Nazar, Rogelio and Acosta, Nicolás (2023) A Lightweight Statistical Method for Terminology Extraction. Journal of Computer-Assisted Linguistic Research 7: 43-59. ISSN: 2530-9455.

A Lightweight Statistical Method for Terminology Extraction

Rogelio Nazar*[1], Nicolás Acosta[2]
[1] Pontificia Universidad Católica de Valparaíso, Chile
[2] Universidad Nacional de Cuyo, Argentina
*Corresponding author: rogelio.nazar@pucv.cl

Received: 29 September 2023 / Accepted: 10 November 2023 / Published: 18 December 2023

Abstract
We propose a method for the task of automatic terminology extraction in the context of a larger project devoted to the automation of part of the tasks involved in the production of terminological databases. Terminology extraction is the key to drafting the macrostructure of a terminological resource (i.e., the list of entries), to which information can be later added at the microstructural level with grammatical or semantic information. To this end, we developed a statistical method that is conceptually simple compared to modern neural network approaches. It is a lightweight method because it is based on term dispersion and co-occurrence statistics that can be computed with basic hardware. For the evaluation, we experimented with corpora of lexicography and linguistics in English and Spanish of ca. 66 million tokens. Results improve baselines in almost 20%.

Keywords: automatic terminology extraction, computational terminology, corpus-based terminology processing, information extraction, dispersion measures

1. Introduction

Ever since the first efforts in the sixties, different software products have tried to optimize terminology processing as tools for the storage, organization, and retrieval of terminological data (Hutchins 1998). The first decades were dominated by the work of Wüster (1979) and his followers (Felber 1984; Arntz and Picht 1995), who relied basically on the method of introspection and their own knowledge as specialists. This changed with the arrival of linguistic corpora in the nineties, when an empiricist turn took place and terminologists began to apply methods imported from computational linguistics (Sager 1990). At present, computer assisted terminology has grown into a large and diversified field. Different tools have been proposed for term management (Steurs, De Wachter and De Malsche 2015), and research has appeared in areas such as terminology extraction (Kageura and Umino 1996; Rigouts Terryn, Hoste and

DOI: HTTPS://DOI.ORG/10.4995/JCLR.2023.20427

Lefever 2022), bilingual terminology alignment (Simões and Almeida 2008; Filippova, Can and Corpas Pastor 2021) and definition extraction (Pearson 1998; Meyer 2001), among other research avenues.

The large body of publications now available reflects the development of the field in the last decades, and terminology professionals are now better equipped to engage in the creation of terminological databases. That said, there is still ample room for improvement. Firstly, there are no tools that can integrate solutions for the different requirements of a terminology project. Secondly, the software currently available for terminology management is time consuming for the user because programs demand repetitive tasks that in many cases could be automated. This would be convenient not only to improve productivity but also to protect users from unnecessary stress.

In this juncture, we devised a project for automating different tasks involved in the creation of terminological databases, including functions for obtaining a draft macrostructure (a lemma-list) and data for the microstructural level such as grammatical (morphosyntactic patterns, gender, formation process) as well as semantic information (hypernyms, equivalents in another language, definitions and synonyms).

With this tool, terminology-related professionals will be able to generate raw material which they can later improve manually by adding or correcting data. Having some quality raw material would hopefully speed-up the process of generating a term database by editing and correcting rather than starting from scratch. The implementation is a freely available web-based prototype[1] that can perform the tasks of corpus processing (file upload, conversion to plain-text format, language detection, POS-tagging and indexing), terminology extraction (with optional human supervision), information extraction (definitions, equivalence in another language, hyperonymy, synonymy, etc.) and database management (editing, storage, retrieval and export~import options in HTML, CSV and TBX).

In this article, we focus on the description and quantitative evaluation of the results of terminology extraction, leaving for future work the evaluation of other functions. As a use-case, we describe experiments with specialized corpora of the domains of linguistics and lexicography in English and Spanish with a total extension of ca. 66 million tokens.

The article follows a canonical structure: Section 2 offers a brief overview of computational terminology techniques with emphasis in terminology extraction. Section 3 presents a description of the proposed method and its implementation as a prototype. Section 4 shows a preliminary evaluation of the method's performance and Section 5 recapitulates the main points and the challenges ahead.

2. Related work

2.1. Terminological Theory and Practice

As already mentioned, most of 20th century terminology was dominated by the work of Eugen

[1] http://www.termout.org

Wüster and his followers (Wüster 1979; Felber 1984; Arntz and Picht 1995). This group of pioneers became known as the Vienna School and consolidated their General Theory of Terminology (GTT), which established the terminology principles in theory and practice. Wüster, in particular, had a leading role in coordinating international efforts to publish standards, dictionaries, and norms, as well as initiating the computational treatment of terminology (Humbley 2022).

