su mayoría universitarios), con procedencias sociales y geográficas diversas, algunos con conocimiento de otras lenguas (inglés, principalmente) y menos de la mitad con estancias previas en países francófonos superiores a un mes. Los datos temporales de las entrevistas son los que ofrecemos en la siguiente tabla:

Nivel MCER	Locutor (m/h)	Duración media entrevista	Duración
A1	10M/3H	11m 46 s	2h 32m 56s
A2	6M/7H	13m 55s	3h 00m 59s
B1	6M/1H	15m 44s	1h 58m 09s
B2	6M/2H	16m 30s	2h 12m 02s
C1	9M/2H	16m 57s	3h 06m 30s
C2	7M/5H	16m 11s	3h 14m 16s
TOTAL	**44M / 20 H**	**15m 11s**	**16h 15m 42s**

Figura 1. Cuadro resumen de las grabaciones realizadas por niveles.

198 En futuras investigaciones prevemos analizar las grabaciones sin partir de un nivel determinado de escolarización, para intentar definir así el nivel real según el discurso, basándonos en los errores y en el conjunto de ítems conseguidos.

3.3. Entrevistas por niveles

CORAF es un corpus oral que, como ya queda indicado, pretende recoger muestras de habla espontánea de una variedad de aprendientes de francés. Queríamos conocer cómo los hablantes de L2 utilizarían la lengua en una interacción comunicativa ordinaria no preparada, y similar en lo posible a la que realizarían en un entorno nativo. Por tanto, nuestra técnica consistió en dejar que el hablante se expresara sin trabas, durante el tiempo que estimara oportuno, y sin corregir sus errores.

Evidentemente, ante una grabadora visible y un entrevistador desconocido, el estudiante se encontraba en una situación diferente a las que estaba acostumbrado. Por eso, con el fin de estimular su expresión, realizábamos preguntas sencillas y de carácter abierto que le resultaban familiares, por referirse a temas relacionados con los contenidos mínimos establecidos en el currículo de los distintos cursos de la Escuela Oficial de Idiomas (EOI), que ya habían sido abordados en clase.

Así, en los niveles básicos (A1 y A2) hay un número menor de preguntas (de diez a doce), pues nuestra experiencia nos aconseja realizar con dichos aprendientes preguntas especialmente claras y sencillas, ya que todavía no son capaces de expresarse con fluidez y tranquilidad durante más de 15 minutos seguidos.

El resto de niveles hay una media de treinta preguntas entre las que se incorporan las de los niveles inmediatamente inferiores, que son complementadas con otras de complejidad creciente, hasta alcanzar los niveles de competencia C1 y C2.

Los contenidos de las encuestas se caracterizan por incluir:

- Una parte de introducción, que sirve para que el estudiante se presente, y también para que se familiarice con la situación, reduciendo así su posible nivel de ansiedad inicial.
- Una serie de preguntas sobre aspectos cotidianos diversos (referencias a la familia, a su trabajo y/o estudios, actividades que realiza en su tiempo libre, descripción de su lugar de origen, experiencias del pasado, planes para el futuro, etcétera).
- Cuestiones sobre su aprendizaje de la lengua: en niveles inferiores, se pregunta si conocen o estudian otras lenguas, por qué han elegido el francés y qué es lo que les cuesta más aprender. En el resto de niveles se hace una reflexión más profunda y se indaga además en sus experiencias de interacción con hablantes nativos, en sus recursos personales cuando no pueden comunicarse eficazmente (gestos, palabras en otro idioma…) y, en sus procesos de aprendizaje, animándoles a que realicen posibles críticas a la metodología de enseñanza y sugieran posibilidades de mejora.

3.4. Proceso de recogida de datos

Ya mencionamos nuestra decisión de contar con las Escuelas Oficiales de Idiomas debido al carácter generalmente voluntario con el que se cursan esos estudios. Otra de las ventajas

que ofrecen estas instituciones es que en ellas hay aprendientes de todos los niveles del MCER, que, por otra parte, son adultos en su mayoría, lo cual implica que la gestión de permisos de grabación y de difusión es menos compleja.

Una vez contactadas todas las EOI de Castilla-La Mancha, se realizan las grabaciones en las tres únicas que muestran un interés explícito en participar en el estudio, completándose posteriormente la muestra con entrevistas a estudiantes de Filología Francesa de la UCLM.

Las entrevistas se graban sin interrupciones, en salas habilitadas a tal efecto en los distintos centros, con la grabadora visible y sin que los entrevistados conozcan las preguntas con anterioridad. Transcurren a modo de conversación distendida con la investigadora, dejando que el entrevistado se exprese naturalmente, sin corregir sus errores y ayudándolo o reformulando las preguntas en momentos puntuales en que da muestras de encontrarse perdido. Todos los entrevistados contestan a todas las preguntas, si bien es cierto que no todas las conversaciones son iguales, puesto que el entrevistador realiza las nuevas preguntas en función de las respuestas anteriores. Esto hace que el orden inicial de las preguntas pueda verse alterado, bien porque el entrevistado se anticipa a cuestiones que estaban previstas para más tarde, o porque toma la iniciativa de ahondar en determinados temas. Todo ello tiene como resultado la espontaneidad que se pretendía como característica esencial de este corpus oral.

4. Desarrollo e implementación del corpus de aprendientes de FLE

Una vez recogidas las muestras de habla necesarias, se inicia una nueva fase, aún en desarrollo, donde se procede al tratamiento y transcripción del sonido y su posterior análisis y etiquetado de errores.

4.1. Tratamiento y transcripción del sonido

El sonido archivado en la grabadora digital se exporta en formato .wav para llevar a cabo unas mínimas tareas de tratamiento que consisten, generalmente, en el corte de ruidos iniciales y finales de encendido y apagado de la grabadora, su conversión a un formato de menor tamaño, reduciendo con ello la muestra a un número menor de Hz (22000 hz) y, en caso necesario, se procede a una leve amplificación del sonido para una mejor audición.

Posteriormente, se transcribe con la ayuda del *software* Transana siguiendo las convenciones propias del Laboratorio de Lingüística Informática de la UAM[199] para los corpus orales, que provienen en origen de la metodología CHAT, modificada en sucesivos proyectos de corpus orales como C-ORAL-ROM, MAVIR y CHIEDE. Las transcripciones se podrán consultar en ficheros de texto plano (txt), y también en formato XML.

Tenemos que señalar que hemos procurado que la transcripción sea fiel al discurso del hablante. Es decir, se respeta todo lo dicho por el entrevistado, reflejándolo de la misma manera en la transcripción, respetando su expresión y los posibles errores de su discurso. Las correcciones se proporcionan en los comentarios de la transcripción y se hace

199 http://www.lllf.uam.es/ESP/

especial hincapié en mantener la fenomenología de la lengua oral, respetando todas sus particularidades y no considerándolas, a diferencia de otros investigadores, errores por su comparación con la lengua escrita.

Con el fin de fomentar al máximo el uso del corpus y las transcripciones, así como de evitar problemas del incremento del número de etiquetas, en principio, se ha optado por hacer dos versiones de transcripción, una donde se encuentre la transcripción en sí, enriquecida por ciertas marcas para señalar fenómenos de la interlengua (neologismos por interferencia de otras lenguas, problemas de conjugación, interferencia de la lengua materna…), y otra que contenga el análisis completo de errores. La primera transcripción será la más ligera con el fin de ayudar al usuario en su consulta y también para favorecer su reutilización en aplicaciones informáticas futuras.

4.2. Anotación y análisis de errores (AE)

La anotación de errores (AE) constituye uno de los elementos más valiosos del corpus oral de aprendientes, sobre todo, por la gran cantidad de información que puede proporcionarnos sobre la adquisición del francés en hablantes de lengua materna española.

La AE se realiza en un fichero diferente, que respeta el formato de la transcripción, pero que incluye una serie de etiquetas donde se señala el error y su forma correcta, desde distintas perspectivas: la categoría gramatical a la que pertenece (criterio gramatical o lingüístico), el mecanismo de cambio que se produce en el error (criterio descriptivo), y las posibles causas del mismo (criterio etiológico). Queremos también atender a aspectos positivos de la lengua del aprendiente, analizando las estrategias de comunicación que emplea para solventar problemas surgidos en la interacción.

5. Primeras conclusiones y trabajo futuro

El corpus CORAF se encuentra aún en fase de desarrollo y, por tanto, sólo disponemos de algunas hipótesis de trabajo, así como de resultados parciales del AE.

Entre ellos, podemos destacar que:

- Un 80% de los entrevistados cree necesaria una mayor dedicación a la expresión oral en sus clases, ya que consideran la pronunciación y la comunicación oral como lo más difícil del aprendizaje del francés.
- Los errores más frecuentes se producen, *grosso modo*, en las categorías de preposición (elección falsa recurrente), nombre (forma errónea en la concordancia en género y número por tendencia a la regularización de la norma), determinantes (forma errónea en la concordancia en género y número) y verbos (problemas en la negación-omisión- y forma errónea de tiempo verbal).
- En niveles básicos, los errores más frecuentes provienen principalmente de la interferencia de la L1, por la tendencia lógica a traducir literalmente desde su L1 aquello que el aprendiente quiere expresar. Posteriormente, aumenta la proporción de errores intralinguales por procedimientos de simplificación y de hipergeneralización.

El trabajo futuro, una vez realizado el análisis pormenorizado de errores, se encaminaría a:

a) Realizar actividades para docentes y aprendientes basadas en el corpus oral (siguiendo la metodología del *data-driven learning*);

b) Mostrar un análisis contrastivo con un corpus oral de hablantes nativos de las mismas características que CORAF;

c) Crear una plataforma digital de aprendizaje del francés basada en corpus orales, tanto de aprendientes del francés como de hablantes nativos, donde tenga cabida el estudio de la gramática en contexto y la práctica de funciones comunicativas esenciales.

6. Referencias Bibliográficas

Aston, G. (2002). The learner as corpus designer. En B. Ketteman, & G. Marko, *Teaching and learning by doing corpus analysis* (págs. 9-25). Amsterdam: Rodopi.

Braun, S. (2005). From pedagogically relevant corpora to authentic language learning contents. *ReCALL 17 (1)* , 47-64.

Courtillon, J. (2003). *Élaborer un cours de FLE.* Paris: Hachette.

Corder, S. P. (1967). The signifiance of Learner's errors. *IRAL, 9* , 161-170.

Influencia del feedback en el alumnado de educación primaria con respecto a su producción oral en lengua extranjera.

Mª Isabel Velasco Moreno

Resumen

La estrecha relación entre feedback y enseñanza hace que muchas investigaciones sobre docencia hayan profundizado en mayor o menor medida sobre este fenómeno. Sin embargo, la mayor parte de estos trabajos ofrece un estudio teórico sobre el tema, centrándose en la figura del docente y las razones que originan el feedback, aportando sugerencias para hacerlo más eficaz.

Nuestra investigación tiene por objeto efectuar un análisis descriptivo de la realidad escolar actual, estudiando el feedback acaecido en el aula de LE independientemente de quien lo haya suministrado. Desde el punto de vista del alumno intentamos descubrir lo que percibe, distinguiendo diferentes tipos de FB para, posteriormente, poder investigar los efectos ocasionados con relación a su producción oral.

Análisis del discurso comunicativo, enseñanza LE, feedback, análisis del corpus lingüístico, discurso del aula.

Abstract

Feedback has always been connected with the teaching-learning process. Many studies have been done about it but most of them are theoretical, studying ways of being administered, reasons to give feedback and how to make it more effective, always focusing on the teacher as the one who gives feedback in class.

Our empirical research shows a descriptive analysis of real classes of English as a foreign leanguage for Spanish children (11-12 years old) in Andalusia. It investigates who gives feedback and it shows types of feedback perceived from students' point of view as well as effects provoked by them.

Communicative discourse analysis, L2 teaching, feedback, corpus linguistic analysis, classroom discourse

1. Introducción

Con frecuencia, la investigación del proceso educativo va acompañada del estudio y la reflexión del fenómeno de FB al estar estrechamente relacionado con el aprendizaje.

A partir de los estudios de Thorndike (1913) comenzó a considerarse tanto el refuerzo positivo como el negativo como dos elementos esenciales para la modificación de conducta. Hasta la década de los 80 del pasado siglo se mantuvo el auge del conductismo, para dar paso posteriormente al cognitivismo destacando especialmente la figura de Krashen (1981) y sus aportaciones sobre la adquisición de segundas lenguas. Su hipótesis de auto monitorización es especialmente relevante al indicar que la identificación y la observación del propio error así como el análisis de las causas que dieron lugar al mismo ayudan al estudiante a aprender del mismo e incluso subsanarlo. Por otro lado, sugiere la necesidad de presentar al alumnado material lingüístico que se encuentre ligeramente por encima de su nivel de adquisición para que sea capaz de comprenderlo y asimilarlo pero consideramos especialmente importante la noción de filtro afectivo que ayuda a activar la atención y receptividad cuando el discente experimenta un sentimiento de confianza en sí mismo. En nuestra opinión, este filtro se activa tras la recepción de FB.

De forma simultánea, los estudios sobre relación social llevados a cabo por Vygotsky (1978) mostraron al niño como ser social que necesita relacionarse con otros seres vivos para aprender. Cobra un papel especialmente relevante para SLA la interacción verbal (Long, 1985). Por otro lado, no sólo se contempla la dimensión cognitiva y social sino también la afectiva, insistiendo en el aprendizaje individualizado y en las múltiples inteligencias que un individuo puede tener (Gardner, 1983). El aprendizaje se halla influenciado por las emociones (Goleman, 2001) y la autoestima juega un papel determinante (Corkille Briggs, 1970). Mackey (2006) advierte de la importancia de la atención selectiva, mostrar atención y ser consciente del error se consideran dos procesos cognitivos imprescindibles en SLA.

Estudios específicos de FB como el de Brrokhart (2008) ponen de manifiesto cómo la información facilitado por el docente junto a la que el alumno posee le ayuda a éste a saber dónde se encuentra con respecto a sus objetivos de aprendizaje y a decidir cómo avanzar. Sin duda alguna, factores como el esfuerzo personal son imprescindibles para estar alerta a los errores e implicarse en la corrección de los mismos (Chaudron, 1993).

La mayoría de los estudios existentes sobre feedback se han centrado en la evaluación del profesor y en niveles universitarios, analizando las correcciones e imprecisiones del profesor (Allwright, 1975); centrándose en lo que se corrige (Chaudron, 1993); si la corrección es o no atendiendo al nivel de interlenguaje del alumno (Allwright y Bailey, 1991; Johnston,1995) o a la autoevaluación (Cameron,2001). Igualmente, se persigue hallar formas más eficaces de proporcionar FB (Hattie y Timperley, 2007), indagando en las causas que provocan el mismo así como pautas para saber administrarlo eficazmente (Brookhart, 2008). Recientemente se aprecia un mayor énfasis en la necesidad de aprender a administrar y recibir FB (Stiggins, 2007) y en la estrecha relación existente entre emociones y factores cognitivo-sociales para adquirir mayor competencia comunicativa en L2 (Bown y White, 2010).

Es aquí donde situamos nuestro estudio intentando descubrir la posible influencia que los sentimientos surgidos tras el FB puedan ocasionar en el aprendizaje de LE.

2. Objetivos

El principal objetivo de este estudio es la investigación de la transmisión de *feedback* en el aula de Lengua Extranjera, centrándose en aspectos tan relevantes como los tipos de FB que tienen lugar en clase; la relación existente entre FB y ratio así como la cantidad y calidad de FB recibido. Creemos que un análisis profundo nos conducirá a conocer la influencia que el FB recibido ejerce sobre el aprendizaje, especialmente sobre la competencia comunicativa oral del alumnado de Lengua Extranjera.

3. Metodología

Esta investigación se ha llevado a cabo con las siguientes fases:

1. * Selección de participantes:

Hemos de comenzar señalando que, dada la reticencia del profesorado y de la dirección de numerosos centros educativos, no ha sido fácil encontrar grupos de participantes para ser observados y video-grabados. No obstante, hemos tenido acceso a la observación y grabación en sistema audio de dos grupos de alumnos correspondientes a 6º nivel de Primaria, con edades comprendidas entre 11/12 años. Ambos grupos son dependientes de la comunidad autónoma andaluza aunque geográficamente distanciados entre sí y en contextos diferentes. G1, con 22 alumnos, está ubicado en una ciudad y G2, con 10 alumnos, en una aldea. Este segundo grupo está habitualmente integrado en un grupo más numeroso con alumnos de niveles educativos inferiores.

* Observación directa y audio grabación

2. * Trascripción de dos clases de inglés como L2 de cada grupo.
3. * Análisis cuantitativo y cualitativo del corpus lingüístico obtenido.
4. *Resultados obtenidos.
5. * Conclusiones

4. Resultados obtenidos

4.1. Análisis cuantitativo

Centramos el análisis cuantitativo en primer lugar en los miembros de la clase que han realizado actos de continuación. A continuación, hemos investigado diversos aspectos como la cantidad de actos comunicativos que se han efectuado, el idioma utilizado y el momento de la clase en el que han tenido lugar.

4.1.1. Actos de continuación (AC) realizados

Para hallar la cantidad de AC realizados hemos contabilizado todos los actos de continuación que han tenido lugar en clase independientemente de quien los haya realizado, hallando notables diferencias entre ambos grupos. La cantidad de AC realizada en el total de la muestra de G1 es 42 y en G2 97, lo cual supone el 5,73% del DC en G1 y el 10,26% en G2. Por sesiones de clase se observa que la cantidad de AC presente en G2 es muy superior a la de G1. En la primera sesión de G2 hay el triple de AC que en G1 y en la segunda esta diferencia alcanza prácticamente el doble. (G1= 22; G2=37).

Finalmente en este apartado hay que destacar el hecho de que el discurso comunicativo generado por los hablantes en su totalidad en G1 (945 AC) sea muy superior al generado en G1 (733).

4.1.2. Emisores de Actos de Continuación

Investigamos los movimientos comunicativos de continuación verbales como no verbales efectuados por todos los miembros de la clase en las dos sesiones analizadas de cada grupo.

4.1.2.1. Actos Comunicativos de Continuación del alumnado

En primer lugar, contabilizamos la cantidad de actos de continuación realizados por los discentes ya sean de carácter verbal (Fs) como no verbal (NVFs) y lo que ello supone dentro de DC. como se muestra en la siguiente tabla.

Tabla 1 Proporción de discurso comunicativo de los discentes destinado a realizar Actos Comunicativos de Continuación

	Grupo 1				**Grupo 2**			
	G1-S1	**% G1-S1**	**G1 S2**	**% G1 S2**	**G2 S1**	**% G2 S1**	**G2 S2**	**% G2 S2**
Fs	3	3,33%	9	4,50%	4	1,49%	1	0,83%
NVFs	0	0	3	1,50%	9	3,35%	1	0,83%
Total	3	3,33%	12	6,00%	13	4,83%	2	1,67%
Discurso Comunicativo del alumnado	90		200		269		120	

Se desprenden varias observaciones. En primer lugar, apreciamos que la contribución al discurso es muy dispar entre las sesiones de los grupos, es decir, que los mismos alumnos participan más activamente en unas clases que en otras. Por otro lado, capta la atención el reducido número de actos comunicativos de continuación hallados en todas las sesiones dado que en el mejor de los casos alcanza el 6% de la contribución de todo el alumnado al discurso.

En la tabla 2 se muestran las preferencias del alumnado para emitir actos verbales de continuación o no verbales, observándose un alto porcentaje de emisiones verbales en G1.

Tabla 2.Actos de Continuación Verbales y No Verbales del alumnado

	% G1-S1	% G1 S2	% G2 S1	% G2 S2
Fs	100%	75%	30,77%	50%
NVFs	0	25%	69,23%	50%

A nuestro juicio, los datos correspondientes a G2 presentan mayor aproximación a la realidad conocida a través de nuestra experiencia docente. Probablemente si hubiéramos podido hacer video grabaciones las NVFs habrían alcanzado valores más altos pero el sistema audio nos ha limitado en este sentido.

4.1.2.2. Actos Comunicativos de Continuación del profesorado

Con respecto al comportamiento de las docentes, los datos obtenidos se muestran en los Figuras 1 y 2.

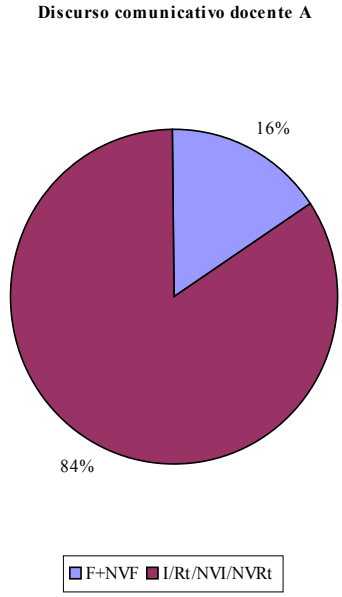


Figura 1. Distribución del discurso comunicativo de docente A

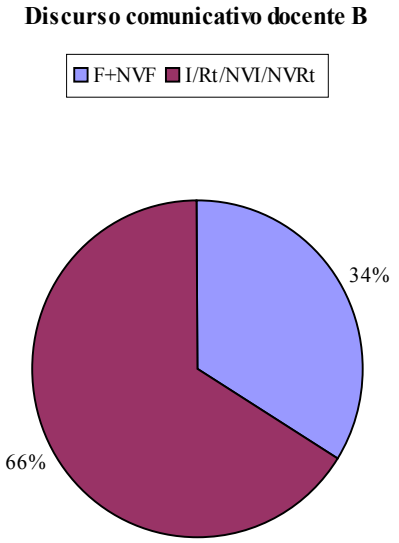


Figura 2. Distribución discurso comunicativo de docente B

Se aprecia que el porcentaje de actos comunicativos realizados en el movimiento de continuación (ACC) es desigual entre el profesorado. Del total de clases analizadas el 16% del discurso comunicativo (DC) de la profesora A se corresponde con actos correspondientes al movimiento de continuación, ya sea de carácter verbal o no verbal frente al 34% de la profesora B. En la tabla 3 se especifican los actos comunicativos de las docentes por sesiones.

Tabla 3. Actos Comunicativos de Continuación (ACC) de docentes por grupos y sesiones

Docente	Sesión	ACC	DC docente	% ACC del DC docente
A	G1-S1	17	134	12,7%
	G1-S2	10	309	3,2%
B	G2-S1	47	194	24,2%
	G2-S2	35	362	9,7%

Destaca la notable diferencia entre los porcentajes de cada docente. Los correspondientes a B son muy superiores a los correspondientes a A, duplicando el porcentaje que ésta obtiene en la primera sesión y triplicándolo en la segunda.

Otro dato relevante que tiene lugar en ambos grupos es la notable diferencia entre las sesiones de cada docente. No parece existir en el profesorado uniformidad al ejecutar ACC pues observamos que los resultados de A en su primera sesión triplican el porcentaje obtenido en la segunda, siendo éste prácticamente insignificante (3,2%). De forma similar sucede en el caso de B.

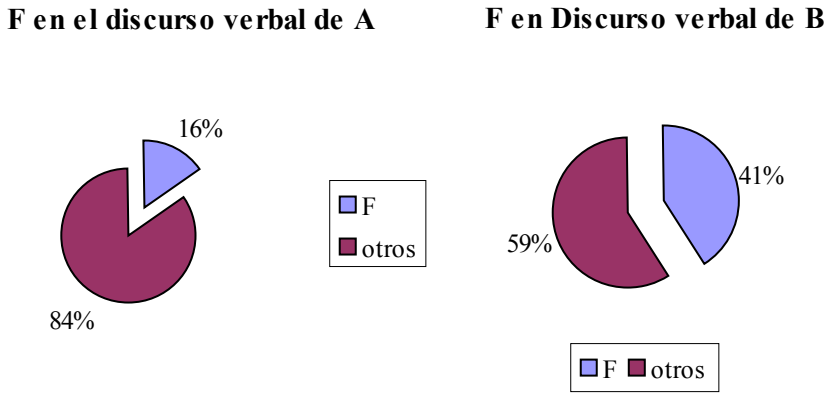
Profundizando en el análisis hacemos distinción entre los Actos de Continuación Verbal (F) y los No Verbal (NVF) observando los datos de la Tabla 4.

Tabla 4. Actos de Continuación Verbal y No Verbal del profesorado

	A-S1	A-S1	A-S2	A-S2	B-S1	B-S1	B-S2	B-S2
F	15	88%	10	100%	43	91%	33	94%
NVF	2	12%	0	0%	4	9%	2	6%
Total ACC	17	100%	10	100%	47	100%	35	100%

Analizados todos los actos comunicativos de continuación realizados por las docentes descubrimos una clara opción por la forma verbal siendo los niveles muy elevados alcanzando el 100% en la sesión G1-S2.

Nos cuestionamos cuál es la parte del discurso verbal de cada docente dedicado a Actos de Continuación. Tomando como marco de referencia exclusivamente las intervenciones orales de las docentes durante las dos sesiones analizadas de cada grupo hemos hallado los datos mostrados en los Figuras 3a y 3b.



Figuras 3a y 3b. Actos Verbales de Continuación (F) en el discurso verbal docente

Distinguimos una clara diferencia en la cantidad de actos verbales de continuación emitidos dado que el porcentaje alcanzado por B casi triplica el de A. En la figura 4 se recoge la distribución del discurso verbal distinguiendo entre Actos Verbales del movimiento de Iniciación (I), del de Respuesta (Rt) y del de Continuación (F).

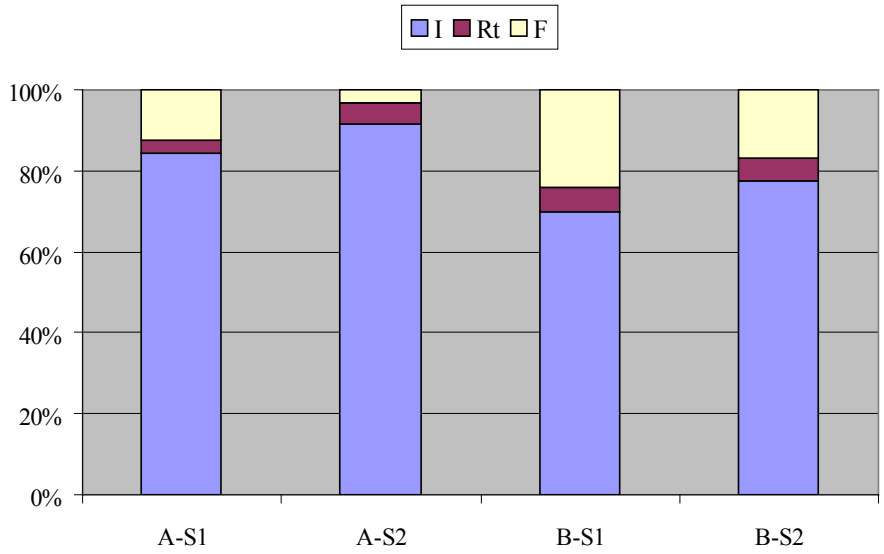


Figura 4. Distribución del discurso verbal del profesorado por sesiones

Se advierte que ambas profesoras realizan fundamentalmente iniciaciones verbales. En segundo lugar, se hallan los actos de continuación verbal y finalmente observamos la respuestas en menor medida.

4.1.3. Idioma seleccionado

Por lo que respecta al idioma, los datos reflejados en la siguiente figura muestran una notable diferencia entre ambas profesoras, siendo L2 la opción más utilizada por B, especialmente en su primera sesión. Sin embargo, en el caso de A en una sesión ha empleado la lengua materna prioritariamente y en la segunda la lengua meta.