The GTT was, however, later challenged by some theoretical proposals that began to emerge around the 1990s (Sager 1990; Cabré 1999). Humbley (2022) summarizes Wüster's posthumous critical reception and reviews the main points of criticism. For example, many new trends advocated for a descriptive rather than a prescriptive point of view. The main reason, however, for a decline of the GTT seems to have been a change in methods. The arrival of corpus linguistics and the incorporation of its tools and methods into terminology practice led to a rapid turn to empiricism, and this naturally led to surprises and the admission of phenomena that were not supposed to exist in terminology, such as polysemy and term variation, among others.

2.2. Terminology Extraction

Aside from database software employed in terminology management (Steurs, De Wachter and De Malsche 2015), technical advances like automatic terminology extraction (ATE) revolutionized the acquisition of terminology from corpora. ATE is a categorization problem in which, for any input unit (i.e. a term candidate), a system must produce an output tag, ideally a term/non-term distinction, similar to the spam/no spam tag in the case of emails. Instead of discrete values, however, researchers in the ATE tradition have opted for a score to sort term candidates (Kageura and Umino 1996; Heylen and De Hertog 2015).

ATE has its roots in previous work in information retrieval (Spärck Jones, 1972), but it was in the 1990s when it gained widespread attention. Some proposals focused on how to detect multi-word terms, sometimes using statistical association measures (Daille 1994; Frantzi, Ananiadou and Mima 2000) or syntactic rules (Justeson and Katz 1995; Bourigault, Gonzalez-Mullier and Gros 1996). Another trend in terminology extraction has been the computation of statistics using reference corpora, a trend associated with the exploitation of keywordness or weirdness, i.e., the comparison of the frequency of a term in a specialized corpus with its frequency in a reference (general language) corpus (Ahmad, Gillam and Tostevin 1999; Anthony 2005; Drouin 2003; Baisa, Michelfeit and Matuška 2017; among others).

A more recent tendency is the application of machine learning techniques (Conrado, Pardo and Rezende 2013; Lang et. al. 2021; Rigouts Terryn, Hoste and Lefever 2022). Machine learning in terminology extraction is not new (Cabré, Estopà and Vivaldi 2001), but an increasing number of researchers are leaning in that direction. A recent survey on terminology extraction seems to confirm this (Tran et. al. 2023). In fact, one can expect to see many new neural network algorithms based on large language models applied to the automation of tasks not only to terminology but lexicography in general (de Schryver and Joffe 2023). This is, however, a tendency that shows ever-growing computational complexity. Neural network algorithms also tend to be unpredictable (Rigouts Terryn et al., 2020), and sometimes it is difficult to interpret

why some errors occur (OpenAI, 2023). This complexity also affects their implementation to offer practical solutions and the scalability for massive data processing taking into account energy consumption and hardware capabilities. In the present scenario, it still feels worthy to continue exploring alternatives, especially if they can be simple and scalable.

2.3. Terminology and Information Extraction

Another important note to add is that the field of terminology extraction has gradually been expanding to other related subfields, such as bilingual terminology alignment using parallel, comparable or unrelated corpora (Simões and Almeida 2008; Lefever, Macken and Hoste 2009; Aker, Paramita and Gaizauskas 2013; Haque, Penkale and Way 2018; Filippova, Can and Corpas Pastor 2021), or the application of text mining techniques to obtain information about the terms from the corpus, mainly definitions (Pearson 1998; Meyer 2001; Zhang and Jiang 2009); hypernyms (Hearst 1992; Bordea et. al. 2015; Shwartz, Santus and Schlechtweg 2017) and synonyms (Ville-Ometz, Royauté and Zasadzinski 2007; Cram and Daille 2016), among others.

3. Methodology and Implementation

3.1. Functions for Corpus Preprocessing

We embarked on developing a free web-based software intended to assist the terminologist in different tasks needed in the creation of term databases. Such routines are the compilation and preprocessing of specialized corpora, the extraction of terms from the text, the extraction of information about the terms and the editing and management of the extracted terminology database before it is exported in an industry-standard format.

We believe a terminology project should begin with the compilation of a specialized corpus. Typically, this will consist of publications of some field, covering a single topic or domain, having the same degree of specialization, some authority, and a minimum extension (Pavel and Nolet 2002; Steurs, De Wachter and De Malsche 2015). As a rough reference, this program will not produce quality results with corpora having less than 300 full natural texts.

Once the users have a corpus at their disposal, they are expected to upload it to the platform, preferably as a single ZIP file. We describe the operations that follow the uploading procedure, which users must apply in sequence and are necessary for the rest of the operations with the corpus.

- **Format detection and conversion**: The prototype accepts various formats such as ZIP, TXT, PDF, PS, DOC, DOCX, ODT, HTML, XML and applies the appropriate conversion routine to obtain UTF-8 UNIX TXT format documents.
- **Language detection**: Once with the corpus in TXT, the program applies an automatic language detection routine, carried out on the basis of text similarity measures with samples of texts of different languages. It detects the main language of the documents and also fragments of text inside that are in a different language.
- **POS-tagging**: With the documents already separated by language, the corpus is submitted to a POS-tagging procedure. For this operation we resorted to an external software,

UDPipe (Straka and Straková 2017).

- **Indexing**: The system makes extensive use of concordance extraction to decide if a word or sequence of words is a term and, if so, to obtain information about it. For this it uses an indexing procedure that produces tables with the positions of the vocabulary units in the corpus.

3.2. ATE Method

As mentioned earlier, ATE is a categorization problem in which, for every term candidate, a system will produce a score that will lead to the acceptance or rejection of said candidate. Our method also obeys this logic and has a battery of filters arranged in increasing order of computational complexity, finishing in a combination of statistical measures.

The initial exclusion rules are computationally inexpensive because they are based on stoplists and morphosyntactic patterns. The core of the method is the later application of a series of statistical measures such as frequency of occurrence in the corpus, dispersion (based on document frequency), and co-occurrence (the analysis of other words that share the same sentences with the candidate).

The first step of the terminology extraction procedure is the creation of lists of word *n*-grams (with *n* defined by the user, with 5 as default). Each of these is treated as a potential term and submitted to the following battery of measures:

- **Stoplist**: This is a set of simple exclusion rules to eliminate *n*grams that begin or end with a member of a list of function words (grammemes such as prepositions, articles, conjunctions). These function words are admitted inside the candidate, however, as it may occur with some *n*grams with $n > 2$ (e.g., in Spanish, the term *dióxido de carbono*, 'carbon dioxide').
- **Morphosyntactic patterns**: In this project we have opted to limit the number of term candidates to those which can be parsed as noun phrases. Candidates including other grammatical categories or patterns, such as verbs or adverbs, are excluded. This is certainly a limitation for users interested in specialized predicates, but these units may require a different methodology.
- **Term frequency**: For any candidate x that survives the previous filters, we calculate a number of statistical measures, among which there is term frequency: $f(x)$. This measure might not be useful in isolation or while analyzing a single document, as most terms in a text will be hapax or dis legomena. That said, term frequency can also be a useful indicator when used in conjunction with other statistical measures and when analyzing a large and coherent collection of specialized documents.
- **Dispersion**: In our method, this measure is defined as a combination of term frequency and document frequency, in turn defined as the number of documents in which a term occurs: $df(x)$. Figure 1 shows the dispersion patterns of two Spanish candidates of similar frequency in a short sample of 235 Spanish linguistics texts. The blue, continuous line corresponds to *equivalencia* (equivalence), an actual Spanish linguistics term, and the red, dotted one to *menor medida*, a fragment of the general vocabulary expression *en menor medida*, (to a lesser extent), which of course is not a term. Genuine terms show a specific dispersion pattern. When a term

occurs in a document, it is likely to occur again, and several times. On the contrary, words of the general vocabulary seem to be more uniformly distributed. This difference in patterns can be explained by the fact that terms that designate specialized concepts are more likely to be part of the topic of discourse.

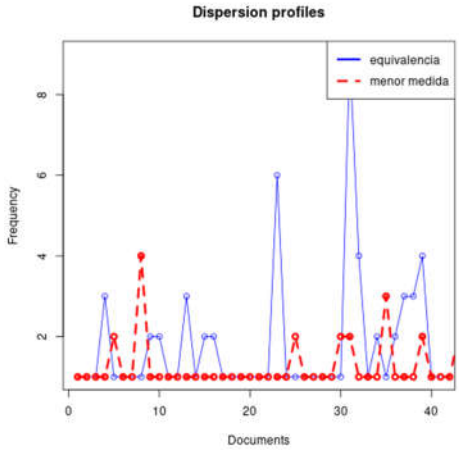

FIGURE 1. DISPERSION PROFILES OF *EQUIVALENCIA* ('EQUIVALENCE') AND *MENOR MEDIDA* ('LESSER EXTENT')

Term dispersion measures, such as TF-IDF, have been in use since the seventies (Spärck Jones 1972). Here, however, we try to find the simplest and fastest method to exploit dispersion patterns to measure how informative a term candidate is, and for that we use coefficient (1). The variable *h(x),* for hapax, is defined (2) as the number of documents in a collection *D* of *n* documents, in which term *x* has frequency 1. This last number could be a parameter, but we will leave that decision for future research.

$$d(x) = 1 - \frac{h(x)}{df(x)} \qquad (1)$$

$$h(x) = \sum_{i=1}^{n} \begin{cases} 1 & frec(D_i) = 1 \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

- **Co-occurrence**: As shown by early distributionalists (Harris 1954; Firth 1957), one can know a lot about a word by observing the company it keeps. In the case of terminology extraction, this means that terminological units are often revealed by their co-occurrence patterns, and this can be used as a robust predictor of the specialized value of a candidate. Terms show a tendency to co-occur with a reduced number of other terms that constitute their semantic field. For instance, one can expect that a linguistic term such as *consonant* will show a tendency to appear in the same sentences with other terms such as *vowel, aspiration, deletion*, etc. Similarly, *equivalencia* will co-occur with others such as *diccionario, definición, traducción, voz*, etc. (dictionary, definition, translation, expression). Figure 2 illustrates this by comparing

equivalencia with the non-term *menor medida*.