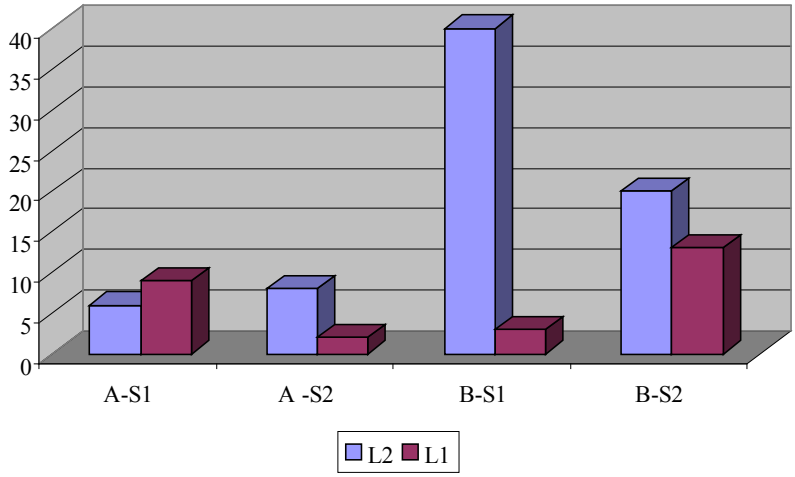


Figura 5. Lengua utilizada para emitir actos de continuación

4.1.4 Episodios con Actos de Continuación Verbal (F)

Aunque todas las clases de LE tienen por objetivo la enseñanza de la lengua inglesa no siempre se realizan las mismas actividades para alcanzar dicho objetivo. Hemos tipificado los diversos momentos educativos presentes en las sesiones así como el porcentaje de F emitidos durante el desarrollo de dicho episodio pedagógico.

Tabla 5. Actividades realizadas por alumnado durante emisión de F

Episodio pedagógico	**A-S1**	**A-S2**	**B-S1**	**B-S2**
Práctica oral	23,53%	83,33%	76,60%	
Audición y comprensión auditiva	58,82%			32,35%
Lectura y comprensión lectora			23,40%	38,24%
Comprobación Teoría (pronunciación de vocabulario)				29,41%
Corrección oral ejercicios escritos		16,67%		
Práctica escrita	11,76%			
Planteamiento tareas casa	5,88%			

El análisis de este corpus muestra que la mayor cantidad de F tuvo lugar mientras los alumnos se encontraban haciendo actividades de práctica oral u otras que requerían la participación oral del alumnado, como por ejemplo actividades relativas a Audición y comprensión auditiva, lectura y comprensión lectora o en Comprobación Teoría (pronunciación de vocabulario*)*. Escasos son los F producidos durante la realización individual de actividades escritas.

4.1.5. Extensión de los actos de continuación verbal

Finalmente, el análisis cuantitativo se ha centrado en la extensión de los enunciados que conforman dichos actos de continuación. En la tabla 6 se advierte cómo la mayoría contiene menos de 5 palabras.

Tabla 6. Extensión de actos de Continuación Verbal

Palabras/F	G1-S1	G1-S2	G2-S1	G2-S2
Una	0	1	19	10
Dos	4	1	13	5
Tres	2	1	1	4
Cuatro	3	0	0	3
Cinco	0	1	1	1
Seis	4	0	0	2
Siete o más	1	1	1	2

La brevedad parece ser un rasgo característico de estos enunciados, destacando el hecho de que el 91,43% de los F emitidos por B en G2-S1 contienen una o dos palabras.

Con respecto a quién van dirigidos hemos podido constatar que el profesorado lo dirige íntegramente al alumnado, ya sea a nivel individual o al grupo. De igual forma, los escasos actos de continuación emitidos por los alumnos (Fs), se dirigen al profesor siendo relativos al proceso de enseñanza.

Una vez realizado el estudio cuantitativo sobre el vehículo de trasmisión de FB, nos parece imprescindible realizar un análisis cualitativo para descubrir el FB que percibe los destinatarios.

4.2.-Análisis cualitativo

Estudiamos en detalle, desde la perspectiva del discente, la retroalimentación suministrada para conocer lo que realmente percibe el alumnado.

4.2.1. Feedback percibido

El corpus lingüístico analizado muestra que a través del lenguaje, el para-lenguaje y la kinesia los alumnos han recibido una información sobre su trabajo y/o actuación en clase. Los alumnos han percibido fundamentalmente dos tipos de FB, uno de carácter positivo y otro de carácter negativo, con dos niveles en cada uno. El eslabón más bajo en esta escalera de cuatro peldaños, ocupa máximo nivel de carácter negativo, el *rechazo (F(refusal*)). En un nivel superior se halla la no aceptación, (F(no ack)), seguido del nivel más bajo del FB positivo, la *aceptación (F(ack))* y finalmente la *aprobación (F(endorsement))* en

el grado positivo más elevado. Hemos agrupado la cantidad de actos comunicativos de aprobación y de aceptación para obtener el FB positivo, pues a través de estos tipos de actos el alumno percibe que su línea de actuación es correcta y su aprendizaje progresa adecuadamente. En otro gran grupo entran los actos que denotan no aceptación y rechazo.

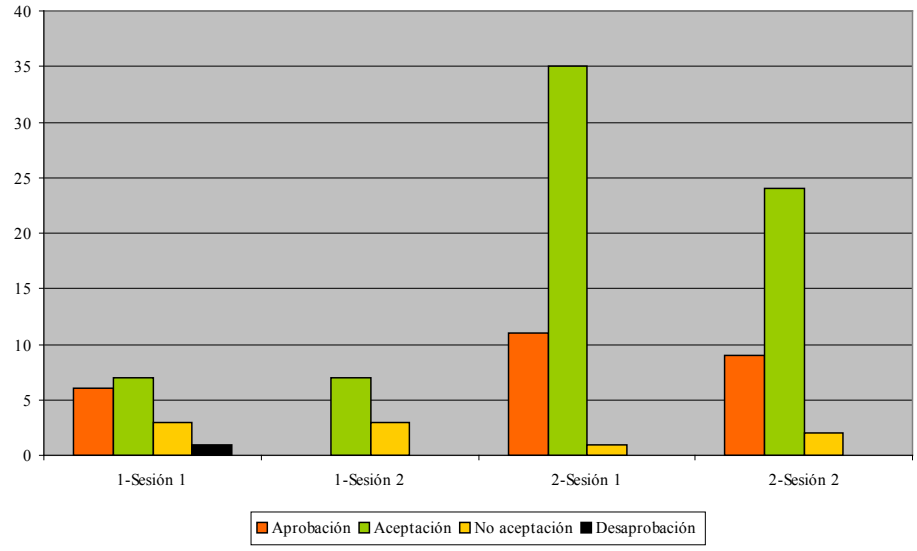


Figura 6. Feedback percibido

El tipo de FB recibido por ambos grupos difiere en gran medida aunque la aceptación es el acto más utilizado. El 76,47% del FB recibido por el alumnado de G1-S1 es positivo frente al 23,52% de carácter negativo. La segunda sesión de este grupo muestra valores muy similares con el 70% de FB+ y el 30% de FB-. Sin embargo, los valores obtenidos por G2 alcanzan cotas más elevadas, hallándose en su primera sesión 97,87% de FB+ frente al 2,13% FB negativo y en la segunda sesión el 94,29% de FB+ y sólo un 5,71% es FB negativo.

4.2.2. Efectos del FB percibido sobre la producción oral del alumnado.

Cuestionados sobre los posibles efectos que el FB percibido pudiera ocasionar en la producción oral del alumnado estudiamos las intervenciones orales realizadas.

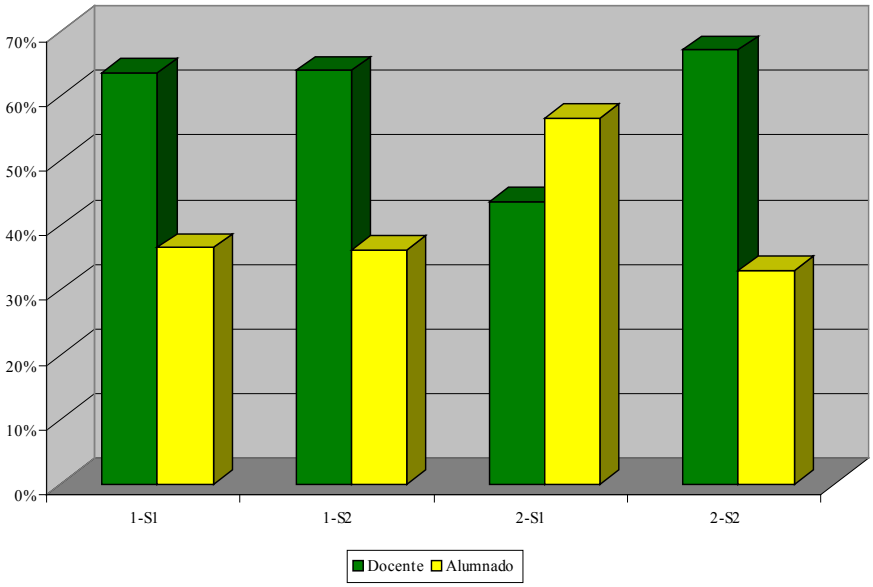


Figura 7. Producción oral por sesiones

En todas, salvo en una sesión, las profesoras contribuyen al discurso verbal en mayor medida que el grupo de alumnos. Es precisamente en G2-S1, la sesión donde los alumnos han recibido mayor cantidad de FB positivo, donde el alumnado ha contribuido al discurso en mayor medida.

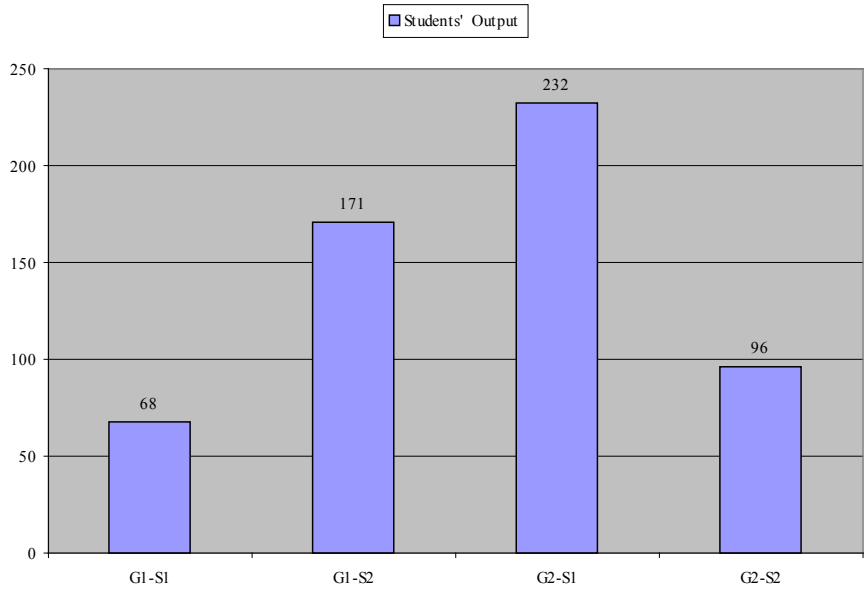


Figura 8. Producción oral del alumnado

Se aprecia que G2 es el grupo con mayor producción oral siendo muy significativa la diferencia entre sesiones. El *output* de G2-S1 triplica a G1-S1.Es posible comprobar no sólo que el grupo que ha participado verbalmente en mayor medida se corresponde con aquel que ha recibido mayor cantidad de FB positivo sino también que la producción oral en lengua extranjera es mayor, superando en más de 60 actos verbales al siguiente grupo, según refleja la Figura 9.

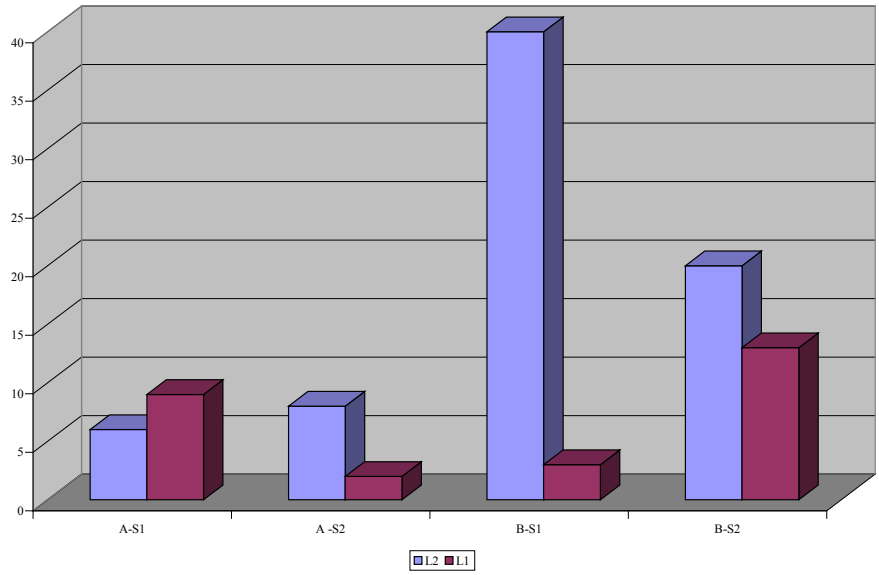


Figura 9. Idioma en *output* del alumnado

Asimismo, hemos constatado que el alumnado de G2 ha emitido hasta ocho veces más actos verbales en L2 que los de G1.

5. Conclusiones

En primer lugar, creemos que esta investigación pone de manifiesto la enorme utilidad que el análisis del discurso aporta al estudio del proceso de enseñanza-aprendizaje de LE al mostrar la posibilidad de investigar el *feedback* transmitido en clase mediante el análisis del discurso comunicativo generado por todos los miembros de la misma.

Parece evidente que tanto alumnos como profesorado proporcionan FB en todas las sesiones aunque el docente en mayor medida. Por otro lado, esta transmisión se canaliza en la interacción a través del movimiento de continuación, mediante actos de carácter verbal o no verbal. Hemos constatado que los alumnos utilizan preferentemente la vía no verbal y escasamente la verbal y sucede al contrario en el caso de los docentes.

Asimismo, existe falta de homogeneidad en la cantidad de actos de continuación suministrados por el profesorado, lo cual sugiere que su emisión esté estrechamente

relacionada con la interacción de éste con el grupo-clase. Hemos advertido que el FB suele aparecer especialmente durante la realización de actividades que requieran la participación oral del alumnado.

El análisis cualitativo nos conduce a concluir que el alumno puede percibir dos tipos de *feedback:* uno de carácter positivo u otro de carácter negativo con dos niveles de concreción cada uno: *aprobación y aceptación* versus *no aceptación y rechazo.* Los sentimientos despertados por el tipo de FB percibido influyen en su participación en clase y en su producción oral. La percepción de *feedback positivo* provoca sensación de bienestar, de progreso en el aprendizaje que eleva la autoestima del alumno impulsándole a superar el reto planteado. En cambio, la única percepción de *feedback negativo* le desanima, disminuyendo el esfuerzo y la voluntad de participación en LE. No queremos sugerir que el alumnado deba recibir única y exclusivamente FB positivo sino que lo deseable sería minimizar los efectos negativos del FB- alternando ambos tipos de FB para poder recuperar la confianza en sí mismo rápidamente.

Dada la importancia que la información suministrada por el profesor así como por sus compañeros sobre su actuación en L2 tiene para el alumnado, creemos imprescindible la educación sobre formas de administrar y de recibir FB adecuadamente. Coincidimos con Stiggins (2007) y Brookhart (2008) al sugerir la necesidad de aprender a evaluar para aprender en lugar de evaluar lo aprendido. Igualmente, el alumno se enriquecería al ser capaz de valorar y aceptar el FB procedente no sólo del docente sino de cualquier miembro de la clase con la finalidad de estimular su aprendizaje y su participación oral en LE.

6. Referencias bibliograficas

Allwright, R. L. (1975) Problems in the study of the language teacher's treatment of learner error in M.K. Burt y H.C. Dulay (eds). On TESOL' 75 *New directions insecond language learning, teaching and bilingual education: 96-109*. Washington D.C..TESOL.

Allwright, D. y Bailey, K. (1991) *Focus on the language classroom*. Cambridge. Cambridge University Press.

Bown, J. y White, C. (2010) A social and cognitive approach to affect in SLA. *IRAL* 48, 331-353

Briggs, D. C. (1970) *Your Child Self-Esteem*. Nueva York. Doubleday comp. (1997) El Niño Feliz: su clave psicológica (16ª edición) Traducción de Oscar Muslera. Barcelona. Gedisa.

Brookhart, S. (2008) *How to give effective feedback to your students*. Association for Supervision & Curriculum Development.

Cameron L. (2001) *Teaching Languages to Young Learners*. Cambridge. Cambridge University Press.

Chaudron, C. (1993) *Second Language Classrooms.* Cambridge. Cambridge University Press.

GARDNER (1983) *Frames of Mind: the Theory of Multiple Intelligencies*. Basic Books.

GOLEMAN (2001) *Inteligencia Emocional*. (43ª ed.) Barcelona: Kairós

HATTIE, J. Y TIMPERLEY, H. (2007) The power of feedback. *Review of Educational Research*, 77, 81-112

JOHNSON, K. (1995) *Understanding Communication in Second Language Classrooms*. Cambridge. Cambridge University Press.

KRASHEN, S. (1981) *Second Language Acquisition and Second Language Learning*. Oxford: Pergamon**.**

LONG, M. H. (1977). Teacher feedback on learner error: mapping cognitions. In Brown, H. D., Yorio, C. A., & Crymes, R. (eds.), *On TESOL '77*. *Teaching and learning English as a Second Language*: *Trends in research and practice* (pp. 278-94). Washington, D.C.: TESOL. Reprinted in Robinett, B. W., & Schachter, J. (ed.), *Second language learning*: *Contrastive analysis, error analysis*, *and related aspects* (pp. 446-65). Ann Arbor: University of Michigan Press, 1993.

LONG, M. (1985) Input and second language acquisition theory, en S. Gass y C. Madden (eds.) *Input in Second Language Acquisition*. 377-93. Rowley, MA: Newbury House.

MACKEY, A (2006) Feedback, noticing and instructed second language learning. *Applied Linguistics* 27/3 405-430.

MEHRABIAN, A. (1971). *Silent messages.* Belmont, CA: Wadsworth.

MEHRABIAN, A. (1971). Nonverbal betrayal of feelings. *Journal of Experimental Research in Personality, 5,* 64-73.

POYATOS, F. (1994) *Paralenguaje, kinésica e interacción*. Madrid: Istmo.

SINCLAIR, J. MCH Y COULTHARD, M (1975) *Towards an analysis of discourse*.Oxford. Oxford University Press.

STIGGINS, R.J. (2007) Assessment through the students' eyes. *Educational Leadership* 64 (8), 22-26.

THORNDIKE, E. L. (1913) *Educational psychology*. Volume 1: the original nature of man. New York. Columbia University, Teachers College.

TSUI, A.B.M. (1994) *English Conversation*. Oxford. Oxford University Press.

VYGOTSKY, L. S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press

Usos y aplicaciones específicas de la lingüística de corpus

Testing the Exception: An Analysis of Eminem's Language Uses from a Corpus-based Approach.

Pedro Álvarez Mosquera
University of Salamanca

The presence of internationally known rapper Eminem in the hip-hop scene has been controversial since his beginnings in the late 90's. His exceptionality as a Caucasian individual in a predominantly African American genre is reflected in the number of records sold and the support from influential figures in the hip-hop world. In our study, by maintaining a sociolinguistic approach, we have used Wordsmith Tools *to analyze Eminem's language uses in his album,* The Marshall Mathers LP *(2000), and compared them to contemporary African American rapper Jay Dilla's album,* Welcome 2 Detroit *(2001)***.** *The analysis of these two similarly sized corpora from both rappers who belong to the same age group, city of origin and gender, allowed us to place ethnicity and language at the center of this study. Our results show that their language uses present significant similarities in relevant aspects of rap related to the communicative role of the African* griot *in the African American tradition, while some important differences were also noted.*

Keywords: *Ethnic corpora, authenticity, Wordsmith Tools, Hip hop studies, African American linguistic patterns*

La presencia del mediático rapero Eminem en el mundo del rap, ha sido controvertida desde su aparición en los años 90. Su pertenencia al grupo étnico blanco en un género musical marcadamente afroamericano ha hecho de su éxito una auténtica excepción, ganándose el apoyo y respeto de importantes personalidades del mundo del rap. En nuestro estudio, mediante el uso de Wordsmith Tools*, mantenemos una aproximación sociolingüística para analizar los usos lingüísticos de Eminem en su álbum* The Marshall Mathers LP *(2000) y los comparamos a los presentados por el rapero afroamericano Jay Dilla en su trabajo discográfico* Welcome 2 Detroit *(2001). El análisis de dos corpus de tamaño equivalente, pertenecientes a dos artistas coetáneos que tienen el mismo lugar de origen, grupo de edad y género, permite situar al elemento étnico diferenciador y al componente lingüístico en el corazón de nuestro estudio. Los resultados obtenidos indican que ambos artistas presentan similitudes significativas en aspectos intrínsecamente relacionados con el papel comunicativo de la figura del* griot *africano en la tradición afroamericana, aunque también se detectan importantes diferencias.*

Palabras clave: *Corpus étnicos, autenticidad, Wordsmith Tools, Hip hop studies, patrones lingüísticos afroamericanos.*

1. Introduction

Eminem is the most famous non-African American rapper. Since the 90's, his way of rapping has been under scrutiny by other rappers, critics and, relevant figures in the rap scene. Considered by many acclaimed African American rappers as a true exception, the reason why he evokes so much controversy stems from the combination of his ethnicity, social origins, usage of cultural and linguistic patterns associated with the African American community, and his own talent. A brief overview of the history of rap reveals a strong connection between this music genre and African Americans (Smitherman, 2000: 275). This strong relationship has been preserved through oral and cultural ethnic patterns, the use of AAVE (*African American Vernacular English*) features, and a strong concept of authenticity. This ethnic component in rap music has undermined the chances of success for many non-African American rappers, leading them to display or perform features associated with blacks to overcome this burden. In fact, there have been noticeable examples, such as Caucasian rapper *Vanilla Ice* (Robert Van Winkle), who invented his origins and past in the ghetto (another site of authentification), revealing the need to validate his status and presence in this music genre (Rose, 1994: 11-12).

When it comes to language choices, this need for validation might lead non-African rappers to reproduce AAVE features, triggering a linguistic phenomenon called *language crossing* (Rampton, 1995). Rampton has shown that *crossers* present important limitations at appropriating linguistic features, and Caucasian rappers are not an exception when it comes to rap (Cutler, 1999; Alvarez-Mosquera, 2009). Importantly, Rampton also emphasized "effective crossing relied on the biographically contingent social and interactional competence and sensibility of particular individuals" (1995: 193). It is precisely the biographical component what has been considered as the source of differentiation between Eminem and other white rappers: "Eminem's upbringing, among poor blacks and whites, was closer to the black and minority experience reflected in rap, lending Eminem the added credibility of similarity. Eminem, without a doubt, is the first white rapper with true street cred to cross over" (Bozza, 2003: 93). His case resembles other noticeable cases described in previous academic works. For instance, Sweetland (2002) notes that a Caucasian teenager is described by the African American community where she grew up as *basically black*. Therefore, it seems that a particular process of socialization can extend the linguistic options over ethnic boundaries, distinguishing Eminem from other Caucasian rappers. In other words, "Eminem is the product of a white background as well as black culture, and he was alienated from both groups when he was growing up" (Bozza, 2003: 172). In fact, in the overwhelmingly African American ghettos of Detroit, the underprivileged social status of AAVE is inversed and most people in this area are easily exposed to this language variety and other cultural patterns associated with the African American community (Alim, 2006; Brewer, 1991; Morgan, 2001: 194; Rose, 1994; among others). As a figment of the socialization process and his own talent, Eminem has won the acceptance and respect of salient African American rappers and critics: "[t]here have been people fucking with hip hop and black culture before, but I don't think anyone has ever gotten the one hundred percent stamp of approval […] he is the first dude to get it" affirms Sasha Frere-Jones (*Village Voice* magazine) (Bozza, 2003: 178). If fact, in July 2000, Eminem was the first Caucasian to appear on *The Source*

magazine's cover. Therefore, there seem to be enough reasons to lead us to believe that Eminem represents a true exception in the hip-hop world. However, is he so exceptional at the linguistic level?

2. Materials and methodology

In order to obtain an objective answer about Eminem's language uses as a sign of his authentic status in the hip-hop arena, we used *Wordsmith Tools* to measure and contrast them in comparison with African American rapper Jay Dilla's words. Therefore, we are trying to rely more heavily on empirical data to describe and evaluate Eminem's language behavior, rather than relying on mere opinions. After considering many rappers, we chose Jay Dilla because he is from the same city, and belongs to the same age group and gender, as Eminem, limiting the influence that these salient social variables could have on the obtained data, and allowing us to place ethnicity at the center of our analysis. Eminem's corpus is formed by more than 4,000 words contained in 7 songs from his album *The Marshall Mathers LP* (2000). As for Jay Dilla, he incorporates many other rappers and includes instrumentals in his album *Welcome 2 Detroit* (2001), so we found it necessary to add some songs from his following CD *Champion Sound* (2003), to reach an equivalent corpora in terms of tokens (over 4,000 words). Also, it is important to note that we only took the chorus into account one time, in order not to taint our data. For instance, if one of the rappers includes the word *killer* a couple of times in a chorus that may be repeated over ten times in a song, this might increase the prevalence of this term in the study, moving it to the top of wordlist provided by *Wordsmith Tools*. Likewise, we omitted any comment or indication that relates to the format of the song, such as the title, chorus, and so forth. For further clarification, we only analyzed the words spoken by both rappers while performing. Finally, by listening to every song and taking into account the information provided by the CDs and the website *Ohhla.com*[200] (*Original hip hop lyrics archive*), we also removed the parts of the songs sung by other artists who collaborated with these two rappers in their albums.

3. Corpus analysis

As we mentioned earlier, in order to observe these two corpora from a sociolinguistic point of view, *Wordsmith Tools* appeared to be the most useful methodological software. This paper only displays a few of the various ways that this program can be utilized. In our case, we processed our corpora through *Wordlist* to obtain the list of words contained in each of the two corpora, organized in terms of frequency (the number of times a particular word has been used). By adding the same *Stopword* list, containing around fifty grammatical words (e.g. the, a, to, of, etc.) to both corpora,[201] and paying close attention to the existence of lemmas, we were able to observe and establish some relevant linguistic patterns that we will discuss after presenting some general results:

200 This website has been used by other authors, such as Morgan (2002), in their studies.

201 These grammatical words are irrelevant for our study. In fact, their high index of frequency would have placed them on the first positions in our lists, making it more difficult to conduct an accurate analysis.

Table 1. WST analysis for rapper Eminem

N	Word	Freq.	%	Texts	%
1	I	213	4,460732937	1	100
2	YOU	164	3,434555054	1	100
3	ME	72	1,507853389	1	100
4	I'M	70	1,465968609	1	100
5	JUST	53	1,109947681	1	100
6	MY	50	1,047120452	1	100
7	YOUR	43	0,900523543	1	100
8	BUT	35	0,732984304	1	100
9	LIKE	35	0,732984304	1	100
10	BE	33	0,691099465	1	100
11	IS	32	0,670157075	1	100
12	DON'T	30	0,628272235	1	100
13	SHIT	25	0,523560226	1	100
14	WHEN	25	0,523560226	1	100
15	GET	23	0,481675386	1	100

Table 2. WST analysis for rapper Jay Dilla

N	Word	Freq.	%	Texts	%
1	YOU	144	3,52854681	1	100
2	WE	62	1,519235492	1	100
3	I	58	1,421220303	1	100
4	LIKE	47	1,151678562	1	100
5	GET	41	1,004655719	1	100
6	ME	37	0,90664053	1	100
7	MY	37	0,90664053	1	100
8	NIGGAS	36	0,882136703	1	100
9	YA	35	0,857632935	1	100
10	YOUR	29	0,710610151	1	100
11	THEY	28	0,686106324	1	100
12	ALL	26	0,63709873	1	100
13	IS	26	0,63709873	1	100
14	DO	24	0,588091135	1	100
15	WANT	24	0,588091135	1	100

A first look at the *Wordlist* results reveals that salient African American patterns associated with rap are equally represented in both corpora. As noted by many authors (Rickford & Rickford, 2000; Smitherman, 2000: 275), rap has its origins in the fertile *Black Oral tradition*, stemming from Africa. Smitherman affirms that rappers act as a "a post-modern African griot, the verbally gifted storyteller and cultural historian in traditional

African society" (Smitherman, 2000: 269). In communities and times before writing was prevalent, the power of words and the role of *griots* were extremely important for their members. In fact, rappers not only must be linguistically and lyrically fluent, but they are also expected to witness what is occurring around them and to tell the truth with wit and cleverness, while still maintaining clarity (Baker, 1993: 91; Costello & Foster, 1990: 96-98). On a preliminary approach, it seems that the two rappers observed in our study reproduce this essential idea.

To expand upon our findings related to these oral patterns, we will cover three main points. First, the fact that the pronouns *I* and *you* are quite prevalent in both rappers' corpora reflects the notion that rappers have a message for their audience that they are trying to convey. More unexpectedly, we observed that the pronoun *I* was used almost four times more frequently by Eminem (213) than Jay Dilla (58). When including lemmas (*I'm*, *I'd*, *I'll*, etc.) the gap increases and confirms this tendency (Eminen: 302; Jay Dilla: 73). Even when looking for other first-person references, such as *me* (Eminen: 72; Jay Dilla: 37) or *my* (Eminen: 50; Jay Dilla: 37), Eminem always presents more instances than African American rapper Jay Dilla. This is a very important outcome, since it inverts the direction of the previous pattern detected in the analysis of Caucasian rappers from the 80's, 90's and 00's (Alvarez-Mosquera, 2010), where African American rappers always used the word *I* more frequently than their white counterparts.

Since rappers address their audiences directly, or interact with the protagonist of their stories from a personal perspective, the pronoun *you* (Eminem: 176; Jay Dilla: 189) is present in the top part of both tables as well. This directly conflicts with the pattern observed in previous cases, where *you* was used less frequently by Caucasian rappers than by African American rappers (Alvarez-Mosquera 2010). This places Eminem closer to the African American standard, in terms of the group's use of this pronoun. In contrast, the last non-lexical word with a relevant role in our study, the pronoun *we*, draws diverging lines that separate both rappers from their respective ethnic groups. As stated earlier the *griot* plays a central role for the community which he belongs (Costello & Foster, 1990: 115), so we expect frequent references to this community as a group. In our study, Eminem follows the *white pattern* detected before, since rapper Jay Dilla uses the pronoun in a much higher number of instances (Eminem 12; Jay Dilla 62). This could illustrate Eminem's difficulty in referring to *his group* in an African American dominated, music world that encompasses rap music (Bozza, 2003: 172).

Continuing with our study, the next item that is relevant to our analysis is the use of the lexical word *nigger*. The controversy surrounding this word makes its study even more relevant from an ethnic point of view. Its significance is due to its complex social implications, which have changed over time. Most notably, *nigger* was rarely used publicly or by the African American community in the past, but nowadays, as Kennedy states, "[b]lacks use the term with novel ease to refer to other blacks, even in the presence of those who are not African American" (Kennedy, 2002: 174). The Caucasian American rappers' eagerness to sound more *black*, or even its use by other ethnic group members as a sign of affection, could have led Eminem to replicate it as well. According to our results, the term *niggas*, ranks in the eighth position in the African American corpus (36

repetitions), while it is not present in Eminem's corpus at all. Actually, the fact that this term is used more frequently in the plural form might reinforce the concept of community, which acts as a message meant for the larger group. The result is quite clear; ethnicity is a critical part of in the rap scene and hip-hop culture. In fact, when taking lemmas into account, this tendency to use the word *nigger* is confirmed by a higher number of instances on the African American side (Eminem: 0; Jay Dilla: 64). Regarding the importance of using certain terms that refer to ethnicity, it is worth noting that Eminem uses the word *wigger*. Although he only uses it once, it is significant because *wigger* is a term used to describe the socially assumed, fake nature of white rappers.

The last linguistic element that this corpus-based approach allows us to explore is the use of violent terminology. In general terms, rap music is believed to have a significant violent component, which has made rap the target of criticism from a wide-range of social groups. This negative view of rap music is detrimental to African Americans, and it is worsened by its increasing presence in the media where: "[rap] has also fulfilled national fantasies about the violence and danger that purportedly consume the poorest and most economically fragile communities of color" (Rose, 1994: 11). To observe the extent that both rappers are consistent with this stereotype, we focused on the use of the following terms: *fight*, *gun*, *murder*, *shot*, *kill*, *dead* (including lemmas). It is our intention to determine if Eminem and Jay Dilla show similar or different rates when using explicitly violent terminology.

Table 3. WST (violent terminology)[202]

EMINEM	Frequency	Position	**JAY DILLA**	Frequency	Position
GUN	5 times	110	GUN	4 times	128
KILL	9 times	61	KILL	4 times	133
DEAD	4 times	136	POLICE	0 times	_
SHOT	2 times	394	SHOT	2 times	347
FIGHT	1 time	686	FIGHT	1 time	597

A simple glance at table 3 demonstrates that violent terminology is used at a higher rate by Eminem. His usage of the words *gun*, *kill* and *dead* are more frequent than Jay Dilla's,

202 In the previous study, we included the term *police* due to a possible association between this group and repression (type of violence). However, no cases of police were found in Eminem or Jay Dilla's corpora. We substituted *police* by the term *dead*.

while they present the same number of instances in the case of *shot* and *fight*. These results correlate with the data obtained for European American rappers in Álvarez-Mosquera (2010), which states that white rappers tend to use a higher number of violent terms in order to sound more *black*. Interestingly, they go further than African American rappers in terms of their usage of this type of terminology in all four cases. Therefore, the data leads us to question the general belief that African American rappers incite violence (at least in terms of frequency), while it creates evidence that supports the case that white rappers overuse violent terminology as a way to resemble more prototypical rappers.

4. Conclusion

Categorizing Eminem as an exception, or a mere product of the music industry in the history of rap, can be done in many subjective ways. With our approach we want to contribute to this debate by presenting objective data on one of the most relevant features of rap: its language. The analysis and comparison of Eminem and Jay Dilla's corpora presents significant similarities and stark differences. On the one hand, Eminem's usage of the pronoun *I* and *you* places him at the same level of his African American counterparts in a particular aspect of rapping that is strongly linked to the roots of this genre: the concept of the African *griot*. None of the rappers studied earlier (Alvarez-Mosquera, 2010) were able to mimic these African American language patterns, which are a core feature of rap. However, Eminem's display of prototypical features goes beyond the African American average when using violent terminology. This pattern, which has been also observed in other white rappers, can be interpreted as a way of demonstrating toughness and achieving authenticity in an environment where they are at an ethnic disadvantage. On the other hand, Eminem's language uses show some important limitations. As we have seen, Caucasian rappers might be able to appropriate or reproduce certain language items (e.g. violent terminology), but displaying ethnicity is much more difficult and controversial. AAVE words and oral patterns are intrinsically interwoven with the *black experience* in the United States, which is defined as "the experience of domination and the hidden transcripts produced in relation to these experiences of domination are culturally coded and culturally specific" (Rose, 1994: 123). In other words, AAVE provides words and oral patterns that are a significant part of the code in the hip-hop scene. For instance, the use of the pronoun *we* to refer to their own group is much more significant in Jay Dilla's corpus. These results not only correspond with the other African American rappers analyzed in Alvarez-Mosquera (2010), but they also underline the importance of their message for their own communities, resembling the traditional African *griot*. This important gap between African American rappers and Caucasian Americans undermines the latter group's chances (including Eminem's) to sound more authentic. Along with the concept of community, the overwhelming use of the word *niggas* in Dilla's corpus, and its absolute absence in Eminem's (and the other three rappers) places ethnicity at the core of rap music. The use of this controversial term by African American rappers draws a significant ethnic line that seems impossible for Caucasian rappers to cross with no social cost.

We are by no means claiming that talent is not needed in order to be a successful rapper. However, language plays a central role in rap and AAVE is still the main point of reference.

Therefore, according to these preliminary results, we can affirm that Eminem's language uses place him closer to African American linguistic patterns than to those exhibited by other Caucasian rappers, which partially confirms his exceptionality in this culture-specific context. However, the fact that Eminem is also unable to reproduce ethnically-marked terms, such as *nigger*, or at expressing cultural patterns like the concept of *community* (replicating Caucasian patterns instead), it illustrates the persistence of ethnic barriers in the hip-hop scene. In order to provide further analysis, we intend to increase the size of our corpora and carry out additional research to explore other language patterns that could help us to confirm, or discard, Eminem's exceptionality as a rapper in an objective way.

5. References

Alim, Sammy H. (2006). *Roc the Mic Right: The Language of Hip Hop Culture*. London: *Routledge*.

Alvarez Mosquera, P. (2009). *El Uso de AAVE por Raperos Blancos: ¿Un Caso Real de Language Crossing?* Unpublished MA dissertation, University of Salamanca.

Alvarez Mosquera, P. (2010). Exploring the use of Wordsmith Tools for sociolinguistics purposes: A case study of cultural loaded language uses in white and black rappers' corpora. En I. Moskowich et al. (Eds.), *Language Windowing through Corpora* (pp. 39-48). A Coruña: Universidade da Coruña.

Baker, H. (1993). *Black Studies, Rap, and the Academy*. Chicago: U of Chicago P.

Brewer, M. (1991). The Social Self: On Being the Same and Different at the Same Time. *Personality and Social Psychology Bulletin*, *17*(5), 475-482.

Bozza, A. (2003). *Whatever You Say I Am: The Life and Times of Eminem*. New York: Three Rivers Press.

Costello, M. & Foster, D. (1990). *Signifying Rappers: Rap and Race in the Urban Present*. New York: Ecco.

Cutler, C. (1999). Yorkville *Crossing*: White Teens, Hip Hop and African American English. *Journal of Sociolinguistics*, *3*(4), 428-442.

Kennedy, R. (2002). *Nigger: The Strange Career of a Troublesome Word*. New York: Pantheon.

Morgan, M. (2001). 'Nuthin' but a G thang:' Grammar and language ideology in hip hop identity. In S. L. Lanehart (Ed.), *Varieties of English around the World: Sociocultural and Historical Context of African American English* (pp. 185–207). Amsterdam: Benjamins.

Rampton, B. (1995). *Crossing: Language and Ethnicity Among Adolescents*. New York: Longman.

Rose, T. *(1994)*. *Black Noise*: *Rap Music* and *Black Culture in Contemporary America*. New England: Wesleyan U P.

RICKFORD, R. & RICKFORD, J. (2000). *Spoken Soul: The Story of Black English*. New York: Wiley.

SMITHERMAN, G. (2000). *Talkin That Talk: Language, Culture and Education in African America*. New York: Routledge.

SWEETLAND, J. (2002). Unexpected but Authentic. *Journal of Sociolinguistics*, *6*, 514-536.

Lexical bundles in US presidential speeches: a corpus-driven study of B. Clinton's, G.W. Bush's and B. Obama's addresses

David Brett

Antonio Pinna
University of Sassari

Patterns of variability in lexical bundles are investigated in a corpus of US presidential addresses. The findings are then compared with those reported in the literature concerning other fields of discourse. The methodology adopted mirrors that of Biber's (2009) study using two corpora: a 4.5-million-word corpus of American English conversation; and a 5.3-million-word corpus of academic prose. The results for the Presidential data display distinct distribution patterns of lexical bundle types when compared with those of the reference corpora. The functions of the most frequent examples of the most fixed pattern are also analysed, showing high proportions of slogans and multi-word collocations.

Corpus-driven, multi-word unit, Presidential speech, lexical bundle

Patterns of variability in lexical bundles are investigated in a corpus of US presidential addresses. The findings are then compared with those reported in the literature concerning other fields of discourse. The methodology adopted mirrors that of Biber's (2009) study using two corpora: a 4.5-million-word corpus of American English conversation; and a 5.3-million-word corpus of academic prose. The results for the Presidential data display distinct distribution patterns of lexical bundle types when compared with those of the reference corpora. The functions of the most frequent examples of the most fixed pattern are also analysed, showing high proportions of slogans and multi-word collocations.

keyword keyword keyword keyword keyword keyword

1 introduction

Corpus linguistic studies have traditionally privileged the investigation of a specific type of Multi-Word Unit (MWU) model, one which is variously known as the *n*-gram (e.g. Stubbs 2007), chain (e.g. Stubbs and Bart 2003), lexical bundle (e.g. Biber *et al.* 1999: 987-1024) or word cluster (e.g. Carter and McCarthy 2006: 828-837). This is a recurrent, continuous sequence of word forms. The most commonly studied form is that composed of four word forms (e.g. Biber *et al.* 2004; Biber and Barbieri 2007; Hyland 2008), as bundles of this length are usually more frequent than longer strings and, at the same time, have a wider assortment of readily recognizable functions than shorter sequences. The studies by Biber *et al*. (2004) and Biber and Barbieri (2007) are particularly important as they identify four main functional roles played by lexical bundle: discourse organization, reference, stance and interaction management. Discourse organizers link prior and forthcoming portions of text; referential bundles identify an entity or a particularly relevant attribute of an entity; stance expressions convey speaker attitude; finally, interactive bundles are typically used to mark politeness or reported speech. These particular functions have been shown to provide a means of differentiation between spoken and written university registers. In particular, Biber and Barbieri (2007: 273, 279) prove that stance expressions are the most frequent in the oral registers (e.g. classroom teaching and class management), while referential bundles are more common in the written ones (e.g. institutional writing and textbooks).

From a theoretical point of view, the central discovery of the studies mentioned above is that these strings, which at first sight may be dismissed as semantically and syntactically incomplete, are characterized by pragmatic integrity, i.e. they demonstrate a considerable degree of functional specialization which may be the main reason behind their remarkable frequency in language use. For example, in their analysis of CANCODE, a 5-million-word spoken corpus, O'Keeffe *et al*. (2007: 70-75) identify a series of pragmatic categories such as discourse marking, hedging, the preservation of face and the expression of politeness, all of which are relevant signals of how the interaction between the interlocutors of a conversation in a given context is unfolding. Biber (2009: 284) summarizes this finding by maintaining that "lexical bundles provide a kind of pragmatic 'head' for larger phrases and clauses, where they function as discourse frames for the expression of new information".

From a methodological point of view, Biber's (2009) approach to the study of MWUs is radically corpus-driven and, following this line of argument, he specifies three general characteristics of this approach: it is based on the analysis of actual word forms (as opposed to lemmas); it only considers sequences of word forms (discarding their syntactic status); finally, it focuses on their frequent, recurrent combinations. Frequency, together with distribution in more than five different texts, is therefore paramount in order to identify lexical bundles, and the result, when interpreted in functional terms, as for the examples shown in Section 3, is far from being a collection of random sequences devoid of linguistic value.

In this study we adopted the methodological approach used by Biber (2009), in which he investigates variability within multi-word units using two corpora: one of American

English conversation and the other of academic prose. The corpus used for the current study is composed of US Presidential addresses and remarks delivered by B. Clinton (1993-2000), G.W. Bush (2001-2008) and B. Obama (2009-2010). As a macro-genre Presidential speeches are monologic texts characterized by being usually prepared to be recited in public. They could therefore be expected to contain features of both written and oral language, possibly tending towards the oral end of the cline. This led us to speculate that our data would fit this picture by showing patterns of variability which positioned Presidential speeches as more or less evenly straddling the oral-written divide as defined by Biber's (2009) findings.

2 MATERIALS AND METHODS

2.1 The corpus of Presidential speeches

Our corpus of Presidential (henceforth P) speeches is composed of addresses by the three most recent U.S. presidents: B. Clinton, G.W. Bush and B. Obama. The data concerning the first two statesmen were divided into two subsections, each corresponding to their respective first and second terms of office. Table 1 provides further details regarding the composition of the corpus.

Table 1. The composition of the corpus of Presidential speeches.

Sub-section	Period	N. Tokens
CLINTON A	Jan 93 – Jan 97	2,168,000
CLINTON B	Jan 97 – Jan 01	4,868,000
BUSH A	Jan 01 – Jan 05	3,459,000
BUSH B	Jan 05 – Jan 09	2,010,000
OBAMA	Jan 09 – Dec 10	1,529,000
TOTAL		14,034,000

As can be seen from Table 1, the number of tokens in each sub-section displays a considerable amount of variation, as CLINTON B, and BUSH A to a lesser extent, are visibly larger than the other sections. However, we decided not to reduce the size of these two sections, as we found that the larger numbers did not appear to lead to a greater occurrence of the phenomenon under investigation. To the contrary, by splitting CLINTON B into one million word sub-sections, we found that the number of instances of lexical bundles that met the 10 instance per million words criteria actually *decreased* as sub-section size increased. To be more specific, the numbers of lexical bundles that met the criteria in the 1, 2, 3 and 4 million word sub-sections were 4076, 3088, 2728 and 2585, respectively. Hence, rather surprisingly, our data suggest that increases in corpus size do not influence the number of lexical bundles per million words directly, if anything there is an inverse relationship.

2.2 Methodology

In order to permit direct comparison between our results and those obtained by Biber (2009), we aimed to mirror as much as possible the methodology adopted by the latter when investigating variability within multi-word units using two corpora: a 4.5-million-word corpus of American English Conversation (henceforth C); and a 5.3-million-word corpus of Academic Prose (henceforth A). Unless stated otherwise the procedures were identical.

Initially, the corpora were searched for 4-grams, discarding sequences with a frequency of less than 10 occurrences per million words. Biber placed a further constraint that each 4-gram be present in at least five different texts. However, the length of the texts was not specified; furthermore, whether these five different texts were to be of different authorship was not mentioned. We decided not to apply this measure for two reasons: only three speakers, with related speech writing teams, were represented, hence attempting to idiosyncrasy appeared pointless; the Presidential texts were organised by semester, and the resulting files were rather large, probably far larger than those analysed by Biber.

Each corpus was then searched for a series of sequences composed of three of the components of each 4-gram, allowing variability in the fourth slot, e.g. *234, 1*34 etc. If the token in a given slot in each 4-gram composed less than 50% of the results for that slot, the slot was deemed to be variable, as opposed to fixed, and marked with an asterisk. If the token constituted 50% or more of the occurrences, the slot was considered to be fixed, and marked with a number. The resulting patterns, identified for each of the selected 4-grams, were then counted, and organised into groups based on pattern similarity. For example, Biber found that variability in one internal slot (1*34/12*4) was six times more common in A than in C, whereas initial and/or final variability patterns (*234/123*/*23*) were roughly twice as frequent in the latter data.

3 RESULTS

3.1 Quantitative analysis

Figure 1 shows the results obtained in our study in comparison with those observed by Biber (2009). The data represented in the graph do not appear to suggest any proximity of P to either of the two registers being used as a yardstick. For example, while the 12**/**34, *23*, 1***/***4 and **** patterns show similarity with A, the 12*4/1*34 pattern type is more similar to C. Furthermore, several pattern types are present in P in proportions totally unlike those of either A or C (1234, 123*/*234, and to a lesser extent *2**/**3*). The high proportion of the pattern which displays the greatest amount of fixedness, 1234, is perhaps hardly surprising, bearing in mind the fact that we are dealing with only three speakers and a limited range of topics.

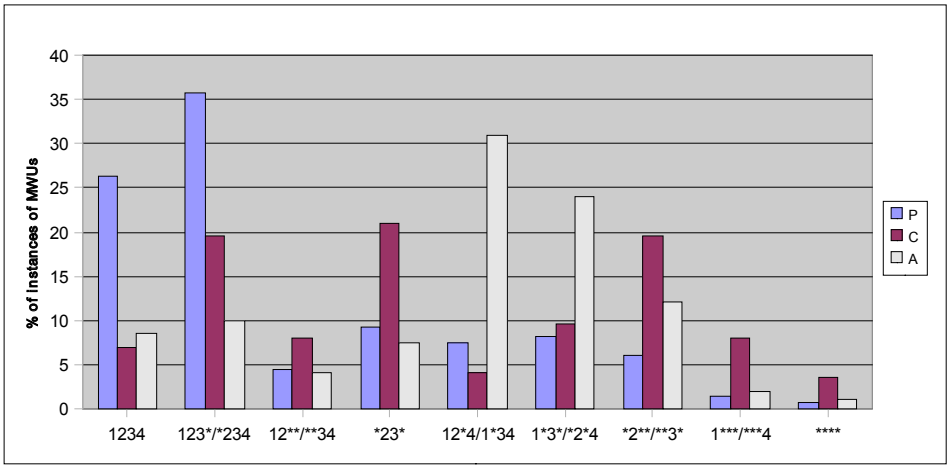


Figure 1. The proportions of MWU variability pattern types in the Presidential (P) data in comparison with Biber's (2009) Conversation (C) and Academic Prose (A) data.

Thus far, the results suggest quite strongly that the P data bear similarity to neither the C nor the A data, and are, on the other hand, equidistant, as it were, from both the latter text types. A further point to be addressed is: if the P text type is indeed distinct from both the reference datasets, forming a separate group, how discrete is this group? How much variation is there within it? Figure 2 shows the proportions of variability types across the sub-sections of the P corpus (note that for the sake of simplicity the results for Clinton A and Clinton B have been merged, as have those for Bush A and Bush B). The similarity in the proportions of the different variability patterns from President to President is nothing short of striking. The only pattern in the dataset which is not practically identical is that of the most fixed pattern, 1234, which constitutes a considerably greater proportion of the total patterns in the Bush sub-section, than in those of his immediate predecessor and successor.

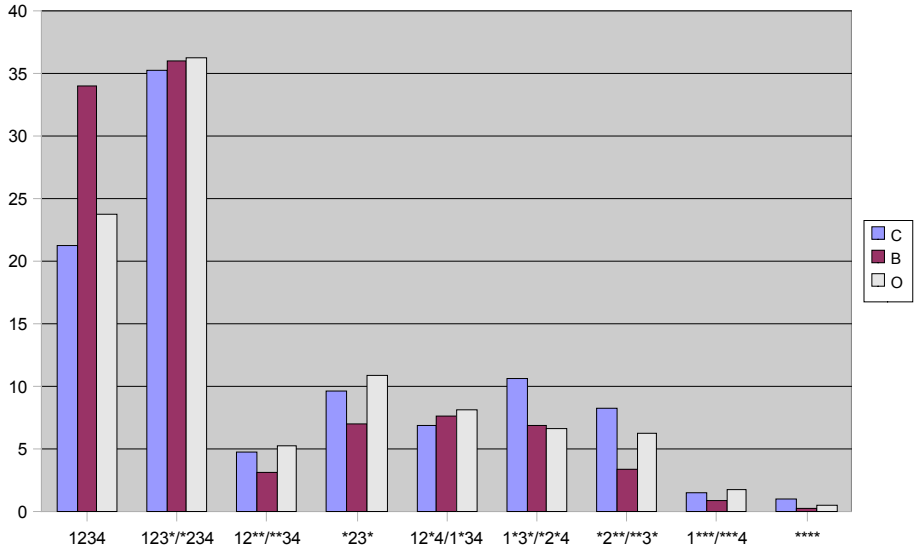


Figure 2. The proportions of MWU variability pattern types in the sub-sections of the Presidential (P) data: C = Clinton A & Clinton B; B = Bush A & Bush B; O = Obama.

Our results hence appear to confirm that variability patterns within lexical bundles differ from genre to genre and therefore their potential as text type identifiers. On this note we return to the data presented in Figure 1, comparing the pattern distributions of P with A and C. Figure 3 presents the same information in a different way, one which is perhaps more effective in underlining the greater or lesser presence of the various pattern types across the different genres.

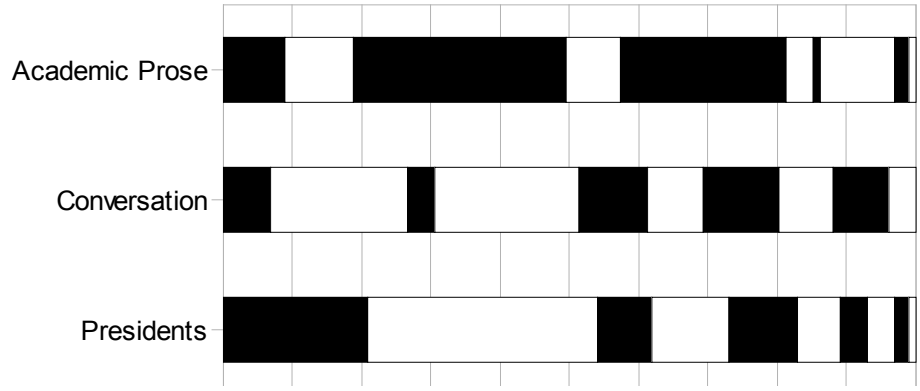


Figure 3. The proportions of MWU variability pattern types in the Presidential (P) data in comparison with Biber's (2009) Conversation (C) and Academic Prose (A) data. Top element in legend corresponds to rightmost element in graph.

3.2 Qualitative analysis

In comparing the proportions of multi-word pattern types in the P, C and A data (Fig. 1), one immediately notices the remarkable difference between Presidential speeches and the other two registers when the most fixed sequence type, 1234, is taken into consideration. In order to explore this in functional terms, the lists comprising the 50 most frequent multi-word units of the 1234 type in Presidential addresses have been studied in their contexts of use and subsequently classified. For reasons of space, we will now focus on data relating to the Bush A sub-section, that which shows the greatest proportion of lexical bundles of this particular type. Fig. 4 shows how these are distributed in the various functional classes, which are illustrated in the Introduction above and follow those identified by Biber (2009).

Our data in Fig. 4 highlight a fundamental distinction between two major categories: multi-word collocations and multi-word formulaic sequences, which should be considered as poles of a continuum (Ibid: 290). The first category consists of multi-word technical terms, typically connected to the topics dealt with in a particular register. Multi-word formulaic sequences, on the other hand, provide a pragmatic frame for the new information that follows.

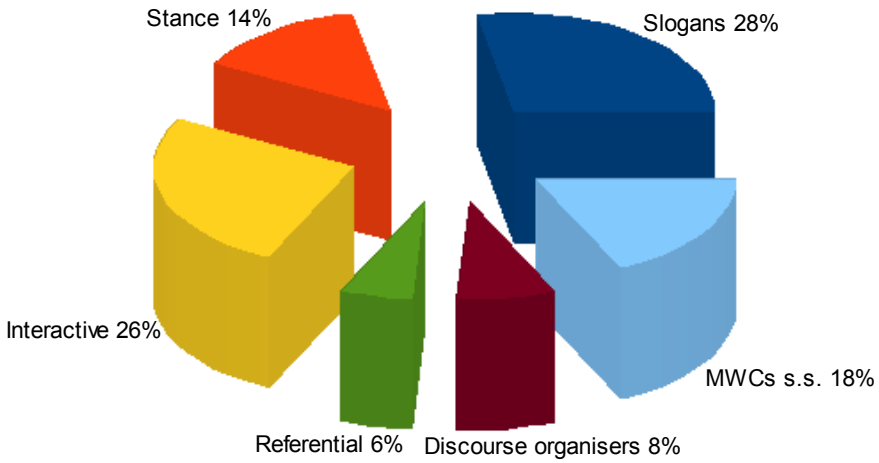


Figure 4. The proportions of functions of the 50 most frequent 1234 patterns in the Bush A subsection.