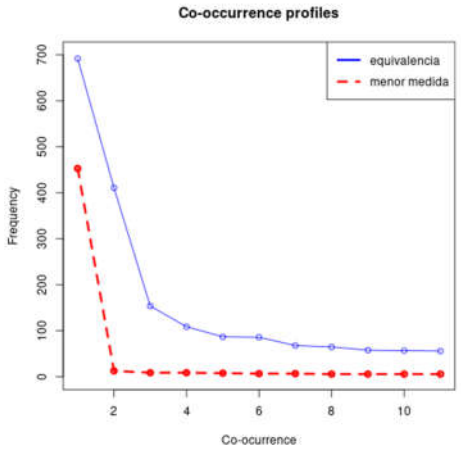

FIGURE 2. CO-OCCURRENCE PROFILE OF *EQUIVALENCIA* AND *MENOR MEDIDA*

$$c(x) = \frac{\log_2 \sum_{i=1}^{k} R_{x,i}}{\log_2 |m(x)|} \qquad (3)$$

We used the co-occurrence measure (3) to exploit this intuition. Here, x is a term candidate; Rx the set of (single) words co-occurring with x; $f(x)$ the frequency of x and Rx,i is the frequency of the ith most frequent co-occurring word in the contexts of occurrence of x. The value k is an arbitrary parameter. In our experiments, k = 20. Larger ks mean longer processing time, but more work should be carried out to find the best compromise between efficiency and efficacy.

- **Extras**: The prototype includes a module for the extraction of references and definitions from the corpus. With the help of these functions, we include a variable $v(x)$ to denote an additional value for x when found in the title of bibliographic references in the corpus and/or when there are definitional patterns in its immediate vicinity. Appearing in titles and being defined are both taken as indicators of the significance of a term.
- **Final score**: The system combines the (above mentioned) statistical measures in a final product score $s(x)$ for the ranking of the candidates (4).

$$s(x) = \sqrt{f(x)}.(1 + c(x)).(1 + d(x)).(1 + v(x)) \qquad (4)$$

After processing the corpus, the prototype presents the results in tables classified by language and shows tables for both the accepted as well as the rejected candidates. This way, upon inspection, users have the possibility of manually correcting the results by deleting false positives or rescuing false negatives from the rejected list. There is also the possibility of eliminating all candidates that show a final score under a given value or those that include any arbitrary component.

As an alternative, the program also offers the users the possibility of uploading a list of terms as examples that are used to narrow down results. This can help users obtain more refined results as the program will promote candidates which frequently co-occur with the examples. In particular, it can benefit those who need terms of a specific subfield (e.g., lexicology) but only have a corpus of the general area (linguistics).

4. Evaluation

There have been some proposals for ATE evaluation datasets in recent years. Among the most cited references, we find the ACL RD-TEC 2.0 dataset (QasemiZadeh and Schumann 2016) and the TermEval (Rigouts Terryn et al. 2020). Both are samples of specialized corpora with exhaustive annotations for both single and multiword terminology, intended to be used for the evaluation of precision and recall in ATE systems. These materials do not include text in Spanish, one of our languages of interest, but they do include English. That said, we can't use these materials anyway because they are both too short for our system, which has been designed to operate with massive collections. Our preliminary attempts evaluating our method with these datasets showed competitive precision but not enough recall. We would have to change the design of the algorithm to operate with such small datasets, perhaps by resorting to external resources like Wikipedia, treating it like a large corpus. While an interesting avenue for future work, at the moment, we opt to evaluate the system as it stands today, and that means to evaluate its performance on large corpora.

Thus, to conduct such a preliminary evaluation of the performance of this ATE method, we used two corpora of familiar areas: one of lexicography and another of general linguistics in English and Spanish. Their size is 34,456,086 and 32,387,689 tokens, respectively. Both corpora were compiled from open-access specialized journals and conference proceedings, and are similar in nature. In the case of the lexicography corpus, we used one compiled by Lindemann, Kliche and Heid (2018). The lexicography corpus, which can be considered a subfield (though not a subset) of the other, includes proceedings of Euralex and other conferences, as well as papers from open-source lexicography and lexicology journals. The linguistics corpus, in turn, consists of a sample of documents obtained from the linguistics journals listed in Table 1.