If we concentrate on the various functions performed by the pragmatic sequences shown in Fig. 4, the relatively high frequencies of the stance and interactive bundles (39% of the total) underline the central role played by the interpersonal aspect of meaning in Presidential addresses. For example, the two most frequent sequences in the Bush A data are: an instance of speaker stance (*I want to thank*) and a politeness marker (*thank you all for*). Their considerable presence with respect to referential and discourse organization bundles reflects the peculiar nature of this type of pre-planned face-to-face

communicative event, where frequent dialogic signals, which are also characteristic of formal conversation, are consistently employed to play down the fundamentally monologic dimension of Presidential speeches in order to create the illusion of dialogue.

Interestingly, as Fig. 4 shows, 46% of the sequences belong to the category of Multi-Word Collocations (MWCs) in the broadest sense. While in conversation and academic prose, these are characterized by lower frequencies than multi-word formulaic sequences (Biber 2009: 209), in Presidential speeches they are consistently present among the most frequent lexical bundles, constituting almost half of the sequences in the list.

However, we felt the necessity to make a further sub-division in this category, distinguishing between MWCs *stricto sensu*, and an analogous class with partially different functions, that of Slogans/Soundbites. As regards the former, among the most frequent examples in Bush's addresses there are those which are explicitly connected to the events and the reactions immediately following 9/11: 'the war on terror', 'weapons of mass destruction', 'Department of Homeland Security' and so on. These are characterised as denoting entities and processes specific to topics and events dealt with in this particular domain and time period.

The Slogan/Soundbite sub-class is actually that with the highest proportions in Fig. 4. These denote goals which are evaluated along the parameter of desirability, i.e. actions and their results which are to be reached or avoided. Slogans also tend to emerge from our data as a result of their propensity to form chains: the concatenation of 4-grams may form strings of up to 12 elements, as shown in Table 2.

Table 2. An example of how a slogan emerges from the concatenation of a series of 4-grams. The rightmost column indicates the occurrences of each 4-gram in the Bush A data.

every	child	can	learn									357
				(to)								
					read	and	write	and				268
						and	write	and	add			256
							write	and	add	and		273
								and	add	and	subtract	275

Slogans/Soundbites are clearly deeply related to the genre under investigation, and have two intertwined fundamental functions: the first is clearly the persuasive function, the second is that of their contribution to "Political Branding" as defined by Fairclough (2006: 101-2).

4 CONCLUSIONS

We may conclude by making the following observations. First of all, the results for the Presidential data are clearly distinct from those for both Conversation and Academic Prose. Furthermore, the proportions of different patterns of variability are remarkably consistent from President to President. This suggests that the methodology may be of use

in differentiating between different genres, as not only has a distinction between groups been recorded, but the groups themselves are seen to be relatively compact.

Another point worthy of mention is the marked tendency towards fixedness in the Presidential data. One factor which contributes notably to this is the high proportion of MWCs, a category we deemed in need of further sub-division into MWCs *stricto sensu* and Slogans, the latter in particular being inseparably linked with the domain under investigation.

One point in the methodology which clearly requires further attention is the threshold at which a given slot in a given MWU switches from being variable to fixed. When setting this, even Biber (2009: 282) appears to acknowledge how arbitrary it is: "a simple cut-off of greater or lesser than 50% was used for each slot in a 4-word sequence". While easy to grasp and pleasingly simple, we must be wary of how much variation may be glossed over by the binary variable/fixed. For instance, concerning the Obama section, amongst those lexical bundles which turned out to have a 1234 pattern, the average values across the four slots ranged from 100% (e.g. *a second great depression*) to 60% (*to see if we*): most would feel that the former constitutes a far "tighter" bundle than the second. Therefore, further research may address this issue, applying different thresholds to data of various types and then qualitatively and quantitatively evaluating the results in order to identify whether an optimum can be reached.

References

Biber, D. "A corpus-driven approach to formulaic language in English". *International Journal of Corpus Linguistics* 14:3 (2009), 275–311.

Biber, D., Johansson, S., Leech, G., Conrad, S., and E. Finegan (1999) *Longman Grammar of Spoken and Written English*. Harlow, Longman.

Biber, D., Conrad, S. and V. Cortes (2004) "*If you look at …*: Lexical bundles in university teaching and textbooks". *Applied Linguistics*, 25:3, 371-405.

Biber, D. and F. Barbieri (2007) "Lexical bundles in university spoken and written registers". *English for Specific Purposes*, 26, 263-286.

Carter, R. and M. McCarthy, (2006) *Cambridge Grammar of English*. Cambridge, Cambridge University Press.

Fairclough, N. (2006) *Language and Globalization.* London and New York, Routledge.

Hyland, K. (2008) "As can be seen: lexical bundles and disciplinary variation". *English for Specific Purposes*, 27(1), 4-21.

O'Keeffe, A., McCarthy, M. and R. Carter (2007) *From Corpus to Classroom*. Cambridge, Cambridge University Press.

Stubbs, M. (2007) "An example of frequent English phraseology: distributions, structures and functions". In Facchinetti, R. (ed.) *Corpus Linguistics 25 Years on*. Amsterdam: Rodopi, 89-106.

STUBBS, M. AND I. BARTH (2003) "Using recurrent phrases as text-type discriminators. A quantitative method and some findings". *Functions of Language*, 10:1, 61-104.

The use of corpus analysis to manage foreign language errors in a bilingual community

Maria Luisa Carrió Pastor

Eva M. Mestre Mestre

Universitat Politècnica de València

Abstract: Worldwide communication is possible nowadays using English as an international language or lingua franca. English is used in countries with different cultural backgrounds, a fact which affects in the use of pragmatic strategies. On occasions, authors who communicate in a foreign language cannot avoid the use of structures that are more common in their mother tongue (L1). In a monolingual community, language errors could be caused by L1 interference; nevertheless the methodology applied in error analysis and in corpus compilation could vary in a bilingual community. The linguistic status of three languages in contact may not be equal; consequently ideological, linguistic and social factors could influence language acquisition. The main objective of this paper was to find out if the methodology used for corpora analysis is adequate for a corpus of learners with different linguistic background. In this article, we used corpus analysis methodology to determine if learners whose mother tongues were Spanish o Catalan varied their errors when learning English. Foreign language acquisition is a universal concept although we consider that the proficiency of some skills could depend on the mother tongue of the learner. In order to analyse the corpora, which included the errors of English texts written by students whose mother tongue was Catalan or Spanish, we conducted an experimental research that included the categories of communicative, grammatical and lexical errors. The results showed that students with different cultural backgrounds produced a dissimilar amount of communicative and lexical errors while both groups produced a similar amount of grammatical errors. As a consequence of this research, we concluded that the methodology used to detect errors should vary depending on the linguistic background of learners.

Keywords: bilingual community; second language; corpus analysis; specific setting.

Resumen: La publicación internacional es posible gracias al uso de la lengua inglesa como internacional o lengua franca. Se utiliza en varios países con antecedentes culturales diversos, un hecho que afecta al uso de las estrategias pragmáticas. En ciertas ocasiones, los autores que se comunican en una lengua extranjera no pueden evitar el uso de estructuras que son más comunes en su lengua materna (L1). En una comunidad monolingüe, los errores pueden estar causados por la lengua materna, pero tenemos que ser conscientes de que este hecho puede variar en una comunidad bilingüe. El estatus lingüístico de de lenguas en contacto puede no ser igual, en consecuencia, los factores sociales, ideológicos y lingüísticos pueden influir en la adquisición de una lengua. El objetivo principal de este estudio es saber si la metodología usada para el análisis de corpus es adecuada para un corpus de aprendices con distintos antecedentes lingüísticos. En este artículo, utilizamos la metodología del análisis de corpus para determinar si los aprendices cuya lengua materna era el español o el catalán variaban sus errores cuando aprendían inglés. La adquisición de una lengua extranjera es un concepto universal aunque consideramos que los conocimientos de ciertas habilidades pueden depender de la lengua materna del aprendiz. Para analizar el corpus, que incluye los errores en inglés de textos escritos por alumnos cuya lengua materna es catalán o español, realizamos un experimento de las categorías de errores comunicativos, gramaticales y léxicos. Los resultados nos demostraron que los estudiantes con antecedentes culturales distintos producen una cantidad desigual de errores léxicos y comunicativos y una cantidad similar de errores gramaticales. Como consecuencia de esta investigación, concluimos que la metodología utilizada para detectar los errores debería de variar según los antecedentes lingüísticos de los estudiantes.

Palabras clave: comunidad bilingüe; segunda lengua; análisis de corpus; campo específico.

1. Introduction

It is a well-known fact that international communication is possible nowadays using English as a lingua franca. English is used to communicate professionally and socially by speakers with different cultural backgrounds, a fact that affects in the use of teaching and learning strategies. On occasions, learners who communicate in a foreign language cannot avoid the use of structures that are common in their mother tongue (L1). In a monolingual community, language errors could be caused by L1 interference. However, when dealing with a corpus compilation and a subsequent error analysis, a different methodology could be necessary in a bilingual community, since different linguistic backgrounds can be expected. In addition, the linguistic status of several languages in contact might not be equal; consequently ideological, linguistic and social factors could influence language acquisition in a specific setting.

1.1. Corpus analysis and methodology

The analysis of corpora can be carried out with the use of different methodologies. Traditionally, corpus linguistics has been linked to quantitative analysis as a support of research studies and hypothesis, aimed at obtaining numerical results. Indeed, different types of corpora are used to validate theories and ideas, and provide examples that support knowledge (Hornero, Luzón and Murillo, 2006). The foundation of language studies in frequencies of use in order to obtain reliable data is widely accepted in the literature. A corpus enables us to investigate different aspects of language use based on hard data. Deeper analysis of linguistic behaviour; for example, the detection of language trends, is also feasible when based, as it is, on real use rather than on what an idealised native speaker might say.

All the same, a corpus can also be analysed using a qualitative research method, which does not rely on statistics analysis to obtain results. However, Richards (cited in Dörnyei 2007: 25) wisely emphasizes: "[…] qualitative and quantitative data do not inhabit different worlds. They are different ways of recording observations of the same world". Indeed, Sandelowski (2003) broadens this idea explaining that qualitative and quantitative research are not always clearly distinguishable, arguing that comparison between these two cannot be made. A mixed method was considered appropriate for the purposes of the present work, coinciding with the recommendations expressed by Miles and Huberman in Dörnyei (2007: 42):

Entertain mixed models. […] Quantitative and qualitative inquiry can support and inform each other. Narratives and variable-driven analyses need to interpenetrate and inform each other. […] think of it as hybrid vigour.

Insisting on the advantages offered by mixed methods of analysis to offer the appropriate support framework to theoretical principles, the Grounded Theory by Strauss and Corbin (1998:34) can be pointed out: "Qualitative and quantitative forms of research both have roles to play in theorising. The issue is not whether to use one form or another but rather how these might work together to foster the development of theory".

In sum, a carefully planned corpus and well-designed analysis are not enough. One must also exercise due care and attention in the interpretation of the results. It is crucial that we plan the way in which we must obtain our corpus, as well as the way in which results derived from it can be interpreted.

With regards the specific setting of Second Language Learning, an area in which corpus linguistics has been widely used is the compilation of grammar rules and sentence structures which help identifying new frames and adapting them to their actual use. Frequency analysis allows us to verify the use of certain structures and helps us to determine the way the elements of a language are sequenced. Sinclair (1991), who was aware of the importance of establishing a bond between sense and structure through lexical analysis, was perhaps one of the first to apply this at a massive scale. As McCarthy (2001: 127) notes:

[Sinclair's proposal] stands as a good example of how a 'neutral' technology can throw up fundamental questions for theory, and how a practical, 'applied' problem, in this case writing a dictionary using computer evidence, can bounce back and challenge theory. We should not doubt that galloping technological change will bring many more such upheavals over the coming decades.

Other specific applications of corpora in language learning are the evaluation of language proficiency, the improvement of teaching methodologies, the design of specific material, the compilation of learner corpora following the different levels of the Common European Framework of Reference for Languages (CEFR), the detection of the interlanguage or variation of different levels in language learning and last but not least, error corpora used for instance to design specific material and approaches adapted to the needs of learners.

The importance of corpora analysis and its application to applied linguistics is beyond doubt, as recent studies can confirm (Biber, Conrad and Reppen, 1998; Cortese, 2002; de Monnink, 1998; Dörnyei, 2007; Hornero, Luzón and Murillo, 2006; Hunston, 2002; Martí Guinovart, 1999; Oostdijk, 2000).

1.2. Error analysis and learner corpora

A concept that should be emphasized by language teachers is the difference between correct and incorrect text production in a second language. An error in a sentence or text in a language implies that there is a more adequate word or expression than the one used by a user. As a consequence, an incorrect production should be replaced by a more suitable one accepted by the majority of speakers. Although error conception is clear, what are the implications of language errors? Errors can be produced by inadequate learning, errors can be caused by mother tongue influence, misunderstanding can cause errors, term confusions can result in errors, etc. The causes can be diverse, but what are the solutions? Can we use error analysis or error-based corpora to identify language proficiency levels and to provide learning strategies? These are the questions we would like to answer in this paper, but first we will review other concepts associated with error analysis, corpora and language learning.

The process of second language learning entails mother tongue influence, at least in the first stages, as linguistic models tend to be copied when integrating new concepts. This creates different ways to express the same discourse or more specifically, variations within the same genre, causing errors or inaccurate language production. As an example, specific texts should reflect the knowledge of users and their own research. Nevertheless, occasionally, learners using ESP do not choose adequate terms and expressions that are crucial to communicate their ideas. In sum, different cultural backgrounds or inadequate language acquisition can produce misunderstandings.

In this paper, we are going to focus on error classification in order to demonstrate that a customised error corpus could be used to improve learners' command of English. Error analysis provides students with learning strategies that can be used to become skilled at a second language in a faster way, although learners' mother tongue should be considered. Different linguistic backgrounds, with the use of the same learning methods could generate a wide range of errors. In this way, an appropriate error classification could help teachers produce specific exercises, which would improve students' learning processes. For this reason, error analysis helps the understanding of second language production, determining the role of the mother tongue and of the learning deficiencies (Carrió Pastor, 2004; Mestre and Carrió, 2010).

In order to obtain more usable results, errors should be classified taking into account different interlanguage levels. Learners acquire a second language in different stages that Selinker (1965) named interlanguage, which should be considered to identify and differentiate errors in order to mark the general levels of language learning. This concept has been considered too vague and difficult to differentiate till now. Recently, the CEFR[203] proposes a series of levels based on language proficiency that could be compared to Selinker's proposal. The levels are defined as "a series of descriptions of abilities which can be applied to any language and can be used to set clear targets for achievements within language learning. It has now become accepted as a way of benchmarking language ability all over the world." It established different levels representing different interlanguages: A1; A2; B1; B2; C1 and C2. Each level is a step in language progress and implies certain skills that students should acquire. An error corpus associated to CEFR levels could be a very interesting tool that could help students to improve their language and provide teachers with the possibility of creating materials adequate to each level.

Error classification has long been considered from the point of view of structuralism; the most important aspect then considered was the error in sentence structure. Some researchers (Al-Jarf 2000; Carrió & Seiz 2000; Levinson, Lessard & Walter 2000; Lee 2004; Carrió Pastor 2005) have established the difficulty in producing an adequate language structure in second language learners of English. James (1998:129) proposes specific criteria for classification: modality, medium and level. Modality refers to the production or reception of the message, medium indicates the kind of production under analysis, spoken or written, and level specifies the error class (substance or medium, text or usage and discourse or use). This proposed classification does not establish the specific causes of the error, so a new classification, linked to the different classes found in a corpus is needed. Error interpretation is essential to know error causes. A thorough

203 http://www.cambridgeesol.org/what-we-do/europe/cefr.html

classification of an error corpus could determine the different causes that interfere in second language learning and allow the production of specific material to avoid incorrect language production in the different CEFR levels.

Learner corpora have been considered crucial to help people who are learning a language. As Hunston (2002: 206) explains: "These corpora can give information about the difference between learners and between learners and native speakers". The most important work in this area was done by Granger whose web site (http://www.uclouvain.be/en-cecl.html) can be visited to observe all the projects that can be developed from learner corpora. Granger (1998) and her team developed several ideas related to general and specific learner corpora, which were used to level and help students. In her studies, the most important aspect of learner corpora is comparison; Standard English is compared with English produced by language learners. Their studies consider the linguistic background of second language learners, although they do not contrast the errors produced by learners with different linguistic backgrounds. Granger's studies on corpora are quantitative rather than qualitative although she considers some quantitative aspects of certain lexical items. Second language corpora are gathered although third language corpora are not considered. Although she collects corpora from bilingual areas, she does not consider if the learner corpora are produced by second or third language learners. In the case of bilingual communities, we consider that qualitative analysis should be more relevant than quantitative analysis. Indeed, they insist that "one important finding emerging from learner-corpus-based studies in general and EAP in particular is that some of the linguistic features that characterize learner language are shared by learners from a wide range of mother tongue backgrounds whilst others are exclusive to one particular learner population" (Gilquin, Granger, & Paquot 2007: 7). The singularities of third language acquisition should also be considered in a learner corpus in order to reflect language learning. Some researchers, as Flowerdew (2000), consider that special attention should be paid to language varieties and bilingual communities, when she studies a corpus of Hong Kong learners' writings.

Although learner corpora could be useful, it is not always easy to see how they could be transferred to help pedagogic issues. More investigation is needed to advise learners. For instance, teachers should be conscious of the precise circumstance of the overuse or underuse of a word or a structure before providing generalizations.

The main objective of this paper was to find out if the general methodology used for corpora analysis is adequate for a corpus of learners with different linguistic background. In this article, we used corpus analysis methodology to determine if learners whose mother tongues were Spanish and Catalan varied their errors when learning English.

2. Methodology

The corpus used in this research included the errors produced in specific English texts written by students whose mother tongue was either Catalan or Spanish. The Valencian Community is a bilingual area. Due to this fact, at Universitat Politècnica de València some courses are taught both in Spanish and in Catalan; that is, the same subjects are

taught in both languages and the students in the Catalan groups are bilingual. Students have been taught Spanish and Catalan at school and they can communicate fluently in both languages, although their mother tongue is Catalan. The students we analysed were enrolled in the electronic Engineering degree at Universitat Politècnica de València. They were tested to check they had a B1 level of English proficiency (following the CEFR).

Once the groups were selected, activities that helped spotting the errors of students were planned. We prepared a specific task for students in both groups in order to analyse language learning and error production. We gathered fifty specific writings produced by the Catalan speakers in English and fifty specific writings produced by the Spanish speakers in English. The activity proposed was the reading of a description of electromagnetic fields and the subsequent proposal of an application. We conducted an experimental research that included the categories of communicative, grammatical and lexical errors. With regards the methodology used, we considered both kinds of methodological research to analyse the results: the quantitative methodology research and the qualitative methodology research. By comparing the results obtained, we could identify the best methodology to apply in the case of texts produced by bilingual learners. The writings were corrected manually by three of the teachers involved in the teaching of the subject. We collected the writings and classified them in the categories and sub-categories established beforehand. Our conclusions provided further evidence regarding the importance of corpora for language acquisition and regarding the fact that relatively small but highly specialised corpora can be used to describe the language of specific communication and to contrast the language production of speakers with different linguistic backgrounds.

3. Results

We analysed the texts and classified the errors in communicative, grammatical and lexical errors. We chose these classifications as they reflected the most appropriate competences for their language proficiency. The communicative classification included errors related to the sequencing of the parts of the texts produced, the beginning of the paragraphs and the focus of the writing. The grammatical classification included the errors associated to the formation of complex noun phrases, the use of verb tenses and the formation and placement of adverbials. The lexical classification included errors produced due to the influence of the mother tongue, the use of false friends and the writing of words in an incorrect way.

The results obtained in the category of communicative errors are explained below, grouped according with the type of error incurred:

Table 1. Communicative errors.

Communicative errors	Spanish speaker	Catalan speaker
Sequencing	15 (57.7%)	21 (41.2%)
Beginning of paragraphs	7 (27.9%)	24 (47.1%)
Focus	4 (14.4%)	6 (11.7%)
Total	26 (100%)	51 (100%)

As it can be observed, more errors were in the category of Catalan speakers and the most frequent ones were incorrect word choice or incorrect sentence part to begin a paragraph. In the group of Spanish speakers we found more occurrences in the sequencing of the text, i.e. they could not follow a correct order in the text. There are slight differences regarding the subject of the text. Most students kept their texts focused on the specific matter proposed.

The results obtained in the category of grammatical errors can be observed in Table 2:

Table 2. Grammatical errors.

Grammatical errors	Spanish speakers	Catalan speakers
Complex noun phrases	35 (35.4%)	37 (38.9%)
Verb tenses	47 (47.5%)	45 (47.4%)
Adverbials	17 (17.1%)	13 (13.7%)
Total	99 (100%)	95 (100%)

The items chosen for the analysis are common characteristics of specific English language. The occurrences obtained showed that both Spanish and Catalan speakers failed in the same grammatical aspects analysed. In both groups we obtained a similar number of occurrences within the same error categories, therefore showing little differences in the grammatical competence of the students.

The last category analysed was lexical errors, in Table 3 we can observe the occurrences found after the analysis:

Table 3. Lexical errors.

Lexical errors	Spanish speakers	Catalan speakers
Mother tongue	7 (38.9%)	16 (41.0%)
False friends	6 (33.4%)	15 (38.4%)
Incorrect words	5 (27.7%)	8 (20.6%)
Total	18 (100%)	39 (100%)

The occurrences of errors obtained in this last category were less frequent than in the two previous categories. It can be seen that very few errors belong to this category. However, we should highlight that Catalan speakers committed more errors than Spanish speakers in this category, in particular errors related to Mother tongue influence and false friends.

We analysed in more detail the subcategories of the Spanish speakers in the corpus that included few occurrences as we considered that a qualitative analysis could be useful for the research. Our research was based on qualitative analysis of the occurrences found as the social and the linguistic facts were relevant in this study. We wanted to analyse if a qualitative analysis was adequate in order to show the different error production of second language learners. We chose a specific corpus as the characteristics to be analysed could be easily differentiated form the global aspects of language production.

As can be seen in the results tables, different cultural backgrounds can derive in different types of problems when facing the learning of a foreign language, resulting in errors of different sort.

4. Conclusion

As a consequence of this research, we concluded that the methodology used to detect errors should vary depending on the linguistic background of learners. Therefore, a previous analysis of the sources used for the corpus could be helpful to identify the extra-linguistic specificities which should be taken into account when designing the Corpus. The need to analyse corpora using qualitative research was highlighted, as it has got some characteristics that are more applicable to learner corpora. First, the design of the research is more flexible and kept open so it may adapt to some errors that are difficult to categorize. Second, cultural meaning or interpretations can be observed as data can be captured in context, i.e. any relevant information can be included. Third, this research is concerned with subjective opinions and experiences, i.e. it is more appropriate to small groups of participants or a reduced area of study.

The results showed that students with different cultural backgrounds produced a dissimilar amount of communicative and lexical errors while both groups produced a similar amount of grammatical errors. A possible interpretation of this is that for Catalan students English is an L3, and they transfer errors from L1 to L2 and from L2 to L3, that is, students translate into Spanish before they produce in English. This could be an interesting point to explore in order to find more appropriate ways for bilingual students to face the learning of an L3, ensuring that there is no mediation of an L2. This could also imply that L1 and L2 do not have the same sociolinguistic status by Catalan speaking students, since they feel the need of translating into Spanish and only then communicate in English.

The exploratory nature of qualitative research has been one of the aspects we have considered more relevant as we have focused on the way language is acquired depending on the mother tongue of the speakers. Our aim was to broaden our understanding of the cognitive mechanisms implied in language acquisition in a specific setting. We considered relevant to detect the differences between speakers with different mother tongues. We are conscious that we cannot generalise provided the small corpus we gathered, but we would like to emphasize that the exploration of specific meaning does not require large samples.

5. References

Al- Jarf, R. (2000) Grammatical agreement errors in L1/L2 translations. *IRAL*, 38, 1- 15.

Biber, D., Conrad, S. And Reppen, R. (1998) *Corpus linguistics: investigating language structure and use.* Cambridge: Cambridge University Press.

Carrió, M. L. & Seiz, R. (2000). La expresión escrita en inglés técnico. Sus errores. In *III Congreso Internacional sobre Lenguas para Finalidades Específicas*. Barcelona: Universidad de Barcelona, 69-73.

Carrió Pastor, M. L. (2005) *Contrastive analysis of scientific-technical discourse: Common writing errors and variations in the use of English as a non-native language.* Ann Arbour: UMI.

Dörnyei, Z. (2007) Quantitative, Qualitative and Mixed Methodologies. Oxford: Oxford University Press.

FLOWERDEW, L. (2000) Investigating referential and pragmatic errors in a learner corpus, in L. Burnard and T. MacEnery (eds.) *Rethinking Language Pedagogy from a Corpus Perspective.* Frankfurt: Peter Lang.

GILQUIN, G., GRANGER, S. & PAQUOT M. (2007) Learner corpora: the missing link in EAP pedagogy, available at http://sites-test.uclouvain.be/cecl/archives/GILQUIN_GRANGER_PAQUOT_2007_Learner_corpora_missing_link.pdf [01/03/2011]

GRANGER, S. (1998) *Learner English on Computer*. London: Longman.

HORNERO, A., LUZÓN, M.J. AND MURILLO, S. (eds.) (2006). *Corpus Linguistics: Applications for the study of English*. Peter Lang: Frankfurt, New York.

HUNSTON, S. (2002) *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.

JAMES, C. (1998) *Errors in Language Learning and Use*. London: Longman.

LEE, I. (2004) Error correction in L2 secondary writing classrooms: the case of Hong Kong. *Journal of Second Language Writing*, 13, 285-312.

LEVISON, M.; LESSARD, G. AND WALKER, D. (2000) A multi-level approach to the detection of second language learner errors. *Literary and Linguistic Computing*, 15-3, 313- 322.

MARTÍ GUINOVART, M. A. (1999). "Panorama de la lingüística computacional en Europa". *Revista Española de Lingüística Aplicada. Volumen Monográfico: Panorama de la Investigación en Lingüística Aplicada*, 11- 24.

MCCARTHY, M. (2001). *Issues in Applied Linguistics.* Cambridge: Cambridge University Press

MESTRE MESTRE, E. M. AND CARRIÓ PASTOR, M. L. (2010) The use of a corpus of lexical errors in second language learning. *Proceedings of the 2nd International Congress on Corpus Linguistics*. La Coruña: Publicaciones de la Universidad de La Coruña.

MÖNNINK, I. DE (1997). "Using corpus and experimental data: a multimethod approach". Http://iris1.let.kun.nl/literature/demonnink.1997.2.html (13-03-98, 12:26)

OOSTDIJK, N. (2000). "Corpus-based English linguistics at a cross-roads". English Studies, 81- 2: 127- 141.

SANDELOWSKI M (2003) Tables or tableaux? The challenges of writing and reading mixed methods studies, in Tashakkori A and Teddlie C (Eds) *Handbook of mixed methods in social & behavioral research*, Sage: Thousand Oaks.

SELINKER (1965) Selinker, L. (1972), Interlanguage. International Review of Applied Linguistics, 10, 209-241.

SELINKER, L. (1992) *Rediscovering Interlanguage.* London: Longman.

SINCLAIR, J. MCH. 1991. *Corpus, concordance, collocation.* Oxford: Oxford University Press.

STRAUSS, A.L. & J. CORBIN (1998) Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory, 2nd Ed. Sage: Thousand Oaks, California.

Utilidad/uso especifica/o de un corpus de definiciones de categorías semánticas.

José María Guerrero Triviño
Rafael Martínez Tomás.
UNED. Departamento de Inteligencia Artificial.

Mª Carmen Díaz Mardomingo
Herminia Peraita Adrados
UNED. Facultad de Psicología. Departamento de Psicología Básica I.

Resumen

El diagnóstico de las enfermedades neurodegenerativas, y en especial la Enfermedad de Alzheimer, es una tarea extremadamente dificultosa, sobre todo cuando la enfermedad es incipiente y de intensidad leve. Este trabajo de investigación, tiene por objeto la caracterización de las enfermedades neurodegenerativas en sus primeros estadios, utilizando como instrumento metodológico de primer orden, un corpus de definiciones orales para el estudio de patologías en relación con el deterioro semántico que producen estas enfermedades. Este corpus de definiciones orales, es la fuente de información de un Sistema Informático que usa técnicas de Inteligencia Artificial, en concreto Redes Bayesianas, para el diagnóstico de las enfermedades neurodegenerativas, y en especial el diagnóstico de la Enfermedad de Alzheimer.

La Red Bayesiana presenta un modelo causal basado en el corpus de definiciones literales de determinadas categorías semánticas –del nivel básico de categorización—tanto de seres vivos, como de seres no vivos.

Palabras clave: corpus de definiciones orales, red bayesiana, enfermedad de Alzheimer.

Abstract

The diagnosis of neurodegenerative diseases, especially Alzheimer's disease, is sometimes an extremely difficult task, especially when incipient and of slight intensity. This research aims to characterize neurodegenerative diseases in their early stages. It uses as first order methodological tools, a corpus of oral definitions for the study of diseases in relation to the semantic deterioration of these diseases. This corpus of oral definitions is the information source of a computer system that uses intelligent artificial techniques, in particular Bayesian networks, for the diagnosis of neurodegenerative diseases, and specially Alzheimer's disease.

The Bayesian networks presents a causal model based on corpus of literal definitions of concrete semantic category –both living beings and artifacts.

Keywords: *corpus of oral definitions*, Bayesian Networks, Alzheimer disease.

1. Metodología.

La metodología empleada es mediante test oral con restricción temporal, aplicado a los pacientes a los que se les solicita la definición de determinadas categorías semánticas. A partir de la producción verbal libre con restricción temporal, se realiza la grabación y posterior transcripción, así como el posterior análisis de la producción en un determinado marco teórico. En este análisis se distinguen dos tipos de variables: unas cuantitativas (las frecuencias de producción de rasgos, para cada categoría, cada dominio, etc.) y otras cualitativas (los diferentes tipos de rasgos según el modelo). La investigación sobre la memoria semántica y la representación del conocimiento ha constatado diferencias cuantitativas en la producción de rasgos entre personas sanas y enfermos con patologías neurodegenerativas. Este análisis y posterior discretización, proporciona las evidencias para la Red Bayesiana.

Del mismo modo, se ha podido evidenciar que existe un deterioro semántico diferencial entre las categorías de seres vivos (entidades biológicas) y seres no vivos (entidades no biológicas), en personas afectadas con determinadas patologías neurodegenerativas (Alzheimer, Demencia semántica, Demencia por cuerpos de Lewy, etc.), traumáticas (traumatismo craneal) e infecciosas (herpes por encefalitis). Las categorías semánticas se derivan de clasificaciones que se llevan a cabo en el mundo que nos rodea y que permiten tratar como equivalentes objetos que en sí son diferentes. Gracias a que nuestra memoria semántica se encuentra organizada en función de dichas categorías, podemos realizar una serie de funciones cognitivas importantes, tales como hacer inferencias, establecer relaciones entre ejemplares, atribuir propiedades a objetos que no conocemos, razonar; todo lo cual se basa en un principio de economía cognitiva. Las personas que sufren los déficit específicos de categoría, muestran una peor ejecución en tareas que afectan, total o parcialmente, el conocimiento del dominio categorial de los seres vivos mientras que el dominio de los objetos o artefactos –seres no vivos— está total o casi totalmente conservados. También existe un pequeño número de casos en el que se da el patrón contrario, hay un mayor deterioro del dominio de los objetos o artefactos, mientras que el dominio de los seres vivos está, en su mayor parte, preservado (Grasso, Díaz, & Peraita, 2009).

Por otro lado, las Redes Bayesianas constan de dos componentes fundamentales: la estructura y los parámetros. La estructura de la red probabilista, es la parte cualitativa de la red, es decir, define las relaciones causales, funcionales e informativas; identificadas en el dominio. Los parámetros son las probabilidades condicionales y utilidades, y constituyen la parte cuantitativa de la red. La parte cuantitativa de la red, expresa la fuerza de las relaciones probabilistas y son representadas por probabilidades condicionales. Normalmente las relaciones causales entre variables suelen ir acompañadas de un factor de incertidumbre o un grado de correlación, que se pueden expresar fácilmente en las redes Bayesianas a través la fuerza de la relación (Kjaerulff & Madsen, 2007). Nuestro modelo cualitativo se ha construido de forma manual y el modelo cuantitativo se ha construido usando técnicas de aprendizaje automático basandonos para ello, en estudios epidemiológicos (Fernández Martínez, Castro-Flores, Pérez-de las Heras, Mandaluniz-Lekumberri, Gordejuela, & Zarranz, 2008) y en una base de casos con personas

diagnosticadas, como pacientes de Alzheimer y personas sanas, tal y como se explica en los experimentos. (Peraita & Grasso, 2010).

En el modelo causal, expresamos que las enfermedades neurodegenerativas son las causas más probables de producir un déficit léxico-semántico-conceptual. Es decir, esta investigación se basa en el uso de técnicas de razonamiento abductivo, donde partimos de unos síntomas, que en este caso es sólo el deterioro semántico, y se busca la causa más probable que explica esos síntomas. En este caso los síntomas son las alteraciones a nivel cognitivo y la posible causa son las enfermedades neurodegenerativas. Además en este modelo causal, se tienen en cuenta factores de riesgo y de protección, como la edad, el sexo o el nivel cultural (Fernández Martínez, Castro-Flores, Pérez-de las Heras, Mandaluniz-Lekumberri, Gordejuela, & Zarranz, 2008).

2. Modelo de Red Bayesiana.

En la siguiente figura podemos observar uno de los modelos causales usados en este trabajo de investigación. Inicialmente hemos empleado un número reducido de factores de riesgo y protección. Tal y como se ha indicado anteriormente, hemos empleado técnicas de aprendizaje automático para el modelo cuantitativo, ya que el número de parámetros necesarios para construir este modelo de Red Bayesiana supera los 1500 y es inabordable realizarlo con valoraciones subjetivas y/o estudios epidemiológicos.

Se puede observar en el modelo de Red Bayesiana, como las variables de información de contexto o factores de riesgo tienen una relación causal con las enfermedades neurodegenerativas. Se ha podido evidenciar, y así lo demuestran los estudios epidemiológicos, que la edad, el sexo y el nivel educativo tienen algún tipo de relación con las enfermedades neurodegenerativas. Además, tal y como hemos indicado anteriormente, la edad y el nivel educativo tienen influencia en las definiciones orales que realizan los pacientes; por ejemplo, las personas más mayores suelen producir menos atributos en las definiciones orales de objetos básicos que las personas más jóvenes. La misma situación se puede producir en personas con un bajo nivel cultural, estas personas suelen producir menos atributos que las personas con un mayor nivel educativo. Estas situaciones las hemos tenido en cuenta a la hora de propagar evidencias por la red Bayesiana e inferir un diagnóstico.

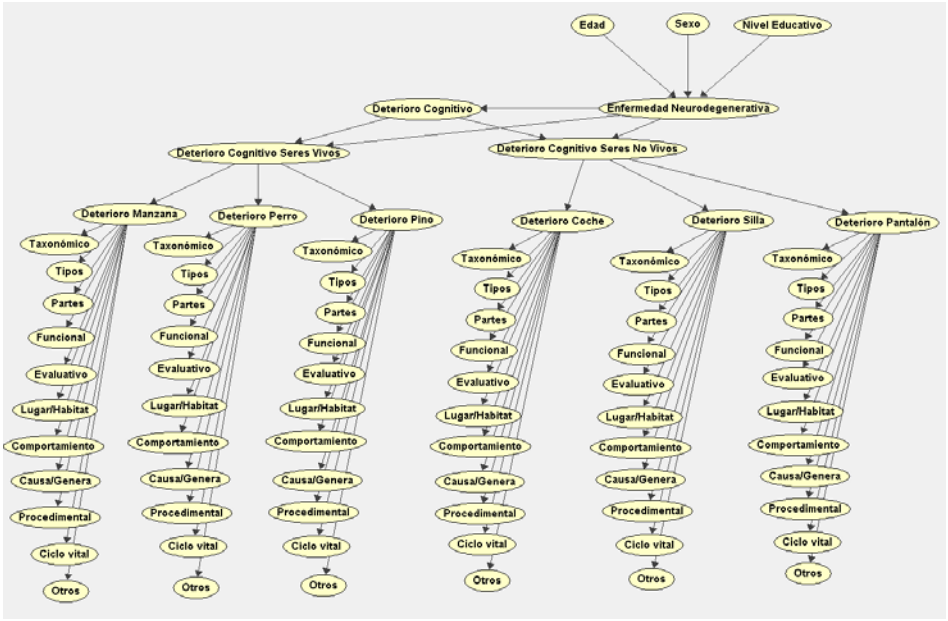


Figura 1.- Red Bayesiana para el diagnóstico de Enfermedades Neurodegenerativas basada en un corpus de definiciones orales.

En esta Red Bayesiana se pueden distinguir cuatro tipos de variables (Kjaerulff & Madsen, 2007):

- Variables de información de contexto o factores de riesgo: Es la información que está presente antes de que ocurra el problema y que tiene una influencia causal sobre la enfermedad. En este grupo de variables tenemos: edad, sexo y nivel educativo.
- Variables de información que representan los síntomas. Variables que representan si el paciente cursa un déficit léxico-semántico-conceptual. Estas variables analizan los rasgos o atributos contenidos en una serie de definiciones orales de categorías semánticas de seres vivos y seres no vivos, y otros específicos para cada una de estas categorías. Los comunes son: taxonómico, funcional, parte-todo, evaluativo, lugar/hábitat, tipos, y los no comunes: procedimiento, actividad comportamental, causa/generación y procedencia. La taxonomía de atributos que sirve de esquema o marco teórico y metodológico de evaluación para esta prueba, al igual que en la segunda prueba, puede verse detalladamente en (Peraita, González-Labra, Sánchez, & Galeote, 2000).
- Variables intermedias. Son variables no observables directamente cuyas probabilidades a posteriori no son de interés inmediato, pero que juegan un papel importante para lograr una correcta dependencia e independencia condicional de las

propiedades y por tanto una inferencia eficiente. En este tipo de variables tenemos por un lado, deterioro cognitivo de seres vivos y seres no vivos, y por otro lado, deterioro cognitivo de manzana, perro, pino, coche, silla y pantalón.

- Variables de interés o hipótesis. Son aquellas sobre las que queremos calcular su probabilidad a posteriori a partir de los hallazgos. Estas variables son: padece deterioro cognitivo (demencia) y padece enfermedad neurodegenerativa.

Merece la pena destacar en este modelo de red Bayesiana, el enlace causal entre el nodo enfermedad neurodegenerativa y, los nodos deterioro cognitivo seres vivos y deterioro cognitivo seres no vivos. Con esta relación causal, se expresa el deterioro semántico diferencial entre las categorías de seres vivos y seres no vivos. En el caso de que el paciente presente este deterioro diferencial, la red Bayesiana, mediante la fuerza de las relaciones, aumentaría la probabilidad de que el paciente esté afectado por determinadas patologías neurodegenerativas, traumáticas o infecciosas.

3. Experimentos.

Actualmente disponemos de una base de casos con 69 personas pertenecientes a una muestra española, y estamos trabajando en ampliar esta base de casos en 142 personas más de una muestra argentina. Al disponer de muestras de dos países, nos permitirá llevar a cabo una investigación transcultural hispano-argentina.

Esta base de casos está clasificada en sanos y enfermos. Los pacientes enfermos han sido diagnosticados como tal, por neurólogos. Posteriormente, a todos los casos, tanto sanos como enfermos, les ha aplicado el test Minimental para detectar y evaluar la progresión del trastorno cognitivo asociado a enfermedades neurodegenerativas como la de tipo Alzheimer. Este test, junto con el diagnóstico de los neurólogos, nos permite clasificar la muestra en sujetos sanos, enfermos leves y enfermos moderados.

Una vez hemos clasificado todos los casos, los pacientes han realizado test orales con restricción temporal, en los que se les ha solicitado la definición de determinadas categorías semánticas. A partir de estos test, se ha segmentado la producción de atributos en once bloques conceptuales básicos para cada una de los objetos de las categorías semánticas seres vivos y seres no vivos, tal y como se muestra en la figura 2. Estos test orales, una vez transcritos y analizados, se incorporan a la base de casos y posteriormente se discretiza con técnicas de agrupamiento. Una vez discretizado la producción de atributos, se incorpora a la Red Bayesiana en forma de evidencias, a través de las variables de información de contexto o síntomas.

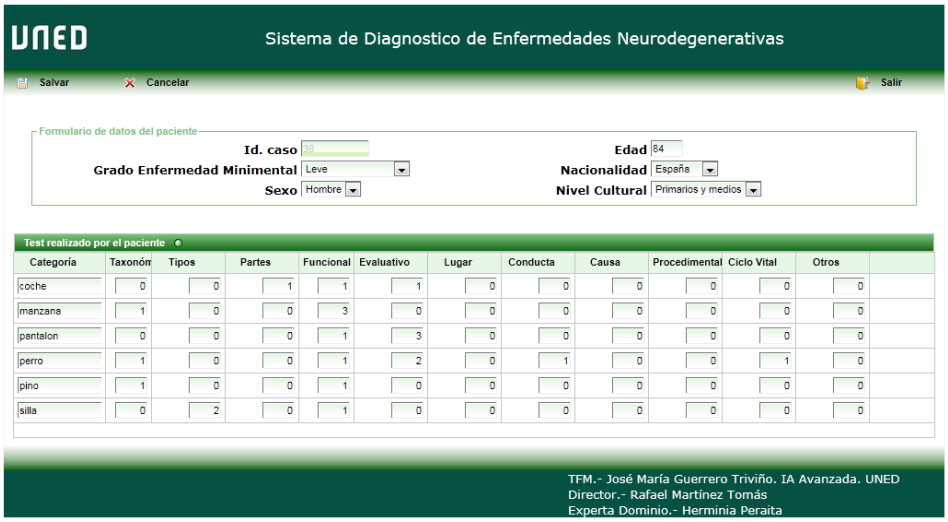


Figura 2. Ejemplo de un caso de la muestra Española.

Este experimento consiste en medir la eficacia del modelo de red Bayesiana que hemos creado para el diagnóstico del deterioro cognitivo compatible con las enfermedades neurodegenerativas, basándonos para ello, en un corpus de definiciones orales. Para este experimento hemos utilizado los mismos casos tanto para la validación, como para el aprendizaje de los parámetros. Para conseguir esto, hemos propagado todas las evidencias de cada uno de los casos, de la base de casos, de forma automática y hemos representado los resultados en tres gráficas, tal y como se puede ver en las figuras 3, 4 y 5. Por cada caso, hemos obtenido como resultado de la inferencia de la Red Bayesiana, la probabilidad de que el paciente esté sano o padezca una enfermedad neurodegenerativa en los niveles leve y moderado. Estamos trabajando en nuevos casos para incorporarlo a la base de datos y poder así, analizar los resultados de esta investigación empleando técnicas cross-validation, es decir, usando un conjunto de casos para el aprendizaje del modelo cuantitativo, y otro conjunto distinto para la validación.

En las figuras 3, 4 y 5, el eje de abscisa representa el número de caso y el eje de ordenada representa la probabilidad de padecer o no padecer alguna enfermedad neurodegenerativa. Cada ilustración dispone de tres curvas que representan la probabilidad de estar sano (color verde), la probabilidad de padecer una enfermedad neurodegenerativa leve (color azul) y por último la probabilidad de padecer una enfermedad neurodegenerativa moderada (color rojo), según los cálculos del proceso de inferencia de la Red Bayesiana. Cabe destacar, que la probabilidad de padecer una enfermedad neurodegenerativa en el proceso de inferencia de la Red Bayesiana, viene determinada por un lado, por los factores de riego y por otro lado, por el corpus de definiciones orales, a partir del que podemos inferir el posible deterioro cognitivo que pudiese presentar el sujeto.

En la figura 3, representamos todos los casos que han sido clasificados como sanos. Las distintas curvas representan la probabilidad generada por la Red Bayesiana de estar sano (curva de color verde), padecer una enfermedad neurodegenerativa leve (curva de color azul) o padecer una enfermedad neurodegenerativa moderada (curva de color rojo).

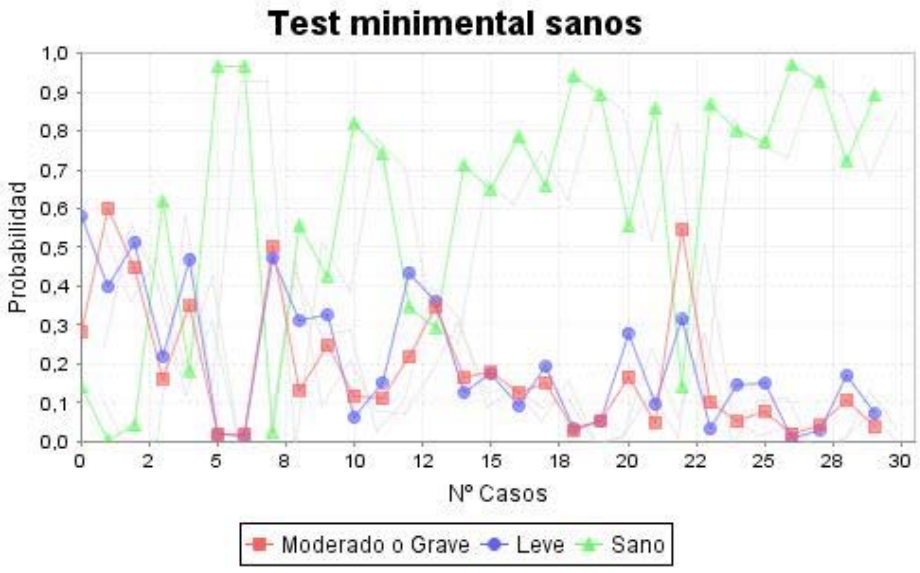


Figura 3. Experimento con toda la base de datos de casos para sujetos que están sanos.

En la figura 4, representamos todos los casos en los que los pacientes han sido clasificados como enfermos que sufren una enfermedad neurodegenerativa leve. Las distintas curvas representan la probabilidad generada por la Red Bayesiana de estar sanos (curva de color verde), padecer una enfermedad neurodegenerativa leve (curva de color azul) o padecer una enfermedad neurodegenerativa moderada (curva de color rojo).

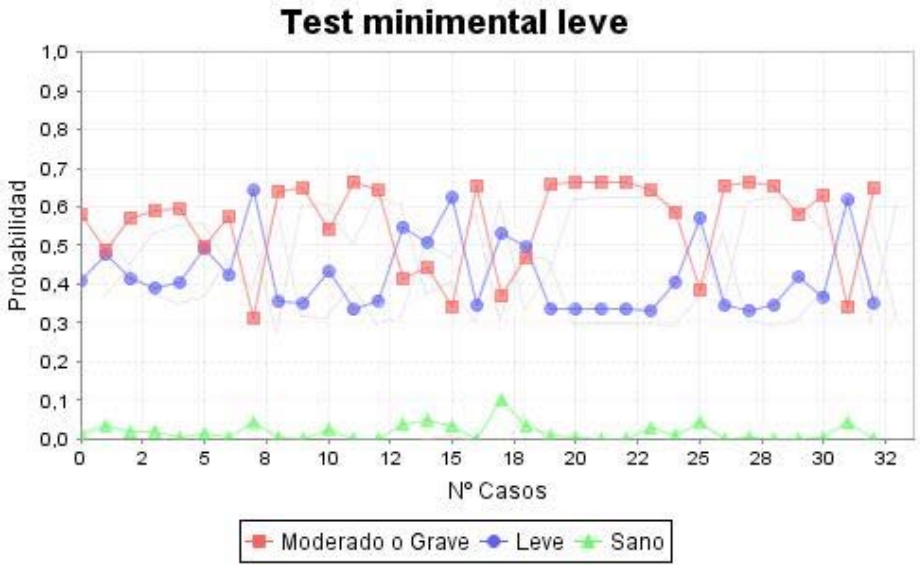


Figura 4. Experimento con toda la base de datos de casos para pacientes que sufren alguna enfermedad neurodegenerativas en nivel leve.

En la figura 5, representamos todos los casos en los que los pacientes han sido clasificados como enfermos que sufren una enfermedad neurodegenerativa moderada. Las distintas curvas representan la probabilidad generada por la Red Bayesiana de estar sanos (curva de color verde), padecer una enfermedad neurodegenerativa leve (curva de color azul) o padecer una enfermedad neurodegenerativa moderada (curva de color rojo).

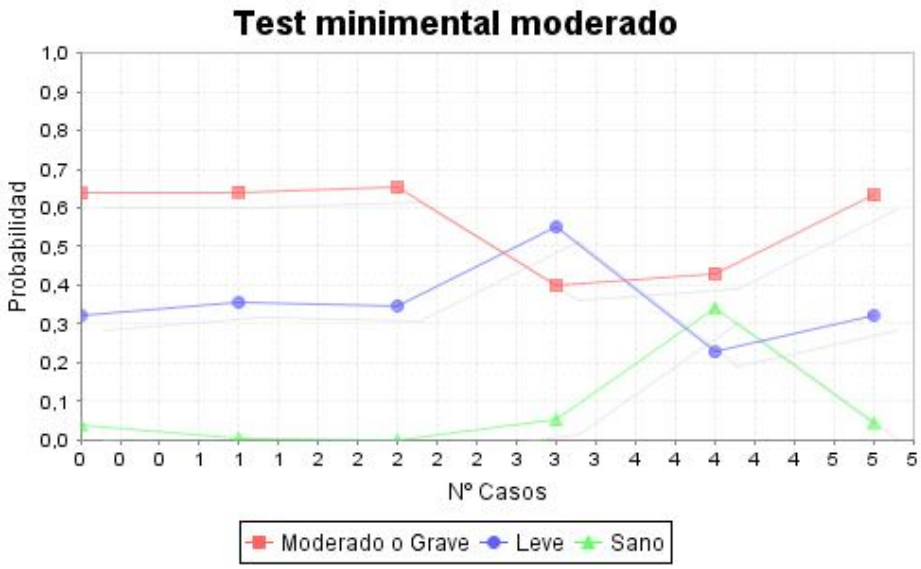


Figura 5. Experimento con toda la base de datos de casos para pacientes que sufren alguna enfermedad neurodegenerativas en nivel moderado.

En este primer modelo de red Bayesiana y experimento, hemos podido comprobar que los resultados son prometedores. Tenemos algún falso positivo, como se puede ver en la figura 3, pero ningún falso negativo, como se puede comprobar en las figuras 4 y 5. No obstante, es necesario hacer más pruebas para determinar el porcentaje de falsos positivos y falsos negativos que pudiera producir el sistema. Estos resultados se podrían combinar con otras técnicas de Inteligencia Artificial, como por ejemplo el análisis de decisiones, los cuales, permiten tomar decisiones de forma normativa y establecer la política de actuación más adecuada, incluso en los casos en que no es tan evidente y el juicio clínico del médico es incapaz de encontrar la mejor solución. Además el análisis de decisiones puede combinar de forma explícita y sistemática, las opiniones de diferentes expertos y los datos experimentales, tales como los datos de estudios publicados en la literatura médica (Díez-Vegas, 2007).

Conclusiones.

Las enfermedades Neurodegenerativas se han convertido un serio problema que repercute tanto en el paciente como en su entorno familiar y social. Su diagnóstico precoz es muy importante, tanto para conseguir una mayor eficacia en los tratamientos farmacológicos, como para comprender su situación. Creemos que esta investigación, que se encuentra en fases iniciales (Valls-Pedret, Molinuevo, & Rami, 2010), puede dar resultados prometedores en el diagnóstico de enfermedades neurodegenerativas y en especial en el diagnóstico de la enfermedad de Alzheimer que es el responsable del 50-70% de los casos de demencia según (Alzheimer Europe).

No obstante, esta investigación intenta poner de relieve la gran aportación que podrían tener el uso de técnicas de Inteligencia Artificial en un problema tan complejo como el diagnóstico de enfermedades neurodegenerativas, ya no sólo para su diagnóstico, sino también en la recomendación sobre los tratamientos más adecuados; haciendo énfasis en los tratamientos cognitivos dónde el uso de Técnicas de Inteligencia Artificial puede llegar a tener un gran impacto.

Bibliografía

(s.f.). Obtenido de Alzheimer Europe: http://www.alzheimer-europe.org

Díez-Vegas, F. (2007). Teoría probabilista de la decisión en medicina.

Neapolitan, R. (2004). *Learning Bayesian Networks.* Prentice Hall Series in Artificial Intelligence.

Fernández Martínez, M., Castro-Flores, J., Pérez-de las Heras, S., Mandaluniz-Lekumberri, A., Gordejuela, M., & Zarranz, J. (2008). Prevalencia de la demencia en mayores de 65 años en una comarca del País Vasco. *Revista de Neurología, 46*(2), 89-96.

Grasso, L., Díaz, M.C, & Peraita, H. (2009). Análisis preliminar de rasgos de definiciones de categorías semánticas del Corpus lingüístico de sujetos sanos y con Enfermedad de Alzheimer. Departamento de Psicología Básica 1. Facultad de Psicología. UNED.

Kjaerulff, U., & Madsen, A. (2007). *Bayesian Networks and Influence Diagrams.* (M. Jordan, J. Kleinberg, & B. Schölkopf, Edits.) Springer.

Peña-Casanova, J. (1999). Enfermedad de Alzheimer. Del diagnóstico a la terapia: conceptos y hechos. *Fundación "La Caixa"*, 10.

Peraita Adrados, H., & Grasso, L. (2010). Corpus lingüístico de definiciones de categorías semánticas de personas mayores sanas y con la enfermedad de Alzheimer. Una investigación transcultural hispano-argentina. *Fundación BBVA*.

Peraita, H., González-Labra, M., Sánchez, M.L., & Galeote, M. (2000). Batería de evaluación del deterioro de la memoria semántica. Psicothema, 12(2), 192-200.

Valls-Pedret, C., Molinuevo, J., & Rami, L. (2010). Diagnóstico precoz de la enfermedad de Alzheimer fase prodrómica y preclínica. *Revista de Neurología, 51*, 471-480.

Corpus and language policy: Iranian language policy towards English loanwords

Katarzyna Marszałek-Kowalewska

Adam Mickiewicz University, Poznań, Poland

Abstract

The aim of this article is to describe Iranian language policy towards English loanwords. It will focus on one particular semantic field – technology. Firstly, the history of Iranian language policy will be defined. Then, English loanwords in Farsi and their classification will be presented. Corpus-based study of technical English loanwords and their Farsi counterparts proposed by Iranian Language Academy is to be described. The objective of the conducted study is to compare the results from the Persian Linguistic Database and corpora compiled by the Author of this paper.The study attempts to verify the differences in usage between certain lexical borrowings and their equivalents. This usage relates to collocations, register and frequency. By means of compiled corpora the question about the successfulness of the Iranian language policy towards this particular semantic group will be addressed.