Journal	**ISSN**
Alfal	2079-312X
Anuario de letras	2448-8224
Boletín de Lingüística	0798-9709
Colombian Applied Linguistics Journal	0123-4641
Cuadernos de Lingüística Hispánica	0121-053X
Forma y Función	0120-338X
Íkala	0123-3432
Lenguaje	0120-3479
Letras	0459-1283
Lexis	0254-9239
Lingüística	2079-312X

Literatura y lingüística	0716-5811
Logos	0716-7520
Núcleo	0798-9784
RLA	0718-4883

TABLE 1. COMPOSITION OF THE LINGUISTICS CORPUS

To conduct this experiment, we followed the steps that regular users would follow to extract terms from their corpus. For each corpus, we uploaded the documents as a single ZIP file containing PDF files of mainly English and Spanish text. The program extracted the ZIP file and then converted each PDF file to plain text, recognizing the main language in the process and then eliminating those fragments of text in other languages. It then submitted the corpus to a tagging procedure and created an index table of the vocabulary.

Once this process finished, we submitted the corpus to the terminology extraction process. There are different parameters for the extraction, such as a minimum term frequency and document frequency. For this evaluation, we used a rather restrictive setting: a minimum dispersion of 10 documents, and a minimum frequency of 10, which benefits precision in detriment of recall.

In order to proceed with the evaluation, we took a random sample of 200 candidates from each of the two corpora. Each sample mixes 100 accepted and 100 rejected candidates. Four specialists, all linguists with experience in lexicography projects, were recruited for the task of evaluating these results. To prevent any possible bias on the part of the human evaluators, they were presented with the sample of terms in alphabetical order, not knowing if each candidate was accepted or rejected by the system. They were asked to mark with a number one next to each candidate that, according to their criterion, should be considered a term. They conducted the task individually and without consultation to external resources.

The measure of inter-rater agreement of the group of evaluators gives a Fleiss' Kappa coefficient of 0.553 (0.776 observed vs. 0.5 expected agreement), which can be interpreted as moderate, close to substantial, agreement. Table 2 shows a fragment of the evaluation matrix that compares the decisions made both by humans and the machine (in columns) for each *n*gram (in rows) for the case of the linguistics corpus with both English and Spanish results. Some decisions of the evaluators, e.g., H4 on *fascicle* or H3 on *fabricación* (manufacturing) may be explained as cases of momentary loss of concentration.

***n*gram**	**M**	**H1**	**H2**	**H3**	**H4**
electronic corpora	1	1	1	1	1
enunciado	1	1	1	1	1
erelt	0	0	0	0	0

escuadrilla	0	0	0	0	0
Estonian language	1	1	0	1	1
etymon	1	1	1	1	1
exemplary sentences	0	1	1	1	1
expresión	1	1	1	1	1
fabricación	1	0	0	1	0
fascicle	1	0	0	0	1

TABLE 2. EXAMPLES OF *N*GRAMS AND THEIR REJECTION OR ACCEPTANCE (MARKED WITH A NUMBER 1) BY HUMANS AND THE MACHINE

Figures 3 and 4 illustrate with dendrograms the similarities between individuals in this group in the corpus of lexicography and linguistics, respectively. To complete the panorama, we added a pair of random selectors. As the figures show, the program is not yet as good as the human specialists, but the distance with them is significantly closer compared with random selections. The figures also show that humans group differently according to the corpus: H1 and H2 are similar in lexicography but dissimilar in linguistics. This is a phenomenon that also deserves further research, but it would be in a different area, that of the measurements of terminological competence in professionals.

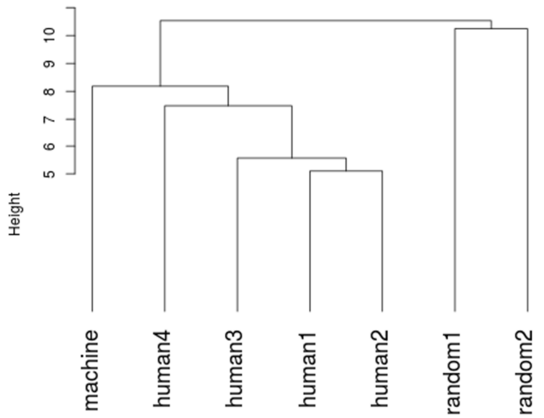

FIGURE 3. DENDROGRAM SHOWING THE SIMILARITIES BETWEEN MACHINE, HUMAN ANNOTATORS AND RANDOM SELECTIONS IN A LEXICOGRAPHY CORPUS

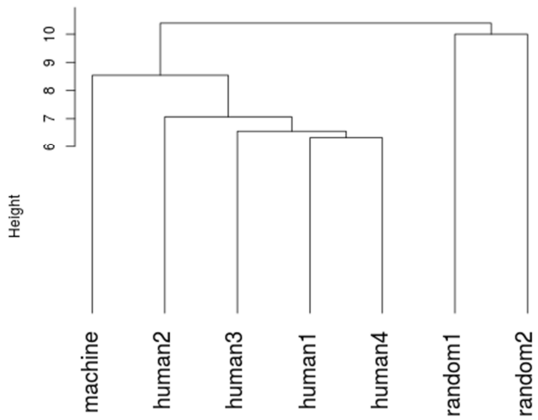

FIGURE 4. DENDROGRAM OF THE SIMILARITIES BETWEEN ANNOTATORS IN A LINGUISTICS CORPUS

We derived a gold standard from the evaluators' agreement to obtain measures of precision and recall. For this, we assumed a term is effectively a term if at least two humans selected it as such. Against this gold standard, we were able to calculate true positives as cases where humans and the machine agree, false positives as those cases when a candidate is promoted only by the machine, and false negatives when humans promote a term that the machine rejects.