Keywords: corpus, loanwords, language policy, purism

Resumen

El objetivo de este artículo es describir la política lingüística persa concerniente a los préstamos del idioma inglés. Se centrará en un campo semántico determinado: la tecnología. Primero se definirá la historia de la política lingüística iraní. Luego, se expondrán los préstamos lingüísticos del inglés al idioma farsi y su clasificación. A continuación, se procederá a presentar un estudio de los préstamos técnicos del inglés en base a un corpus documental, y sus homólogos persas propuestos por la Academia de la Lengua Persa. La realización de este estudio tiene como objetivo comparar los resultados de la PLDB con el corpus copilado por la autora de este trabajo. El presente estudio intenta examinar las diferencias de uso entre ciertos préstamos y sus equivalentes. Mediante el corpus documental, se podrá plantear la cuestión del éxito de la política lingüística persa relativa a este campo semántico en particular.

Palabras Clave: corpus, préstamo léxico, politica lingüística, purismo

1. Introduction

the reason for undertaking the subject of Iranian language policy comes from the Author's interests in loanwords in Farsi which resulted in writing MA thesis "English borrowings in Farsi: a lexicography and corpus-driven study of technical vocabulary".

The Persian language, also called Farsi is said to have borrowed more than fifty percent of its vocabulary. The great majority of these loanwords is of Arabic origin. What is more, many of the Arabic loanwords have already become so established in Farsi that they are no longer perceived as borrowings. Apart from borrowings from Arabic, there are also early Turkish and Greek borrowings. Moreover, Farsi has also been influenced by European languages such as French, Russian, and finally, English. As the main topic of this paper is Iranian language policy and its attitude towards English loanwords it would be helpful to describe the history of Iranian language policy first.

2. History of the persian language policy

Persian language policy is characterised by heavy linguistic purism - the desire to get rid of foreign elements from the language and at the same time to mantain the language pure. The begining of language purism in Iran (former Persia) can be dated to 10th/11th century when philosophers as Avicenna or al Biruni called for limited usage of Arabic elements. However, due to space limitations this paper will not go into detales in the description of remote history of Iranian language policy but it will focus on the XXth century. Basically, the history of Persian language policy can be divided into two main periods: before and after the Islamic Revolution. The key to this division is the main target of Iranian purists. Before the Islamic Revolution, the Arabic language was the main culprit contaminating the Persian language and it was the main target of purification activities. In 1924, Shah Reza Pahlavi started his reign from the modernisation of army. As it soon turned out, the Persian language lacked a lot of words for denotating new technological inventions. Thus, one of the first Shah's order was to establish a committee that would coin new Persian terms to fill that gap. The committee's performance was quite impressive as during one year it created almost 400 words, the majoraty of which is still in use today. And what is important, none of its members was a linguist. These were the fundations for the language academy. In 1935 on Shah's order the first Language Academy (*Farhangestane avval*) was established. Its main aims were to create a dictionary, study Iranian dialects and first of all replace foreign elements with native Persian ones. The six steps of the Academy proposed by its first president Foroughi, clearly describes Iranian language policy:

- Avoid an Arabic word whenever there is a close Persian word.
- When you have a common borrowing and an unknown Farsi equivalent, the latter should be popularised.
- If there is no equivalent in Persian, Farsi word should be created.
- If there is no equivalent, use the borrowed until Persian equivalent is not created.

- When there is no Persian equivalent and the concept expressed by a borrowing belongs to the material domain[204], accept the loanword.
- If there is a foreign word that belongs to the spiritual domain, then a Persian equivalent should be manufactured.

Foroughi was forced by the Shah to resign at the end of 1935. Pahlavi was dissatisfied with the slow progress the Academy made in order to purify the Persian language (Mehrdad, 1998: 21).

Presidents of the Academy changed very often. What is more, none of its aims was actually achieved. In 194 Shah had to abdicate and flew form the country. Consequently, Langugae Academy stopped working. In 1971 Shah's son established *Farhangestane davvam* – second Language Academy. Yet, due to hectic political situation in Iran, second langugae academy did not achieve anything notable.

The year 1979 – the year of the Islamic Revolution in Iran had a great impact on the political scene and as a result on the language policy. The 1979 Revolution (or the Islamic Revolution) resulted in overthrowing Shah's regime and introducing the Islamic Republic under Ayatollah Khomeneini. Ayatollah became the Supreme Leader and in October 1979 the country approved new, theocratic constitution that in articles 15 and 16 states the position of Persian and Arabic languages. Article 15 of the Constitution of the Islamic Republic of Iran states:

> The official language and script of Iran, the lingua franca of its people, is Persian. Official documents, correspondence, and texts, as well as text-books, must be in this language and script. However, the use of regional and tribal languages in the press and mass media, as well as for teaching of their literature in schools, is allowed in addition to Persian.

It does confirm the importance and superiority of Farsi as national language. Article 16 of the Constitution confirms the use and importance of the Arabic language:

> Since the language of the Qur'an and Islamic texts and teachings is Arabic, and since Persian literature is thoroughly permeated by this language, it must be taught after elementary level, in all classes of secondary school and in all areas of study.

Thus, Arabic being so far the target of purists' activities, in 1979 became officially recognized language being indispensable part of Iranian education system and therefore life.

Third Language Academy *Farhangestane sevvam* was established after the Third Supreme Council of the Iranian Revolution in 1990. The members, twenty-five language experts and professors, among whom were also two Tajiks, were preoccupied with studying grammar, orthography, manuscripts and various Iranian dialects. They also identified the channels responsible for frequent neologisms. The influence of the Internet and the media was given as the main culprit. The policy of the third Language Academy is as follows:

204 According to Mehrdad there are two domains shaping Iranian identity: material and spiritual one. The former reflects technology whereas the latter is identified with Shii Islam and language.

1. In coining and choosing a new word, Persian phonetic rules and learned speakers' way of talking and Islamic points of views should be regarded as the main criterion.
2. Phonetic rules should be obeyed according to the Persian way of talking.
3. New words should follow the Persian grammatical rules for coining nouns, adjectives, verbs and so on.
4. New words should be chosen or coined out of the most common or frequent words that have been used since 250 AD.
5. New words can be chosen from among the most frequent and common Arabic words as used in Persian.
6. New words can be chosen from the Middle and Old Persian stages of the language.
7. There should be only one equivalent in Persian for any of the Latin words, particularly technical ones.
8. It is not necessary to adapt or create new Persian words for those Latin words which have been used internationally and globally (Farhangestan-e Zaban 2001 as quoted in Monajemi, 2010: 5).

It is important to point to the fifth point on this list. As has been stated, the main target of the first and second Academy was the Arabic language. The current language policy is, for obvious reasons, no longer hostile towards this language. Iran is the Islamic Republic and Arabic is the holy language of Islam. Today's purism perceives Arabic as a way of purifying Farsi. So far, the Academy has been successful in issuing seven lists of *Collection of Terms Approved*. These are, as the name suggests, words that are allowed to be used and, what is more, should be used by the speakers of Farsi. The first collection was published in 2003 and the last one in May 2010. Thus, without judging the successfulness of this work, it has to be stated that purists in the Islamic Republic in Iran are quite productive. In 2006 the Iranian president Mahmoud Ahmadinejad ordered the government and all official Iranian bodies to use only Persian words approved by the Academy of Language instead of foreign ones. Changes introduced by Ahmadinejad are mandatory for all schoolbooks, documents and newspapers (Dujardin, 2006).

3. Research

In 2008 the Author of this paper decided to check if the Iranian language policy is succesfull in replacing English loanwords with native Persian ones. To this end the list of English loanwords in Farsi had to be compiled. As the Author is not native speaker of Farsi and does not feel in power to decide which word can be counted as an English loanword, dictionary of western vocabulary in Persian *Dictionnaire de l'Européanisme Persan,* compiled by an Iranian lexicographer Monshid Moshiri, was used. Unfortunaltely, this dictionary comes from 1993 so it is not very up-to-date. Yet, it is the only dictionary in which English loanwords were listed. On the basis of this dictionary, a list of 586 English loanwords was prepared. Compiling the list, it turned out that loanwords tend to

fall into certain semantic grups: technology, education, kitchen devices, taboo, sport, food, medicine, vehicles and car devices. What follows now are examples of each semantic field:

Food:

- آوکادو [āvokādo] 'avocado'
- استارت [estārt] 'starter'
- استیک [estesk] 'steak'

Sport:

- کریکت [keriket] 'cricket'
- گل [gol] 'goal'
- گلر [goler] 'goal keeper'

Vehicles and car devices:

- اتوکار [otokār] 'autocar'
- تراموا [terāmvā] 'tram'
- جک [jak] 'jack'

Education:

- ام.اس [em-es] 'Master of Science'
- پی.ای.چ.دی [pi-eyč-di] 'Ph.D.'
- اسکول [eskul] 'school'

Kitchen devices:

- تندرایزر [tenderāyzer] 'tenderizer'
- چاپاستیک [čāp-estik] 'chopsticks'
- سینک [sink] 'sink'

Technology:

- آی.سی [āy-si] 'integrated circuit'
- بایت [bāyt] 'byte'
- بایناری [bāynari] 'binary'

Medicine:

- آی.سی.یو [āy-si-yu] 'intensive care unit'
- اودیپ [odip] 'Oedipus complex'
- اورولوژیست [orologist] 'urologist'

Months:

- اپریل [eypril] 'April'
- آگوست [āgust] 'August'
- جولای [julāy] 'July'

Taboo:

- ویسکی [wiski] 'whiskey'
- بالکتهای [nāyte-kelāb] 'nightclub'
- کلایمکس [kelāymaks] 'climax'

Then the decision was taken to study only one group – technological vocabulary. The next step was to compare list of technical English loanwords with the *Collections of Terms Approved* by the Academy and to look for Persian equivalents. On this basis ten pairs of loanwords and their equivalents were prepared. They are: *computer, digital, file, site, data, frezer, printer, user, lens* and *fax*. The decision was taken to check the frequency, register and collocations in order to answer the question if the Iranian language policy is succesfull in attempt to purify Farsi. The tool used was Persian Linguistic Database (PLDB) – corpus of general modern Farsi cosisting of 50 milion words prepared by profesor Mustafa Assi at the Institute for Humanities and Cultural Studies (IHCS), Iran. The general results of the study were as follow: we can observe the advantage of English loanword over its Farsi counterpart in two cases:

- Firstly, when loanword was incorporated into morphological and phonological system of Farsi,and
- Secondly, when it was contrasted with newly coined Farsi counterpart.

On the other hand, Farsi counterpart had the advantage over a loanword when an already existing Persian word extended its meaning. It also should be noted that all corpus hits were from press register. Last year the Author decided to conduct the research again, yet this time on her own corpus of the Persian lanugae. Finally, two corpora were compiled. The first one is press corpus of 8 million words. It contains articles from newspapers, magazines and news agencies from 2010. Second corpus consists of 360,000 words and comprises technical blogs, technical magazines and technical journals again only from 2010. The idea of compiling two corpora was to check whether the proportion in usage differs in more general, press corpus and specialised, technology corpus. The procedure was as follows:

1. Compare hits from PLDB from 2008 with hits from press corpus from 2010
2. Compare hits from press corpus with hits from technology corpus (both from 2010)

The overall results are presented in the table below:

Table 1. Comparison of corpora results

Nr	Studied term	2010 Press corpus hits	2010 Technical corpus hits	2008 PLDB hits
1.	کامپیوتر [kāmpyuter]	115	277	111
	رایانه [rāyāne]	277	10	304
2.	دیجیتال [dijital]	0	0	0
	رقمی [raqmi]	209	11	65
3.	فایل [fāil]	64	98	46
	پرونجا [parvanjā]	0	0	0
4.	سایت [sāyt]	1598	516	229
	ایستگاه [istgāh]	656	11	119
5.	دیتا [deytā]	91	8	3
	داده [dādeh]	7484	475	2605
6.	فریزر [frizer]	11	0	25
	یخ زن [yakh zan]	0	0	0
7.	پرینتر [printer]	0	5	0
	چاپگر [čāpgar]	28	5	9
8.	یوزر [yuzer]	0	0	1
	کاربر [kārbar]	731	618	65
9.	لنز [lenz]	40	19	15
	عدسی [adasi]	16	2	7
10.	فاکس [fāks]	33	4	62
	دورنگار [durnagār]	2	0	0

However, taking into consideration corpora sizes it was not possibile to compare the raw frequencies. Thus, z-test for statistical significance was adopted. It measures the proportions from two independent groups to determine if they are statisticaly different from one another. The formula of the z-test is as follows:

$$Z = \frac{\left|\frac{O_{11}}{O_1} - \frac{O_{21}}{O_2}\right|}{\sqrt{\frac{O_1}{n}\frac{O_0}{n}}} \sqrt{\frac{O_1 O_2}{n}}$$

Only the results with 95% confidence level were counted as statisticaly significant.

The overall results of the comparative research of 2008 and 2010 showed statistical significance in four cases. In two of them, namely *user* and *fax* we can observe the advantage of English loanword. In other two, *site* and *data* Farsi word proved to do better.

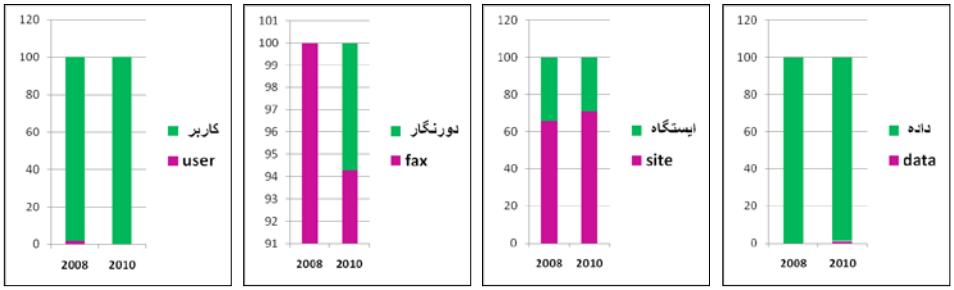


Figure 1. Comparison of statisticaly significant 2008 and 2010 hits.

When it comes to the comparision of press and technology corpus, z-test showed statistical significance again in four cases, all showing advantage of English loanwords. They are *computer, lens, printer* and *site*.

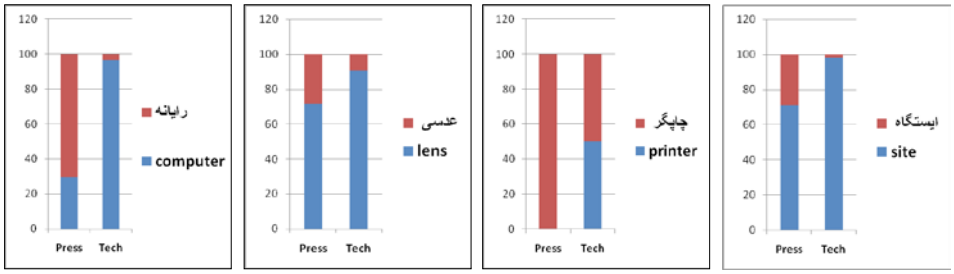


Figure 2. Comparison of statisticaly significant press and technology hits

4. Conclusion

Originally this study was supposed to examine the successfulness of the Iranian language policy towards technical English loanwords. As the research proceeded it turned out that there are certain problems that can make the results vague. The most important problem is the existence of homonymy and polysemy. The fact that the Farsi word داده [dādeh] has

more readings than the studied *data* makes the examining of corpus examples impossible. What is more, it seems to be connected with another problem: in case when an already existing Persian word extended its meaning, are we actually talking about semantic extension or about semantinc borrowing? If the latter was to be true, in all examples that were examined here, we would observe advantage of English loanwords. To be more precise, examples of *data* and *site* in the comparison of PLDB and press corpus hits, showed the advantage of native word. Yet, these Persian words had existed in Farsi long before they gained the studied meaning. The word داده [dādeh] is the past participle of the verb دادن [dādan] 'to give' One solution to this problem could be studying radial basis function network[205]. To be more specific, if a studied meaning (in this case *data*) can be incorporated or linked with the central meaning of a given word (in this case '*given'*) it may count as an extension of meaning. If not, it could be thought of as a semantic borrowing. The same case is with the Polish word *dane* 'data'. It is, as in Farsi, past aprticiple of the verb *dawać* 'to give'. Both Polish and Farsi *data* seem to be linked with the central meaning of the past participle. Yet, the idea of radial basic function network in examining semantic extension and semantic borrowing needs to studied in more details.

References

Assi, M. Persian Linguistic Database. (http://pldb.ihcs.ac.ir/). Constitution of the Islamic Republic of Iran. (1979).

Dujardin, G. (2006). Elastic Loaves! Get Yer Fresh Hot Elastic Loaves!.

(http://dujardin.blogspot.com/2006/07/elastic-loaves-get-yer-fresh-hot.html) (date of access: 1 Mar. 2009).

Mehrdad, K. (1998). Persian Nationalism and the campaign for language purification. *Middle Eastern Studies* 34, 2: 9-36.

Monajemi, E. (2010). Can ethnic minority languages survive in the context of global development? (http://www.sil.org/asia/ldc/parallel_papers/ebrahim_monajemi.pdf) (date of access: 5 Dec. 2010).

Moshiri, M. (1993). *Dictionnaire de l'Européanisme Persan*. Teheran: Alborz. The Academy of Persian Language and Literature. (2005). *A collection of termas approved*. [No indication of publisher].

205 The "Author is grateful to her colleague Barbara Konat who suggested using radial basic function network.

Utilización de córpora textuales para la extracción de modificadores contextuales de valencia para tareas de Análisis de Sentimiento

Antonio Moreno Ortiz, Chantal Pérez Hernández, Rodrigo Hidalgo García

Facultad de Filosofía y Letras
Universidad de Málaga

ABSTRACT

El desarrollo de herramientas de Análisis de Sentimiento requiere de la codificación del lenguaje evaluativo en términos de polaridad. Una de las principales dificultades es la existencia de variables que modifican la interpretación positiva o negativa de lo expresado. Si bien algunos fenómenos discursivos, como la ironía, son muy difíciles de definir en términos formales, otros modificadores contextuales de la valencia sí muestran rasgos identificables. Un claro ejemplo es la negación, que invierte la polaridad de la unidad léxica a la que modifica. Sin embargo, hay formas de inversión de la valencia bastante más sutiles, cuya detección no es tan fácil y requieren igualmente ser codificadas en forma de reglas de contexto en las herramientas de análisis automatizado. En este trabajo describimos nuestra experiencia en el empleo de córpora en la consecución de este objetivo dentro del proyecto Sentitext: una herramienta de Análisis de Sentimiento para el español.

Keywords: análisis del sentimiento, minería de opinión, modificadores contextuales de valencia, expresiones multi-palabra, colocaciones.

ABSTRACT

The development of Sentiment Analysis tools requires the codification of evaluative language in terms of its polarity. One of the main difficulties is the existence of variables that modify the polarity, either positive or negative, of what is being said. Although some discourse phenomena, such as irony, are difficult to account for in formal terms, there are many other Contextual Valence Shifters (CVS) that do show certain formally identifiable traits. Negation is probably the most straightforward example of valence inversion, but there are certainly more subtle means in which valence inversion can be conveyed, which also need to be accounted for and defined in terms of Contextual Valence Shifters. In this paper, we describe our experience in the use of corpora in order to attain this objective for Sentitext: a Sentiment Analysis tool for Spanish.

Keywords: sentiment analysis, opinion mining, contextual valence shifters, multiword expressions, collocations

1. Análisis de Sentimiento

El análisis de sentimiento es una vertiente de la minería de textos que ha recibido gran atención en los últimos años debido en gran parte a la aparición de la Web 2.0 y la explosión que ésta ha supuesto en cuanto a grandes cantidades de texto evaluativo generado por los usuarios de la Red en torno todo tipo de productos y servicios, cuyo análisis y escrutinio es de obvio interés para empresas y organizaciones de diversa naturaleza. En un sentido amplio, consiste en el empleo de herramientas computacionales para la detección y análisis del lenguaje valorativo, en términos de polaridad.

Pang y Lee (2008) repasan el avance de la investigación en el campo del análisis de sentimiento en la última década, donde el principal foco de atención ha sido el análisis de valoraciones de usuarios de películas (Pang, Lee & Vaithyanathan, 2002), productos electrónicos de consumo (Dave et al., 2003), pero también el análisis del mercado financiero (Das y Chen, 2001), de discursos políticos (Thomas et al., 2006) o las valoraciones de usuarios en webs de evaluación de productos de carácter general, como Epinions (Turney, 2002; Taboada y Grieve, 2004). En español la actividad en el campo es mucho más reducida, aunque hay notables trabajos, como los de Cruz et al. (2008) y Boldrini et al. (2009). Moreno Ortiz et al. (2010a) analizan las valoraciones que los usuarios de la web *tripadvisor.es* hacen de los hoteles que visitan a través de una nueva herramienta llamada Sentitext (Moreno Ortiz et al., 2010b), la cual utilizamos también para el estudio que llevamos a cabo en este artículo.

La visión más extendida en el ámbito del análisis de sentimiento es que éste es altamente dependiente del dominio (Aue & Gamon 2005). Es decir, determinadas palabras y unidades fraseológicas que en un dominio determinado serían contempladas como afectivamente neutras pueden adquirir una determinada *polaridad*, u *orientación semántica*, en ese dominio determinado. Por tanto, los esfuerzos han estado enfocados a construir recursos léxicos específicos de dominios determinados, empleando para ello técnicas de aprendizaje automático, con o sin supervisión (Turney & Littman, 2002). Por el contrario, la herramienta que hemos desarrollado y que usamos en este artículo, Sentitext, es independiente del dominio y no usa ningún algoritmo de aprendizaje como tal sino que cuenta con bases de datos léxicas de extensa cobertura cuyo contenido se ha adquirido de forma manual.

Sentitext se nutre de tres fuentes de datos fundamentales: (1) el léxico de palabras individuales, (2) el léxico de frases y (3) las reglas de contexto. Las tres han sido desarrolladas de forma manual, pero utilizando herramientas creadas al efecto para facilitar el proceso de adquisición. El lexicón de palabras individuales consta actualmente de unas 10.000 palabras con carga afectiva, marcadas con su polaridad con uno de estos 4 valores: -2 (muy negativa), -1 (negativa), 1 (positiva), 2 (muy positiva).

Una parte de especial importancia en nuestro trabajo y que determina en gran parte la efectividad del sistema de análisis de sentimiento ha sido la inclusión de locuciones o expresiones multi-palabra. Como señalan Wilson et al. (2009), a veces se usan palabras positivas en frases negativas (p. ej.: "abuso de confianza") y viceversa. En nuestro caso, decidimos desde un principio adquirir un lexicón de locuciones de amplia cobertura y no

específico del dominio. Dicho lexicón contiene en la actualidad más de 17.000 entradas, aunque, a diferencia del lexicón de palabras individuales, no todas tienen carga afectiva.

El tercer componente de nuestro sistema, las *reglas de contexto*, está inspirado en gran medida en la propuesta de Polanyi y Zaenen (2006), en lo que ellas denominan *Context Valence Shifters (CVS)*, o modificadores contextuales de la valencia. Estas reglas sirven para formalizar la modificación de valencia que trae consigo una estructura determinada y codifican además el resultado de la aplicación de cada regla. De entre las 350 reglas de contexto que hemos compilado hasta la fecha, algunas son muy simples y generales (por ejemplo, la presencia del adverbio "muy" precediendo a un adjetivo con carga afectiva intensifica su valencia en el mismo sentido), aunque otras son mucho más complejas y específicas.

2. Selección de palabras clave o seed sets

Diversos autores han ensayo métodos diferentes de asignación automática de la polaridad a las palabras, haciendo uso de diferentes algoritmos de aprendizaje.
Hatzivassiloglou & McKeown (1997) usaron un sistema automatizado para asignar polaridad a las palabras basado en patrones de co-aparición estadísticamente frecuente. Wiebe (2000) emplea córpora etiquetado con anotaciones manuales para obtener, al igual que en nuestro caso, no sólo polaridad sino también gradación. Otro enfoque ha consistido en la utilización de un conjunto inicial de palabras (*seed set*) a partir del cual se genera un léxico de gran cobertura mediante su utilización junto a un diccionarios de sinónimos (Pitel & Grefenstette, 2008).

Nuestra aproximación particular estuvo motivada precisamente por esta última metodología. En primer lugar, adaptamos al español el *seed set* de Pitel & Grefenstette (2008), originalmente en francés, consistente en 44 pares de palabras antónimas con carga afectiva y, en segundo lugar, integramos este conjunto inicial en una aplicación junto a un diccionario de sinónimos, en concreto el Open Office Thesaurus, trabajando de forma cíclica con un proceso de adquisición semi-áutomático y supervisado.

La base de datos de palabras de Sentitext no sólo asigna polaridad a la palabra o expresión multi-palabra en cuestión sino que también distingue gradación (Hatzivassiloglou & McKeown 1997) en el rango -2 a +2. En una etapa posterior, las palabras de la base de datos fueron clasificadas y asignadas a una emoción básica, siguiendo la clasificación propuesta por Parrott (2001), de forma que es posible encontrar todas las palabras positivas/negativas pertenecientes a la emoción *rabia, dolor, orgullo, deseo*, etc.

3. Adquisición de reglas de contexto

El proceso de adquisición de las reglas de contexto fue la que desde el principio ha estado guiada por el análisis exhaustivo de córpora. Para el análisis que aquí presentamos hemos seleccionado una serie de palabras clave, en su mayoría sustantivos, de nuestro *seed set* original. Estos sustantivos se han empleado como términos de búsqueda en la

identificación de patrones de co-aparición con un verbo o sustantivo deverbal, con el objetivo de analizar, en el corpus, cómo ciertos verbos pueden invertir, intensificar o modificar la polaridad de estos sustantivos. Todos ellos estaban incluidos en la base de datos de Sentitext con emoción etiquetada según Parrott (2001), tenían una valencia alta [+2 o -2] y una alta frecuencia en nuestro corpus de español:

Tabla 1. Sustantivos analizados en corpus con indicación de valencia y clasificación según Parrott (2001)

Tipo de Emoción	Sustantivos y valencias asignadas
rage	violencia [-2], ataque [-2]
neglect	ignorancia [-2], deterioro [-2]
nervousness	amenaza [-2], preocupación [-2]
pride	dignidad [2], éxito [2]
irritation	enfado [-2], queja [-2], provocación [-1]
cheerfulness	alegría [2], felicidad [2]
lust	ambición [-1]
suffering	dolor [-2], malestar [-1]
sadness	tristeza [-2], desolación [-2]
affection	solidaridad [2], empatía [1]
contentment	tranquilidad [2], satisfacción [1]
exasperation	aburrimiento [-1], cansancio [-1]
horror	pánico [-2], devastación [-2]
disgust	corrupción [-2], desprecio [-2]
shame	culpa [-2], equivocación [-1]
relief	alivio [2], seguridad [2]
optimism	confianza [1], esperanza [1]
envy	envidia [-2], celos [-2]
disappointed	decepción [-2], fracaso [-2]
sympathy	ayuda [1], compasión [1]
longing	ilusión [2], ensueño [1]

Aunque existen multitud de modificadores contextuales de la valencia dentro del contexto lingüístico más inmediato a la palabra que determinan cambios en la polaridad, como por ejemplo la introducción del operador "no" (*esto es bueno à esto no es bueno*), hay muchos otros que no son tan evidentes, ya que no todas las formas de inversión de la valencia son tan fácilmente identificables como una simple negación. Esto hace necesario el trabajo de detección y catalogación de modificadores contextuales de la valencia que hemos llevado a cabo, ya sea con las combinaciones de verbo + sustantivo (por ej., "carecer de dignidad", "vulnerar las leyes", "hacer frente a la crisis" o "superar el problema"), o las expresiones multipalabra y colocaciones en las que alguno de sus componentes tiene carga afectiva,

tales como "ser un rayo de luz para el enfermo", "hacer un flaco favor al progreso", etc. Si prestamos atención a la diversidad de construcciones gramaticales y léxicas implicadas, incluso en estos pocos ejemplos, parece obvio que la tarea de identificar, clasificar y definir estos modificadores contextuales de valencia, no es trivial.