As a general reference, we compare this performance with some baselines: the frequency of the terms in the analyzed corpus (i.e., to accept the most frequent half of the sample) and another based on the notion of weirdness or keywordness, as proposed by various authors (Section 2). This baseline promotes those terms that are frequent in the analyzed corpus but infrequent in a reference corpus of the same size (i.e., a collection of newspaper articles). In this way, this baseline will accept a candidate *x* if the ratio between term frequency *tf(x)* and its reference frequency *rf(x)* is greater than an arbitrary threshold *k* (Equation 5). As reference corpus we used the one offered by the Leipzig Corpora Collection (Quasthoff, Goldhahn and Eckart 2014). Table 3 shows results for both corpora.

$$weirdness(x) = \begin{cases} 1 & \frac{tf(x)}{1+rf(x)} > k \\ 0 & \text{otherwise} \end{cases} \qquad (5)$$

Linguistics	***tp***	***fp***	***fn***	***tn***	**Pre**	**Rec**	**F1**
prototype	77	23	37	63	**.77**	**.67**	**.71**
frequency	61	41	53	45	.59	.53	.56
weirdness	56	45	58	41	.55	.49	.52
Lexicography	***tp***	***fp***	***fn***	***tn***	**Pre**	**Rec**	**F1**
prototype	71	29	36	64	**.71**	**.66**	**.68**
frequency	58	40	49	53	.59	.54	.56
weirdness	56	45	51	48	.55	.52	.53

TABLE 3. EVALUATION OF THE TERM EXTRACTION PROCESS

As shown in Table 3, results in both corpora are significantly better than the baselines, with 19 points over weirdness in the case of linguistics. Results in linguistics and lexicography are also relatively similar in F1 (.71 ~ .68), but linguistics is slightly better, especially in terms of precision, with 6 points (.77 against .71). One explanation for the higher number of false positives in the lexicography corpus is the fact that common vocabulary words, completely unrelated to the domain, are mentioned frequently in these papers as part of the object of study. Working in a domain like this proved to be a great challenge for an ATE system. If we were to work with, say, a psychiatry corpus, then the topics would likely be mental disorders or treatments, all concepts that would be designated with specialized terms. In the case of a lexicography corpus, however, one frequently observes common words incorrectly promoted as terms because they are often the topic of discourse (e.g. *abdication, bagpipe, beaver, cocktail*). To complicate things further, many texts in this corpus describe bilingual dictionaries, which implies that there will be words in other languages (e.g., French words *chien, coeur*, Italian *acqua*, Czech *jazyka*, Solevene *koga*, Georgian *mazdar*, Dutch *lopen*, and so on).

5. Conclusions, Work in Progress and Future Plans

In this article, we offered a brief description of a web-based prototype for computer assisted-terminology. In particular, we covered the ATE-related functions and included a quantitative evaluation of the system's current performance.

After our assessment, we conclude that the methods here described show promising results and can be helpful for professional terminologists. The current implementation of the method is available as a demo operating in English and Spanish. Among its advantages, we highlight that it consists of conceptually simple algorithms, dispersion, and co-occurrence statistics that can be produced with cheap hardware.

As already pointed out, there is a lack of freely available and easy to use software products that can provide solutions for terminology projects, especially terminology extraction from large corpora. For this reason, we also believe the prototype can be used as a tool for educational purposes in terminology, particularly in settings where commercial applications are less likely to be available. A tool that can help students learn by doing might help university professors who struggle to maintain their audience's interest in terminology theory.

To mention our work in progress, we are currently implementing new functions to complement the ones explained in the present article. Once the ATE process is finished, the prototype offers functions to populate the terminology database with the following fields:

- **Grammatical data**: this includes inflected forms, gender, part of speech / morphosyntactic pattern and other morphological information.
- **Semantic categorization**: it generates full hypernymy chains for each term in each language, based on an algorithm that combines co-occurrence statistics and morphosyntactic patterns.
- **Semantic clustering**: it produces clusters (groups) of terms that are semantically related. In this case, the semantic relation is operationalized as co-occurrence associations. The method we implemented for this is based on co-occurrence graphs.

- **Definitions**: extracts definitions of the terms from the corpus based on a list of manually curated definitional patterns in English and Spanish.
- **Bilingual alignment**: it produces a bilingual alignment of the terms by applying a combination of dispersion and co-occurrence association measures, also including an orthographic similarity coefficient for the cognates.
- **Term variants**: it detects term variants, i.e. terms in the same language which have different forms but the same meaning. Our proposal to address this problem is based on the intuition that two terms in the same language *i* and *j* can be considered term variants if they consistently show a tendency to share the same equivalence in the other language.