4. Resultados del análisis

Hemos estudiado las búsquedas de nuestras palabras clave (Tabla 1), ordenándolas alfabéticamente a partir de la segunda palabra precedente a la palabra clave (*Sort* 2L) para ayudar en nuestra labor de localización de patrones de co-aparición de verbos con sustantivos con más facilidad. A continuación, hemos analizado todos los casos en los que un verbo precede al sustantivo y modifica su valencia, clasificando estas combinaciones de palabras en alguna de las siguientes categorías:

Tabla 2. Patrones de co-aparición (verbo + sustantivo) modificadores de valencia del sustantivo.

Modificación de valencia en sustantivos positivos
CO-APARICIÓN CON VERBO O SUSTANTIVO DEVERBAL. INVERSIÓN [+ a -]
CO-APARICIÓN CON VERBO O SUSTANTIVO DEVERBAL. NEUTRALIZACIÓN [+ a 0]
CO-APARICIÓN CON VERBO O SUSTANTIVO DEVERBAL. INTENSIFICACIÓN [+ a ++]
Modificación de valencia en sustantivos negativos
CO-APARICIÓN CON VERBO O SUSTANTIVO DEVERBAL. INVERSIÓN [- a +]
CO-APARICIÓN CON VERBO O SUSTANTIVO DEVERBAL. NEUTRALIZACIÓN [- a 0]
CO-APARICIÓN CON VERBO O SUSTANTIVO DEVERBAL. INTENSIFICACIÓN [- a --]

La primera dificultad con la que nos encontramos es que hay palabras clave cuya clasificación como palabras positivas o negativas es incierta y altamente dependiente del contexto. Consideremos los siguientes ejemplos:

a) *El nuevo presidente tiene una enorme ambición*.

b) *una asquerosa ambición por el dinero*.

En el ejemplo a), la palabra *ambición* está ligada a la emoción *optimism* mientras que en b) tiene un significado más próximo a la codicia (*lust*). El hecho de que Sentitext clasifique esta palabra como negativa se rige por la mayor frecuencia de casos negativos, pero habrá que observar con detenimiento cómo el contexto modifica su valencia. Tras analizar la totalidad de las palabras clave seleccionadas, hemos obtenido extensas listas de combinaciones de *verbo* + *sustantivo* que modifican la valencia del sustantivo.

Resumimos a continuación las categorías más destacadas.

4.1 Verbos modificadores de la valencia exclusivos de sustantivos positivos/negativos

Una de las primeras conclusiones que podemos extraer es que, si bien algunos modificadores contextuales de la valencia son comunes a los sustantivos positivos y negativos, existe un numeroso grupo de verbos que modifican la valencia de sustantivos exclusivamente positivos, así como otro grupo modificador de sustantivos exclusivamente negativos. En general, el verbo que modifica la valencia de un sustantivo positivo indica *oposición*, de forma que invierte la polaridad del sustantivo a negativo, por ejemplo, *romper el acuerdo*. El verbo *romper* pertenece a la categoría de verbos que invierten de positivo a negativo, pero nunca al revés. De hecho, los verbos que invierten de negativo a positivo suelen ser verbos de *superación*, por ej. *resolver el conflicto.*

Otro de los verbos más comunes de inversión de positivo a negativo (INV [+ a -]) es *perder*. Este verbo precede en el corpus a sustantivos positivos, como por ejemplo: *puestos de empleo o confianza*. Sin embargo, si revisamos la lista de modificadores INV [- a +], no encontraremos *perder,* pero sí *vencer (violencia, pánico)*. Es decir, existen parejas de verbos modificadores de la valencia que apuntan en sentido opuesto, unos hacia *oposición* y otros hacia *superación*. A continuación se muestran algunos ejemplos de verbos comunes de *superación* (Tabla 3) y *oposición* (Tabla 4) que invierten la valencia del sustantivo con el que aparecen:

Tabla 3. Verbos modificadores de valencia (INV [– a +]).

Modificación de la valencia del sustantivo por verbos específicos de inversión [- a +] (Indican superación, esfuerzo por cambiar o invertir la situación):	
VERBO	SUSTANTIVO
acabar con	crisis, problema, conflicto, violencia, dolor, corrupción
afrontar	crisis, problema, conflicto, violencia, ataque
combatir	inflación, delincuencia, crisis, enfermedad, violencia, amenaza, ataque, ignorancia, dolor, aburrimiento, corrupción
evitar	inflación, delincuencia, crisis, problema, violencia, amenaza, deterioro, provocación, dolor, cansancio, pánico, devastación, corrupción, fracaso
superar	crisis, problema, malestar, dolor, tristeza, pánico, corrupción, decepción
resolver	crisis, problema, conflicto, paradoja, violencia, amenaza, queja

Tabla 4. Verbos modificadores de valencia (INV [+ a -]).

Modificación de la valencia del sustantivo por verbos específicos de inversión [+ a -] (Indican oposición, trabar o dificultar):	
VERBO	SUSTANTIVO
dañar	alianza, paz social, dignidad, confianza
destruir	puestos de empleo, dignidad, felicidad, confianza
romper	acuerdo, tregua, paz, dignidad, éxito, solidaridad, tranquilidad, seguridad, confianza
perder	Protagonismo, paz, dignidad, alegría, seguridad, confianza, esperanza, ilusión

4.2 Verbos modificadores de la valencia comunes a sustantivos positivos/negativos

La categoría semántica de los verbos que configuran este grupo puede considerarse diferente de los grupos semánticos de *oposición/superación* mencionados en 3.1. Se trata, por un lado, de verbos que expresan *eliminación* y por otro lado *disminución*. En ambos casos nos encontramos con estos modificadores de la valencia comunes a nuestro *seed set* de palabras positivas y negativas (véase sección 2). El motivo es que *la eliminación, la desaparición o el fin* de algo bueno, automáticamente convierte este momento en negativo y viceversa. De esta misma forma, *la reducción o disminución* de algo malo nos da como resultado una inversión de la valencia a positivo. Estos son sólo algunos ejemplos de entre muchos otros.

Tabla 5. Verbos de inversión bidireccional con sentido de eliminar.

Modificación de la valencia del sustantivo por verbos de inversión bidireccional [+ a -] y [- a +] (Indican eliminación):	
VERBO / SUSTANTIVO DEVERBAL	SUSTANTIVO
desaparecer / desaparición de	puestos de empleo, solidaridad, confianza, esperanza [+ a -] barreras, violencia, amenaza, dolor [- a +]
eliminar / eliminación de	puestos de empleo, ilusión [+ a -] dolor, cansancio, pánico, corrupción [- a +]
finalizar	acuerdo, tregua, paz [+ a -] violencia, dolor [- a +]

Modificación de la valencia del sustantivo por verbos de inversión bidireccional [+ a -] y [- a +] (Indican disminución):	
VERBO	SUSTANTIVO
disminuir	salario, ayuda, seguridad, confianza [+ a -] violencia, dolor, aburrimiento, fracaso [- a +]
reducir	puestos de empleo, salario, éxito, alegría, esperanza [+ a -] dolor, cansancio, riesgo, ineficiencia, amenaza, corrupción [- a +]
finalizar	acuerdo, tregua, paz [+ a -] violencia, dolor [- a +]

Tabla 6. Verbos de inversión bidireccional con sentido de disminución.

Por último, deberíamos detenernos en una categoría de verbos que tienen diferente efecto al anteponerse a sustantivos positivos o negativos. En concreto, se trata de *verbos aumentativos*. Consideremos los siguientes ejemplos:

1) exceso de (éxito, alegría, felicidad, ambición, solidaridad, tranquilidad, seguridad, confianza)

2) exceso de (dolor, corrupción, celos)

Mientras el exceso de algo bueno, como en 1, puede llegar a poner la situación en contra (INV [+ a -]), el exceso de algo malo, si bien menos frecuente como observamos en el ejemplo 2, va a producir un efecto intensificador, en este caso INT [- a - -]. Los límites de este artículo no nos permiten describir la intensificación que determinados verbos producen en el componente valorativo de los sustantivos con los que co-aparecen, sin embargo, en general podemos entender que, irónicamente, el exceso de algo bueno nunca se entiende como algo positivo: "exceso de confianza" o "cegarse de felicidad".

5. Conclusiones

Hemos de tener en cuenta que este trabajo se centra fundamentalmente en un tipo particular de modificación de la valencia, concretamente la inversión de [+ a -] y [- a +], y aun así, hemos podido producir y clasificar multitud de patrones de co-aparición verbo-sustantivo del que sólo hemos podido mostrar unos pocos. Además, nos hemos centrado únicamente en la modificación de valencia por medio de un verbo o sustantivo deverbal, dejando de lado una multitud de posibilidades de modificación postpositiva, no sólo del sustantivo sino aplicable también a otras categorías de palabras. Parece evidente, por tanto, que la integración de la reglas de contexto a Sentitext suponen una ingente cantidad de trabajo adicional, si bien es cierto que ello implicaría, a la vista de nuestros resultados,

una mejora muy sustancial del rendimiento de la herramienta. Por otro lado, Sentitext pasaría a convertirse en la primera herramienta de análisis de sentimiento del español que hace uso de los CVS como reglas de contexto, un trabajo que podría hacerse extensible en el futuro a otras lenguas.

Referencias

Aue, A., & Gamon, M. (2005). Customizing Sentiment Classifiers to New Domains: A Case Study. Presented at the Recent Advances in Natural Language Processing (RANLP), Borovets, Bulgaria.

Cruz, F., Troyano, J. A., Enriquez, F., & Ortega, J. (2008). Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. *Procesamiento del Lenguaje Natural*, (41), 73-80.

Das, S. R., & Chen, M. (2001). Yahoo! for Amazon: Sentiment Parsing from Small Talk on the Web. *Proceedings of the 8th Asia Pacific Finance Association Annual Conference*.

Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: opinion extraction and semantic classification of product reviews. *Proceedings of the 12th international conference on World Wide Web* (págs 519-528). Budapest, Hungary: ACM.

Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics* (págs 174-181). Madrid, Spain: Association for Computational Linguistics.

Boldrini, E., & Balahur, A., Patricio Martínez-Barco & Montoyo, A. (2009). EmotiBlog: an Annotation Scheme for Emotion Detection and Analysis in Non-traditional Textual Genres. *Proceedings of The 2009 International Conference on Data Mining* (págs 491-497). Presented at the DMIN 2009, Las Vegas, USA: CSREA Press.

Moreno Ortiz, A., Pérez Pozo, Á., & Torres Sánchez, S. (2010a). Sentitext: sistema de análisis de sentimiento para el español. *Procesamiento de Lenguaje Natural*, *45*, 297-298.

Moreno Ortiz, A., Pineda Castillo, F., & Hidalgo García, R. (2010b). Análisis de Valoraciones de Usuario de Hoteles con Sentitext: un sistema de análisis de sentimiento independiente del dominio. *Procesamiento de Lenguaje Natural*, *45*, 31-39.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10* (págs 79-86). Association for Computational Linguistics.

Parrott, W. (2001). *Emotions in Social Psychology*. Philadelphia: Psychology Press.

Pitel, G., & Grefenstette, G. (2008). Semi-automatic Building Method for a Multidimensional Affect Dictionary for a New Language. *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008*.

Polanyi, L., & Zaenen, A. (2006). Contextual Valence Shifters. *Computing Attitude and Affect in Text: Theory and Applications* (págs 1-10). Dordrecht, The Netherlands: Springer.

Taboada, M., & Grieve, J. (2004). Analyzing Appraisal Automatically. *AAAI Technical Report SS-04-07* (págs 158-161). Presented at the American Association for Artificial Intelligence Spring Symposium on Exploring Attitude and Affect in Text, Stanford.

Thomas, M., Pang, B., & Lee, L. (2006). Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. *Proceedings of EMNLP* (págs 327-335). Presented at the EMNLP.

Turney, P. D. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)* (págs 417-424). Presented at the ACL 2002, Philadelphia, USA.

Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.*, *21*(4), 315-346. doi:10.1145/944012.944013

Wiebe. J. M. (2000). Learning subjective adjectives from corpora. *Proceedings of the 17th National Conference on Artificial Intelligence* (pág 268–275). Menlo Park, CA: AAAI Press.

Wilson, T., Wiebe, J., & Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, *35*(3), 399-433.

Using computer- based corpora to create learning materials for tourism (ESP)

Alicia Ricart Vayá, María Alcantud Díaz
Universitat de València Estudi General, Departmento de Filolgía Inglesa y Alemana

Abstracts

The present article adopts a computerized frequency-driven approach in the analysis of frequently-used prepositions. Our purpose is to identify the errors made by first-year students of Tourism when writing essays. Our students were asked to write a film review of about 200 words after watching the film "The Terminal. We decided to investigate the errors made when using the prepositions at, in and on. The corpus was composed of 50 student's essays, which we analyzed using WordSmith Tools 5 *both quantitatively and qualitatively. That is, we retrieved the prepositions analyzing their frequencies and concordances in order to look for non-native combinations. Our final aim was to create a series of exercises by using the occurrences of the prepositions as the main basis of our exercises. As a general conclusion, we believe that corpus analysis could be an effective tool in order to create tailor-made activities and teaching materials.*

Key Words: Tourism, corpus analysis, WordSmith Tools*, teaching materials,* Exeleraning

El presente artículo adopta un sistema computarizado de análisis de frecuencias basado en el estudio de preposiciones usadas frecuentemente. Nuestro objetivo es identificar los errores realizados por estudiantes de primero de Turismo al escribir redacciones. Pedimos a nuestros estudiantes que escribieran un resumen tras ver la película "La terminal". Decidimos investigar los errores que cometían al usar las preposiciones at, in y on. El corpus estaba formado por 50 redacciones de los estudiantes y fue analizado con WordSmith Tools 5 *tanto cuantitativa como cualitativamente, es decir, identificamos las preposiciones analizando sus frecuencias y concordancias para buscar combinaciones no nativas. Nuestro objetivo final fue crear una serie de ejercicios tomando como base para ellos las incidencias de las preposiciones. Como conclusión general, pensamos que el análisis de corpus puede resultar una herramienta efectiva para la creación de las actividades hechas a medida y para materiales didácticos.*

Palabras clave: Turismo, análisis de corpus, WordSmith Tools*, materiales didácticos,* Exelearning

1. Introduction

This article springs from the observation that there is a salient amount of research on the detection of preposition errors of non-native speakers of English due to the fact that prepositions represent a considerable proportion of all grammatical errors by English as Second Language learners. Many studies have been carried out in the field of corpus linguistics; however there is hardly any study done using corpus linguistics based on error analysis approach and especially little research is to be found on prepositions (Eeg-Olofsson et al.,2003; Izumi et al.,2004; Lee & Seneff, 2006; Chrodorow & Tetreault, 2008). Hence, it is our intention to fill this gap.

Thus, the purpose of this paper is twofold. Firstly it aims at discovering the main mistakes made by Spanish students when using English prepositions in their essays. Secondly, we aim at producing teaching materials that, in our view, could be remedial to the spotted weaknesses.

Therefore, our students were required to watch the film "The terminal" by Spielberg (2004) as a compulsory task which supplemented two of the units in their student's book of English language (Walker & Harding, 2006): "Airport departures" and "The airline industry". After teaching our students the film review conventions as a genre they were asked to write a film review of about 200 words. We decided to investigate the errors made when using prepositions. Our corpus was composed of 50 student's essays, which we analyzed using *WordSmith Tools* 5 (Scott, 2010) both quantitatively and qualitatively. That is, we retrieved the prepositions analyzing their frequencies and concordances in order to look for non-native combinations.

However, in our view, the amount of corpus analysed seemed not to be sufficient to achieve our aim. Consequently, we took the decision of creating a larger corpus by adding written material extracted from a second task set during the course. This task was a storytelling script on a promotional campaign related to tourism.

Our final aim was to create a series of exercises by using the occurrences of the prepositions as the main basis of our filling the gaps and multiple choice exercises. In this way, our students were provided with exercises based on their own errors when using prepositions in writing. These exercises created with the *Exelearning* programme for the design of learning objects, were uploaded in the form of online activities for future students. As a general conclusion, we believe that corpus analysis could be an effective tool in order to create tailor-made activities and teaching materials. This research has been carried out within the frame of the Tur-i-Tic research team (Anglotic) from the University of Valencia.

This article is structured as follows: in the next section, we describe corpus linguistics analysis, namely its use in the field of Second Language Acquisition. Then we overview the difficulty in learning English preposition usage; in Section 3 we explain our methodology.

In section 4 we show our results and eventually in section 5 we draw some conclusions based on our experience.

2. Corpus linguistiscs

Computing tools specially designed for the analysis of a corpus have enabled the study of recurrent patterns and grammatical structures. Thus, the time and effort spent in the elaboration of a corpus is worth for the fact that this corpus can be used as database material in order to analyze the language used by a certain group of students and find out what the most typical mistakes are.

The main requirements to be taken into account for the elaboration of a corpus according to different experts in corpus linguistics (cf. Ricart, 2008) can be summarized as follows: (i) representativeness, Biber (1998:246) states that a corpus is not merely a gathering of texts but a selection of them according to specific criteria and to a specific aim. (ii) Diversity: "a well designed corpus must represent the different registers of the language", that is "register variation" as Biber (1998:249) explains. (iii) Subject matter is another aspect to be considered for a corpus because frequency of words clearly varies depending on the subject matter. (iv)Data in a corpus must be authentic, a considerable amount systematically organized and presented in electronic format.

In terms of the size of the corpus, Sinclair (1991) supports that a corpus should be as large as possible because this enables the linguist to reach a quality of evidence not previously attainable. For him quality in the case of corpus linguistics goes hand in hand with quantity. Looking at a lot of data allows the linguist to come to conclusions on frequency.

Leech (1991:10) affirms that a collection of machine-readable text does not make a corpus. Therefore, he claims that the size of a corpus is not so relevant. The sample has to be representative according to our purpose of analysis. Swales (2006:20) advocates for specialized corpora and affirms that "bigger may not always be better, and size may not win all". Lee (2001:37) emphasizes that a small specialized corpus is preferable because it is more homogeneous and more suitable for genre-based studies since "it can take into account interactional, pragmatic and contextual features in addition to purely linguistic ones".

Regarding Computer tools for corpus analysis, there is a wide range of these types of tools; we will focus on concordance programs. Concordance programs allow us to look for specific target words in a corpus, providing us with a complete list of the occurrences of one word in particular in a certain context. Some of these programs are available at small or no cost, like *Word Smith Tools* 5 which, as their author Mike Scott (2001:47) explains, are composed by different tools for different tasks. The *Concord* will help us to study in depth any word centred within its concordance lines (composed of a variable amount of context at either side) (Baker, 2006: 125; Baker et al, 2008: 278).

However, it should not be forgotten that the computer program is used as a simple sorting and counting data tool and it is the linguist who performs serious data analysis.

2.1. Corpus linguistics for teaching

According to Fuster and Clavel (2011: n 3 quoting to Barlow (1996) teachers can use a corpus in different ways. One possibility is analysing the corpora themselves for material

design but they can decide to introduce them in the classroom in order to with different purposes:

i. determine frequency patterns in specific domains;

ii. enrich language knowledge;

iii. produce 'authentic data;

iv. generate teaching materials.

In our study, we took the first option, i.e., we analysed the corpora ourselves to create some exercises, but these teaching materials were not thought to be exploited in class but to be used on-line.

The Data driven learning (DDL) methodology (Johhns, 1991), supported by Biber (1998: 170-197) turns students into detectives of the language making learning more attractive. They learn from authentic material finding out the rules of the language by themselves. DDL allows students detect language patterns by studying corpus data displayed in concordance lines. If a student does not know the preposition which follows a certain verb or adjective he can find it out by using the information in the corpus. The advantage for the teacher is therefore, that they can teach those aspects of language that seem problematic for a certain group of students. In this way, the teacher can select the difficult prepositions, for instance, and edit concordance lines to show the students how to use them. Other typical uses of line concordances occur with frequent items or important items in a certain subject area.

One advantage of concordance lines is that they can facilitate the hard task of searching for errors regarding a particular grammatical item or regarding digitalized tasks handed in by students. In this way, we could create for instance an exercise to lead learners to deduce what the prepositions collocate with and to predict other cases with the same pattern, that is, other collocations of a certain preposition with the same kind of word. For example, if the preposition *in* appears with summer, the student might deduce that all seasons need the same preposition. Students with a higher command of the language could also pay attention to longer prepositional phrases to extract a pattern such as "be angry with someone for something".

3. Methodology

The sample of our study was made of 50 students at the University of Valencia. They were doing their first year of the new degree of Tourism from October to December 2010. Students were required to watch the film "The terminal" by Spielberg (2004) and write a film review. Additionally, they had to hand in a storytelling script, as explained before. Both tasks had to contain a determined structure and number of words following the teacher's instructions.

Initially, we focused our analysis only on the film review, however, in our view, the amount of corpus analysed seemed not to be sufficient to achieve our aim. Consequently,

we took the decision of creating a larger corpus by adding written material extracted from the storytelling script.

The first task was carried out individually whereas the second one was performed in groups of three or four. The length of the writings was of 200 words for the film review and of 300 for the digital storytelling script.

In this way, we collected two small corpora composed of 65 student's essays, which we analysed using *WordSmith Tools* 5 (Scott, 2010) both quantitatively and qualitatively. That is, we retrieved the prepositions analyzing their frequencies and concordances in order to establish the main problems students are faced to when writing compositions.

The underlying goal for the present study is that students are prone to make errors while writing. This is so, in our opinion, because they do not realize the errors that they make. Therefore, we believe that making them aware of these mistakes will allow them to correct them in the future.

The process of our analysis was as follows: firstly, once both corpora had been created and converted into text without format, we used the *Concordance* so as to study the occurrences whose search words were the prepositions *in*, *at* and *on*. Then, we searched for errors in the use of the above mentioned occurrences with regard to prepositions. Our next step was to collect all those sentences containing a misused preposition and blank the preposition out. By doing so, we created two kinds of self-correcting activities: on the one hand a filling the gap activity and on the other hand a multiple choice exercise. Finally, by means of the program *Exelearning,* we converted these activities into available on-line activities which we uploaded to the Anglotic web page, being the later the research group in the department of English and German Studies at the University of Valencia.

4. Results

Concordance provided us with 130 the occurrences being *in* the search word, 28 occurrences of *on* and 44 of *at*, which made a total of 158 concordance lines. The search of the prepositions was carried out separately. Some of the examples of each are displayed in tables 1, 2 and 3 below.

Table.1. Example of concordance lines with the preposition on

N	Concordance
1	, that was it!.Moreover, the smell... Strange things kept on happening-It could be the fatigue, the exhaustion of five
2	the company of real pirates - The trip includes two nights on the boat and one on the island, with plenty of shows where
3	pirates - The trip includes two nights on the boat and one on the island, with plenty of shows where you can interact
4	the NewYear's Eve like you never have done it before, get on the biggest plane in the world, Airbus A380. Howto
5	the Singapore culture. And the 3rd party will be celebrated on the 3rd floor with Indian decoration . * Timetable or
6	I took my last look in the mirror and found a strange mark on my forehead above my nose, between my eyes.I cleaned it
7	a true pirate party - In the morning you have time to relax on our beaches. You'll also appreciate the fauna and flora of
8	a cruise where we will have dinner and we will be in a party on the cruise and we will sleep on the cruise too, in the
9	and we will be in a party on the cruise and we will sleep on the cruise too, in the cabins. The saturday, we will have a
10	in the cabins. The saturday, we will have a relaxing morning on the beach where who needs gather strength can rest and

Table.2. Example of concordance lines with the preposition at

N	Concordance
1	the features that make it different from the other cities. At first, we introduce the topic describing the city (where is
2	the Australian way. An hour later the Singapore party starts at the second floor and because of the hour change and the
3	una foto del brochure) Curiosity made me look at the packaging in which it had come and this was
4	travel to a destination that was unknown to me, but familiar at the same time. My first vacation, my first break, and my
5	travel to a destination that was unknown to me, but familiar at the same time. Now, I identified all of those signals I
6	New Year's Eve, a great sensation and a new experience. At 24:00 we welcome the New Year as the Australian way.
7	to be at the airport at 20:00h to check in and board the plane at Sidney Airport, after the take off the dinner will be at the
8	Indian's party India Passengers have to be at the airport at 20 00h to check in and board the plane at Sidney Airport,
9	plane at Sidney Airport, after the take off the dinner will be at the first floor at 22:00h. At 23:30h the open bar starts and
10	after the take off the dinner will be at the first floor at 22:00h. At 23:30h the open bar starts and the pool room for the

Table.3. Example of concordance lines with the preposition in

N	Concordance
1	to celebrate three times New Year's Eve in one night? Flying in the opposite direction of rotation of the earth we achieve it.
2	Is the largest passenger airliner in the World. It s 73m long and 21,4 m high. Is one of the
3	Place Action Airspace 20:00 Sidney Airport Check in Australia 21:00 Sidney Airport Take off 22:00 1º floor Airbus
4	from www.beemp3.com Characters: Background Script: Get in to our time machine and celebrate three times the end of
5	like you never have done it before, get on the biggest plane in the world. Airbus A380. How to celebrate three times New
6	Airbus A380. How to celebrate three times New Year's Eve in one night? Flying in the opposite direction of rotation of the
7	Passengers have to be at the airport at 20:00h to check in and board the plane at Sidney Airport, after the take off the
8	Also, we try to point the most important characteristics out in order to convince the client to visit it. After that, we
9	for give more information to the people who are interested in the trip. STORYBOARD
10	already written my destiny. I studied hard, I graduated in Law and Economics and got to be successful. Yes, yes, a

As mentioned before, we collected those sentences containing a misused preposition and blanked the preposition out giving rise to two kinds of self-correcting activities: on the one hand, a filling the gap activity and on the other hand a multiple choice exercise. To illustrate this with an example (see table 4):

Table.4

- An hour later the Singapore party starts ~~at~~ the second floor
- An hour later the Singapore party starts_____ the second floor

- In the morning you have time to relax ~~on~~ our beaches.
- In the morning you have time to relax _____ our beaches.

- ~~At~~ the evening we will [have] fun (us) with a (season)[session] of monologues.
- _____ the evening we will have fun with a session of monologues.
- in b. on c. at

Finally, by means of the program Exelearning, we converted these activities into available on-line activities which we upload to the Anglotic web page[206] *(see tables 5 and 6).*

Table. 5. Activity 1 in the Anglotic page.