Regarding future developments, we expect to adapt the system to work with other languages such as French, Catalan, German and eventually other languages. This should be feasible as the method is fundamentally based on statistical and language-agnostic algorithms. We will also continue evaluating performance in different domains and using datasets provided by other researchers. Another future plan is to improve computational efficiency and reduce processing times. This is because the method, although conceptually simple, entails some computational cost, mainly due to the detailed analysis of the contexts of occurrence of every analyzed term candidate. This poses a challenge for its practical use as a web platform when dealing with massive corpora (tens of millions tokens). There is certainly work to be done in that direction, but the margin for improvement remains, for the moment, unknown.

References

Ahmad, Khurshid, Gillam, Lee, and Tostevin, Lena. 1999. "University of Surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder)". In *TREC*, volume 500-246 of NIST Special Publication. National Institute of Standards and Technology (NIST).

Aker, Ahmet, Paramita, Monica and Gaizauskas, Rob. 2013. "Extracting bilingual terminologies from comparable corpora". In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 402-411.

Anthony, Laurence. 2005. "Antconc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom". In *2005 IEEE International Professional Communication Conference Proceedings, IPCC 2005*: 729–737.

Arntz, Reiner, and Picht, Heribert. 1995. *Introducción a la Terminología*. Madrid: Pirámide. Fundación Germán Sánchez Rupiérez.

Baisa, Vít, Michelfeit, Jan, and Matuška, Ondˇrej. 2017. "Simplifying terminology extraction: Oneclick terms". Paper presented at Corpus Linguistics 2017 Conference, University of Birmingham, July 25-28, 2017. https://www.birmingham.ac.uk/documents/college-artslaw /corpus/conference-archives/2017/general/paper385.pdf

Bordea, Georgeta, Buitelaar, Paul, Faralli, Stefano, and Navigli, Roberto. 2015. "Semeval-2015 task 17: Taxonomy extraction evaluation (texeval)". In *Proceedings of the 9th International*

Workshop on Semantic Evaluation (SemEval 2015): 902–910. Denver, Colorado: Association for Computational Linguistics.

Bourigault, Didier, Gonzalez-Mullier, Isabelle, and Gros, Cécile. 1996. "Lexter, a natural language processing tool for terminology extraction". In *Proceedings of the 7th EURALEX International Congress*: 771–779. Göteborg: Novum Grafiska AB.

Cabré, María Teresa. 1999. *La Terminología: Representación y Comunicación*. Barcelona: Institut Universitari de Lingüística Aplicada.

Cabré, María Teresa, Estopà, Rosa, and Vivaldi, Jorge. 2001. "Automatic term detection: A review of current systems". In *Recent Advances in Computational Terminology*, edited by Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme, 53–87. Amsterdam: John Benjamins.

Conrado, Merley, Pardo, Thiago, and Rezende, Solange. 2013. "A machine learning approach to automatic term extraction using a rich feature set". In *Proceedings of the 2013 NAACL HLT Student Research Workshop*: 16–23, Atlanta, Georgia: Association for Computational Linguistics.

Cram, Damien, and Daille, Béatrice. 2016. "Terminology extraction with term variant detection". In *Proceedings of ACL-2016 system demonstrations*: 13–18.

Daille, Béatrice. 1994. "Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques". PhD dissertation. Université Paris Diderot.

de Schryver, Gilles-Maurice, and Joffe, David. 2023. "The end of lexicography, welcome to the machine: On how chatGPT can already take over all of the dictionary maker's tasks". Talk presented at *20th CODH Seminar*, at Center for Open Data in the Humanities, Research Organization of Information and Systems, National Institute of Informatics, Tokyo.

Drouin, Patrick. 2003. "Term extraction using non technical corpora as a point of leverage". *Terminology*, 9(1): 99–115.

Felber, Helmut. 1984. *Terminology Manual*. Paris: United Nations Educational, Scientific and Cultural Organization, International Information Centre for Terminology.

Filippova, Darya, Can, Burcu, and Corpas Pastor, Gloria. 2021. "Bilingual terminology extraction using neural word embeddings on comparable corpora". In *Proceedings of the Student Research Workshop Associated with RANLP 2021*: 58–64.

Firth, John. 1957. "A synopsis of linguistic theory, 1930-55". In *Studies in Linguistic Analysis*, 1–31. Oxford: Blackwell.

Frantzi, Katerina, Ananiadou, Sophia, and Mima, Hideki. 2000. "Automatic recognition of multi-word terms: The c-value/nc-value method". *International Journal on Digital Libraries*, 3(2): 115–130.

Haque, Rejwanul, Penkale, Sergio, and Way, Andy. 2018. "Termfinder: Log-likelihood comparison and phrase-based statistical machine translation models for bilingual terminology extraction". *Language Resources and Evaluation*, 52(2): 365–400.

Harris, Zellig. 1954. "Distributional structure". *Word*, 10(2-3): 146–162.