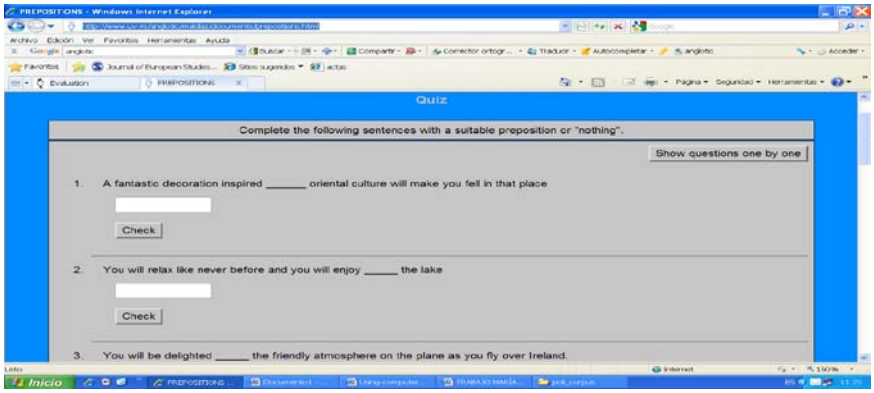
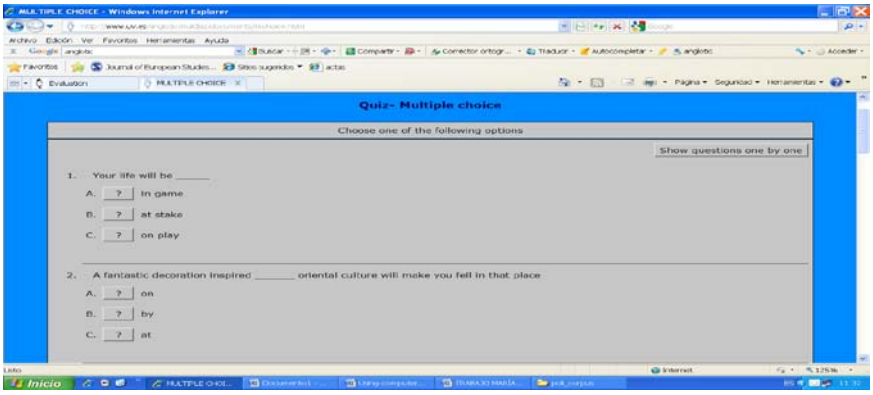


Table. 6. Activity 2 in the Anglotic page.



5. Conclusion

The present research work shows that students in their first year of the new degree of Tourism need practice in the use of the prepositions *in*, *on*, *at*, as they still make a significant number of mistakes when using them. It is evidenced that the creation of a corpus of written material allows teachers to recognize difficult aspects of language for their students. At the same time it enhances the importance of TICs nowadays, as it enables the creation of on-line exercises that will help students recognize and practice prepositions. A further line of research could focus on other linguistic problems faced by students when learning a foreign language such as student's problems with false friends, collocations, adjective order, etc. It would also be interesting to carry out this analysis in other fields of ESP or even in other disciplines.

206 http://www.uv.es/anglotic/maldiaz/documents/prepositions.html

6. References

Baker, Paul (2006) *Using Corpora in Discourse Analysis.* London: Continuum International Publishing

Baker, Paul; Gabrielatos, Costas; KhosraviNik, Majid; Krzyżanowski, Michał; McEnery, Tony and Wodak, Ruth (2008) "A useful methodological synergy? Combining critical Discourse Analysis and Corpus Linguistics to examine discourses of refugees and asylum seekers in the UK press. Discourse & Society May 2008 19:273-306

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English.* Harlow: Pearson Education.

Biber, D., Conrad, S. & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on esl student writing. *Journal of Second Language Writing*.

Chrodorow, M., & Tetreault, J. (2008). The Ups and Downs of Preposition Error Detection in ESL Writing, COLING '08 *Proceedings of the 22nd International*

Conference on Computational Linguistics (Volume 1) (pp. 865-872).

Dalgish, G. (1985). Computer-assisted ESL research and courseware development. *Computers and Composition*.

Eeg-Olofsson, J. & Knuttson, O. (2003). Automatic grammar checking for second language learners – the use of prepositions. Nodalida.

Fuster, M. & Clavel, B. (2010). Corpus Linguistics and its Applications in Higher Education. Revista Alicantina de Estudios Ingleses, 23, 51-67

Fuster-Márquez, M. & Clavel-Arroitia, B. (2010). Second Language Vocabulary Acquisition and its Pedagogical Implications. In Pérez Ruiz, L., Parrado Román, I. & Tabarés Pérez, P. (eds). Estudios de Metodología de la Lengua Inglesa (V). Valladolid: Secretariado de Publicaciones Universidad de Valladolid, (pp. 205-212)

National Center for Educational Statistics (2002). Public school student counts, staff, and graduate counts by state: School year 2000-2001.

Izumi, E., Uchimoto, K. & Isahara, H. (2004). The overview of the SST speech corpus of Japanese learner's English and evaluation through the experiment on automatic detection of learners' errors. LREC.

Izumi, E., Uchimoto, K., Saiga, T., Supnithi, T. & Isahara, H. (2003). Automatic error detection in the Japanese leaners' English spoken data. ACL.

Johns, T. (1991). Should you be persuaded – two samples of data-driven learning materials. In T. Johns & P. King, eds., (1991) Classroom Concordancing. Birmingham University: English Language Research Journal, 4: 1-16.

LEE, D. (2001). Genres, registers, text-types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle. Language Learning and Technology, 5, 37-72.

LEE, J. & SENEFF, S. (2006). Automatic grammar correction for second-language learners. Interspeech.

LEECH, G. (1991). The State of the Art in Corpus Linguistics. In Aijmer & Altenberg (eds), 8-29.

LEVIN, B. (1993). English verb classes and alternations: a preliminary investigation. Univ. of Chicago Press.

MURATA, M. & ISHARA, H. (2004). Three English learner assistance systems using automatic paraphrasing techniques. PACLIC 18.

RATNAPARKHI, A. (1998). Maximum Entropy Models fornatural language ambiguity resolution. Ph.D. thesis, University of Pennsylvania.

RATNAPARKHI, A., LEE, J. & SENEFF, S. (2006). Automatic grammar correction for second language learners. Interspeech.

RICART, A. (2008) An ESP Compartive Analysis in Medical Research Articles: Spanish – English. ProQuest Dissertations & Theses.

SINCLAIR, J. (1991). Corpus concordance collocation. Oxford University Press.

SCOTT, M. (2010). Word Smith Tools 5, Oxford University Press.

SWALES, J. (2006). Corpus Linguistics and English for Academic Purposes in Information Technology in Languages for Specific pusposes, pp.19-33 E. Arnó Macià, A. Soler Cervera and C. Ruaeda Ramos: Springer.

WALKER, R. & HARDING, K. (2006). Oxford English for Careers Tourism 1 Student's book. Oxford University Press.

Exploiting corpus evidence for automatic sense induction

Rema Rossini Favretti

Fabio Tamburini

Andrea Zaninello
Department of Linguistics and Oriental Studies

University of Bologna

Abstract

In this paper, we discuss the application of a sense induction procedure to data from CORIS, a well-balanced reference corpus of Italian. The method considered discriminates between the different senses of a word by analysing the relationships between its collocates and suggesting collocate clusters, each of which corresponds to one sense of a word. The collocate clusters are represented as 3D-graphs in a semantic space. We show that for some examples the method can satisfactorily induce the senses of the chosen node; however, we also show that for some controversial instances human interpretation of the results is needed. We thus conclude that, although powerful, automated systems still require human knowledge both for the analysis and the interpretation of language phenomena, and that an integration of the two methodologies is desirable.

Keywords: sense induction, corpora, CORIS, Italian, polysemy, 3D-graph

1. Introduction

The aim of this paper is to explore how statistical analysis of corpus evidence can contribute to sense disambiguation in non-annotated text. We focus on collocations as a source of surface evidence automatically extracted from corpora through positional and association-based procedures following probabilistic criteria.

Our basic assumption is in line with the Firthian tradition and the classical Harrisian distributional hypothesis, which assumes that 'similar' linguistic items (in particular, *semantically* similar items) will have similar distributions in naturally occurring texts.

More specifically, we hold that most characteristic collocates of a (potentially polysemic) word are a good indicator of its meaning(s) and therefore distributional 'closeness' between them can be seen as a hint of semantic similarity. Significant co-occurrence frequencies can be used to discriminate between the different senses of a word by grouping its collocates according to their distributional behaviour analysed through statistical association measures.

Our paper is organized as follows: firstly, we sketch the methodological background underlining our research and present a brief description of CORIS, the corpus of written Italian used in our study. Secondly, we describe the analysis tools and procedures exploited in our research. Thirdly, we present some case studies focusing on polysemic words in Italian. Finally, we present and discuss our results and conclude with some remarks and perspective work.

2. Background and data

2.1 Background

Our research is based on the traditional *contextual meaning theory* as exemplified in early works by Firth (1957) and further developed by his pupils (see, for example, Sinclair, 1991). In this framework, collocations are defined as *co-occurrences of words within a determined unit of information* (from next neighbours to whole texts) *where linguistic items appear together significantly,* in other words, with greater probability than one would expect if the relationship between them were completely random. In this framework we use the term "random" in its statistical sense, as we agree that in real language randomness is a nonsensical concept.

The meaning of a word is defined *contextually,* i.e. by the relations between the word itself and the other linguistic items that represent its context. This view dates back to the early Saussurian approach, where the function of an item is only defined *differentially* and the value ('*valeur*') of any sign only exists by virtue of the (differential) relationships that it holds with other items in the language.

This view has been variously applied in recent works and in particular the analysis of distributional similarity has been widely exploited in Word Space Models to measure the

semantic or functional similarity between different words (e.g. Lenci, 2008). As Sahlgren (2008) points out, there exist two approaches to a distributional study of meaning: one consists in considering, as distributional evidence, the words that surround the target word, another is based on outlining distributional profiles based on the text regions where a word appears. These approaches, albeit often considered as equivalent, in fact rely on different types of distributional data exploiting, on the one hand, *paradigmatic relations,* and *syntagmatic relations,* on the other hand.

The methodology that we applied here is based on approaches of the first kind; however, differently form these approaches, we will exploit semantic similarities between the collocates of a node as evidence for discovering clusters of collocates, each of which should correspond to as many different senses of the node. As a computational method for our study, we follow the work by Heyer (2001; 2002) for the construction of co-occurrence graphs which exploits the visual representation of a word's collocates to induce its different senses.

2.2 The Corpus

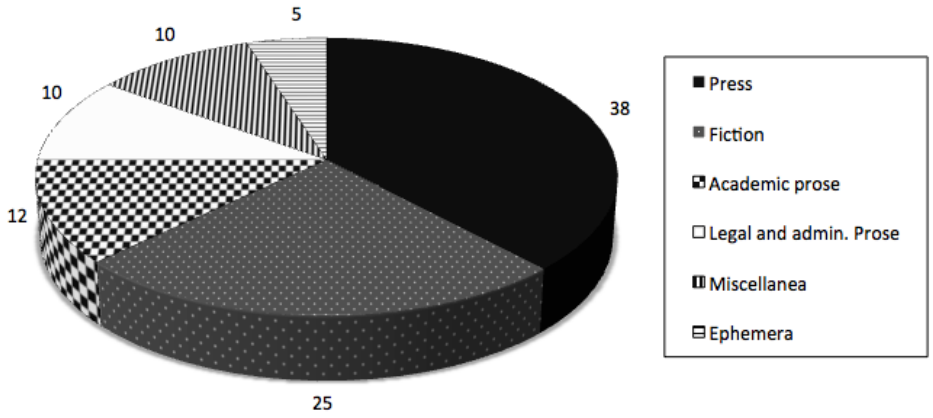
The corpus used in our study is the 120-million word corpus CORIS - Corpus di Italiano Scritto - an electronically-based reference corpus of contemporary written Italian. The corpus is the result of a research carried out at the University of Bologna since 1998 and it was designed, developed and implemented according to linguistically motivated criteria aimed at assuring that the corpus is a representative and well-balanced sample of standard Italian (Rossini Favretti, Tamburini & De Santis, 2002).

The corpus contains written texts from the 1980s to the present and is updated through a monitor corpus every three years. It contains texts from a wide range of varieties of Italian, chosen by virtue of their representativeness. In particular, the selection and proportion of texts was based on external, internal, as well as comparability criteria and led to the following structure (see Table 1) and proportions between different subcorpora (see Table 2).

Table 1. Structure of the CORIS corpus.

Subcorpus	**Sections**	**Text types (examples)**
PRESS	newspapers, periodic, supplement	national, local specialist, non-specialist connotated, non-connotated etc.
FICTION	novels, short stories	Italian, foreign for adults, for children crime, adventure, science fiction, women's literature etc.
ACADEMIC PROSE	books, reviews	human sciences, natural sciences, experimental sciences, popular history, philosophy, arts, literary criticism, economy, biology, etc.
LEGAL AND ADMINISTRATIVE PROSE	books, reviews, documents	legal, bureaucratic, administrative etc.
MISCELLANEA	books, reviews, documents	books on religion, travel, cookery, hobbies, etc.
EPHEMERA	letters, leaflets, instructions	private, public printed form, electronic form etc.

Table 2. Proportion of the subcorpora in CORIS (numbers represent percentages).



3. Procedure

The procedure exploited in this study and applied to CORIS was originally presented in Heyer *et al.* (2001) and was formerly used to modulate register variation (Rossini & Tamburini, 2009). The procedure creates collocation sets for the selected node through an automatic, iterated process of collocation analysis based on association measures

and recursively applied to the collocates. The statistical measure used for collocation extraction is the log-likelihood ratio and the context window considered was the entire sentence.

The results are represented as co-occurrence graphs. This representation allows one to single out clusters of collocates connected at different strengths, and thus define different meaning areas providing a visualisation of polysemy through a representation of the collocates' distribution in a semantic space.

The visualization in the 3D-graph structure highlights areas where stronger relations between collocates are grouped together: homogeneity in the graph is shown when the collocates are interconnected (distributionally and therefore semantically), while separation in the representation is displayed when collocates are not connected.

In the next section we present some relevant examples of application of this procedure to some nodes, leading to different results, which will be eventually discussed.

4. Case studies and results

In this section we present the application of the procedure outlined in the previous section onto two case studies, and we present different results for the nodes considered.

4.1 Case study 1: Risoluzione

We applied the above procedure to a highly polysemous word, '*risoluzione*', which can be variously translated into English by 'resolution', 'decision', 'cancellation' etc., according to its different senses. Below we present a snapshot of the concordances for the node, highlighting its possible uses (see Table 3):

Table 3. A snapshot of the concordances for the node risoluzione.

STAMPAQuot: : " Dal punto di vista politico la **risoluzione del consiglio di sicurezza** dell ' *STAMPAPeri*: itiche e strategiche , la bozza di **risoluzione presentata al Consiglio** di sicure *NARRATRoma*: allo studio dei particolari e alla **risoluzione dei problemi logistici** , politici *PRGAMMDocu*: ti) , salvo che il diritto alla risoluzione sia stato espressamente stipulato *PRGAMMDocu*: lità per l ' utente di ottenere la **risoluzione del contratto** entro un termine ra *PRGAMMDocu*: , con il sostegno dell ' UE , una **risoluzione d ' urgenza sul lavoro** forzati in *PRGAMMDocu*: ne rapporto , spettante in caso di **risoluzione del rapporto di lavoro ,** è discip PRGAMMDocu: gli estremi della giusta causa di risoluzione , che davano diritto di ottenere *PRGAMMVolu*: o agli aeroporti dell ' area . 106 Risoluzione 753 (18 maggio 1992) Raccomanda *MISCRivist*: CCD , quindi è evidente che questa **risoluzione viene raggiunta** per interpolazion *MISCRivist:* asferimento di **immagini con alta risoluzione** e alta dinamica , quali radiograf EPHEMIstru: li le **dimensioni di pagina , la risoluzione** e così via , e non al suo contenu

The co-occurrence graph resulting from the application of the procedure to the node is displayed below (see Figure 1). The graph displays four distinct homogeneity areas that can be identified as clusters corresponding to as many senses of the word. In particular:

1. The first cluster (top-left) makes reference to *risoluzione* as the level of detail of an image (as in *alta risoluzione* = 'high resolution');
2. The second cluster (bottom-right) refers to the meaning of "solving" - e.g. of problems, cases, etc. (as in *risoluzione di problemi* = problem solving) ;
3. The third cluster (centre-right) makes reference to *risoluzione* as "decision" - e.g. of a formal body (as in *risoluzioni del Consiglio di Sicurezza* = Security Council Resolutions);
4. The fourth cluster (bottom-left) refers to *risoluzione* as "cancellation" – e.g. of a contract (as in *risoluzione di un contratto* = rescission of a contract).

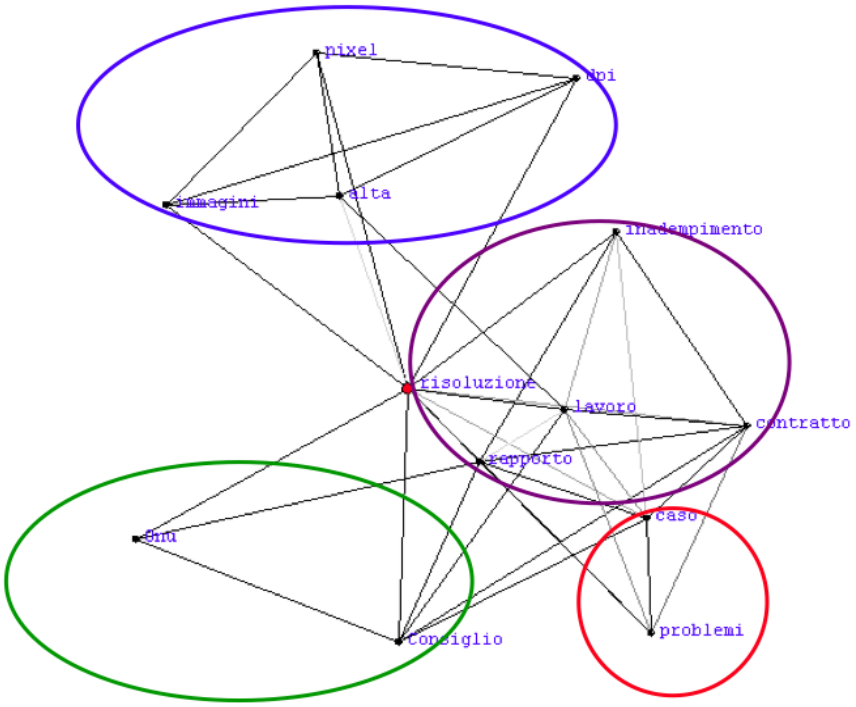


Figure 1. Co-occurrence graph for the node risoluzione.

As we can see in this example, the graph visualisation is able to separate between the different senses of the word and is fairly exhaustive, having identified all the possible uses that were highlighted by the analysis of the concordances as in Table 3.

4.2 Case study 2: Calcio

As a second example, we analysed the collocates of the node "*calcio*" which, according to the analysis of the concordances, are organised around three main axes, corresponding to as many senses of the word (cf. Table 4).

Table 4. A snapshot of the concordances for the node calcio.

STAMPAQuot: straordinaria come per i **mondiali di calcio** ' 90 " . I comunisti aggiungono di

STAMPAQuot: questa annata . [BEGINDOC] 09/06/1997 CALCIO FLASH PLAYOFF E PLAYOUT DI C . Play STAMPAQuot: gli stadi , cedendoli alle **società di calcio** ; ma nello stesso tempo i club chie STAMPAQuot: nuove mosse . Ma se si preferisce il calcio , con Fifa 2000 ci si può cimentare STAMPAPeri: Qualche nome : Ferrarelle (**368 mg di calcio per litro**) , Sangemini (327,8) , STAMPASupp: lizzano l ' automotivatore squadre di calcio come il Milan , campioni come Alber NARRATTrRo: vantarsi di aver atterrato con un **calcio sulle zampe** un giovane rinoceronte NARRATTrRa: superiori perché sostengono che il calcio consiste in ventidue imbecilli che PRACCVolum: transitorio innalzamento **del tasso di calcio** , la tubulina si libera anche dalla MISCVolumi: (400 mg %) ; fosforo **(800 mg %) ; calcio (700 mg %)** . Non è presente lo io

The node presents three main uses: *football* (as in *mondiali di calcio* = world champions), *kick* (as in *calcio sulle zampe* = kick on the paws), *calcium* (as in *tasso di calico* = level of calcium).

However, when we applied the co-occurrence graph visualisation procedure to the node (see Figure 2), different results were returned. In particular, the meaning clusters induced by the procedure are organised around three main clusters:

1. the first cluster (top-left) corresponds to the '*football'* meaning, as in *squadra di calcio* = 'football team'.
2. the second cluster (centre-right and across) makes reference to *calcio* as the chemical element 'calcium', as in *carbonato di calcio* = 'calcium carbonate'
3. the third cluster (bottom-right) corresponds to a very specific use of the node as it refers to a 'cranberry collocation' making reference to the title of a popular TV show about football *Quelli che il calcio*... (lit. 'those who football...').

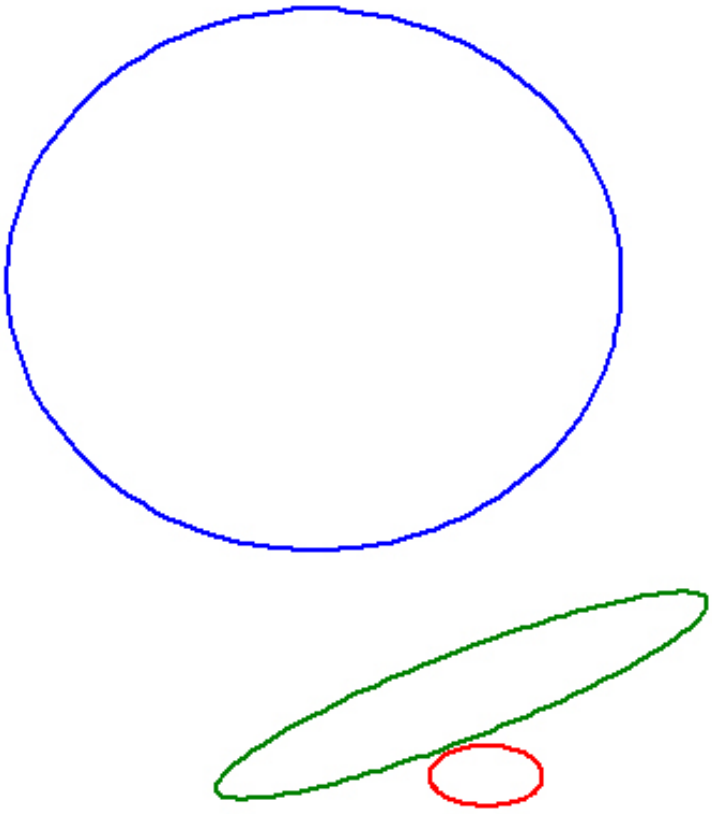


Figure 2. Co-occurrence graph for the node *calcio*.

5. Discussion, conclusions and future work

As we outlined above, the 3D-graph in the first case study returned all four senses of the word previously observed in the analysis of concordances and can thus be said to provide an accurate representation of the polysemy of the node. The second case study, on the other hand, showed a mismatch between the analysis of concordances and the 3D-graph. In the first case, three different senses had been found, including the 'basic' sense ('kick') of *calcio*. However, the 'kick' sense is surprisingly absent in the graph, a result that depends on a too low representation of this use within the computational methodology applied. In return, another 'meaning' nucleus is highlighted in the graph

and corresponds to a very specific use of the word *calcio* (the title of a TV show about football) that is represented as a separate cluster form the football one (although weakly connected to it) because it appears in a 'cranberry collocation', i.e. a very idiosyncratic structure which does not conform to the normal rules of syntactic composition in Italian.

It could be said that it is controversial whether this use actually represents another sense of the word *calcio,* (since, taken alone, it has got the same reference to football as the other cluster), however, if we consider the whole context where the term appears, we believe that it is correct to separate it from the 'football' cluster, since the word is used in a very idiosyncratic way and represents one sense on its own (which we might label the 'TV' sense). Therefore, we showed that the methodology can in many cases correctly identify the main senses of a node through the observation of the resulting graphs.

In addition, we conclude our work with some final remarks and observations on the results obtained.

i) Firstly, as exemplified by the second case study (in particular by the mismatch between the senses found in the concordances and those induced in the graph), we would like to point out that, since our methodology is based on the observation of naturally occurring data, results of the clustering procedure depend on the distribution of the data because computational methods are sensible to the frequency of the phenomena under investigation. The absence of a phenomenon does not necessarily ensure that the latter does not exist in the language, but we believe that negative evidence is a very important clue about the significance of language facts, based on and validated by statistical significant measures applied on naturally occurring texts (and thus not induced through deductive reasoning).

ii) Secondly, the case study on the node *calcio* showed that the system creates different clusters, and therefore indicates two separate meanings for the 'TV' and the 'football' sense. From an extra-linguistic point of view, of course, the senses in both examples are semantically connected, as they refer to the same *referential* meaning of the word. However, we believe that it is desirable to keep them separate, as explained before, because we are mainly interested in the different *uses* of one node and thus it is very useful to be able to distinguish between a very common, unmarked instance of the word (as in the football sense) as opposed to its idiosyncratic use.

iii) Thirdly, it is worth pointing out that the sense induction procedure does not separate between different kinds of meaning relations and identifies uniquely general macro-relations of *semantic similarity/dissimilarity.* In particular, we showed that, on the one hand, the methodology treats in a unified manner semantic compatibility relations, such as polysemy (as in the case of *risoluzione*), homonymy (as in the case of *calcio*), synonymy etc., as it does, on the other hand, for semantic incompatibility relations such as antonymy, complementarity etc. As a matter of example, the kind of semantic relation between the sense clusters of *calcio* would probably be interpreted by linguists as an instance of homonymy rather that polysemy (especially for the 'football' and the 'chemistry' senses), as the word happens to have the same form in the two senses but its meanings are not connected in any way, as opposed to the meanings of *resolution.*

As a suggestion for future work, we believe that the system could be positively implemented by the integration of linguistic as well as extra-linguistic information. This could be done exploiting human knowledge as well as making use of language resources such as electronic dictionaries, ontologies etc. This would allow one to *label the senses* as well as to overcome some of the problems discussed above (such as a separation between true polysemy and usage idiosyncrasies). Moreover, we believe that this procedure may be expanded to the historical dimension in order to study the evolution of a word's senses across time. We plan to do so by applying the procedure to the diachronic corpus DiaCORIS (cf. Onelli *et al.*, 2006), a representative and balanced collection of Italian written language ranging from the National Unification of Italy - 1861 - to the end of the Second World War - 1945.

6 References

Firth, J. R. (1957). Papers in Linguistics 1934-1951. OUP. Oxford.

Heyer, G., Lauter, M., Quasthoff, U., Wittig, T., Wolff, C. (2001). Learning relations using collocations, Proceedings of the IJCAI Workshop on Ontology Learning, Seattle, USA.

Heyer, G., Quasthoff, U., & Wolff, C. (2002). Automatic analysis of large text corpora - A contribution to structuring WEB communities. Lecture Notes in Computer Science, 2346, 15-26.

Lenci, A. (ed.) (2008). From context to meaning: distributional models of the lexicon in linguistics and cognitive science, Italian Journal of Linguistics, 20/1.

Onelli, C., Proietti, D., Seidenari, C., Tamburini, F. (2006). The DiaCORIS Project: a diachronic corpus of written Italian. Proceedings of the 5th International Conference on Language Resources and Evaluation - LREC 2006. Genoa: 1212-1215.

Rossini Favretti, R., Tamburini, F. (2009). Exploring register variation through corpus evidence, in Abstracts of DGfS 2009 Workshop on Corpus, Colligation, Register Variation, Osnabruck, p. 155.

Rossini Favretti, R., Tamburini, F., De Sanctis, C. (2001). A corpus of written Italian: a defined and dynamic model. In Proceedings of Corpus Linguistics 2001 Conference, Lancaster,UK.

Sahlgren, M. (2008). The Distributional Hypothesis. From context to meaning: Distributional models of the lexicon in linguistics and cognitive science (Special issue of the Italian Journal of Linguistics), Rivista di Linguistica, volume 20, numero 1, 2008.

Sinclair. J. (1991). Corpus, Concordance, Collocation. Oxford University Press.