Hearst, Marti A. 1992. "Automatic acquisition of hyponyms from large text corpora". In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*.

Heylen, Kris, and De Hertog, Dirk. 2015. "Automatic term extraction". In *Handbook of Terminology. Volume 1*, edited by Hendrik J. Kockaert and Frieda Steurs, 203–221. Amsterdam: John Benjamins.

Humbley, John. 2022. "The reception of Wüster's general theory of terminology". In *Theoretical Perspectives on Terminology. Explaining Terms, Concepts and Specialized Knowledge*, edited by Pamela Faber and Marie-Claude L'Homme, 15–36. Amsterdam: John Benjamins.

Hutchins, John. 1998. "The origins of the translator's workstation". *Machine Translation*, 13(4): 287–307.

Justeson, John S., and Katz, Slava M. 1995. "Technical terminology: Some linguistic properties and an algorithm for identification in text". *Natural Language Engineering*, 1(1): 9–27.

Kageura, Kyo and Umino, Bin. 1996. "Methods of automatic term recognition: A review". *Terminology*, 3(1): 259–289.

Lang, Christian, Wachowiak, Lennart, Heinisch, Barbara, and Gromann, Dagmar. 2021. "Transforming term extraction: Transformer-based approaches to multilingual term extraction across domains". In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*: 3607–3620.

Lefever, Els, Macken, Lieve and Hoste, Veronique. 2009. "Language-independent bilingual terminology extraction from a multilingual parallel corpus". In *Proceedings of 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, 496-504.

Lindemann, David, Kliche, Fritz, and Heid, Ulrich. 2018. "Lexbib: A Corpus and Bibliography of Metalexico graphical Publications". In *Proceedings of EURALEX 2018*, 699–712.

Meyer, Ingrid. 2001. "Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework". In *Recent Advances in Computational Terminology*, edited by Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme, 279–302. Amsterdam: John Benjamins.

OpenAI. 2023. "Gpt-4 technical report". Last revised March 27, 2023. *arXiv:2303.08774 [cs.CL]*.

Pavel, Silvia and Nolet, Diane. 2002. *Manual de Terminología. Translation Bureau*. Québec: Public Works and Government Services.

Pearson, Jennifer. 1998. *Terms in Context*. Amsterdam: John Benjamins.

QasemiZadeh, Behrang and Schumann, Anne-Kathrin. 2016. "The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods". In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16), Portorož. European Language Resources Association (ELRA).

Quasthoff, Uwe, Goldhahn, Dirk, and Eckart, Thomas. 2014. "Building large resources for text mining: The Leipzig Corpora Collection". In *Text Mining: From Ontology Learning to Automated Text Processing Applications*, edited by Chris Biemann and Alexander Mehler, 3–24. Cham: Springer.

Rigouts Terryn, Ayla, Hoste, Veronique, Drouin, Patrick, and Lefever, Els. 2020. "TermEval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (ACTER) dataset". In *Proceedings of the 6th International Workshop on Computational Terminology*, 85–94, Marseille. European Language Resources Association.

Rigouts Terryn, Ayla, Hoste, Veronique, and Lefever, Els. 2022. "D-terminer: online demo for monolingual and bilingual automatic term extraction". In *Proceedings of the Workshop on Terminology in the 21st century*: 33–40.

Sager, Juan C. 1990. *A Practical Course in Terminology Processing*. Amsterdam: John Benjamins.

Shwartz, Vered, Santus, Enrico, and Schlechtweg, Dominik. 2017. "Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection". In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*: 65–75.

Simões, Alberto and Almeida, José João. 2008. "Bilingual terminology extraction based on translation patterns". *Procesamiento del Lenguaje Natural*, (41): 281–288.

Spärck Jones, Karen. 1972. "A statistical interpretation of term specificity and its application in retrieval". *Journal of Documentation*, 28: 11–21.

Steurs, Frieda, De Wachter, Ken, and De Malsche, Evy. 2015. "Terminology tools". In *Handbook of Terminology. Volume 1*, edited by Hendrik J. Kockaert and Frieda Steurs, 222–249. Amsterdam: John Benjamins.

Straka, Milan and Straková, Jana. 2017. "Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe". In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*: 88–99. Vancouver: Association for Computational Linguistics.

Tran, Hanh Thi Hong, Martinc, Matej, Caporusso, Jaya, Doucet, Antoine, and Pollak, Senja. 2023. "The recent advances in automatic term extraction: A survey". *arXiv:2301.06767 [cs.CL].*

Ville-Ometz, Fabienne, Royauté, Jean, and Zasadzinski, Alain. 2007. "Enhancing in automatic recognition and extraction of term variants with linguistic features". *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 13(1): 35–59.

Wüster, Eugen. 1979. *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*. Wien: Springer.

Zhang, Chunxia and Jiang, Peng. 2009. "Automatic extraction of definitions". In *Proceedings of the 2009 2nd IEEE International Conference on Computer Science and Information Technology, ICCSIT 2009*: 364–368.