Nottingham 2011

The Call Triangle: student, teacher and institution

Correcting erroneous N+N structures in the productions of French users of English

Marie Garnier[*]

Equipe Cultures Anglo-Saxonnes, Université Toulouse Le Mirail, 5 allées Antonio Machado, 31058 Toulouse, France
Institut de Recherche en Informatique de Toulouse, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse, France

Abstract

This article presents the preliminary steps to the implementation of detection and correction strategies for the erroneous use of N+N structures in the written productions of French-speaking advanced users of English. This research is carried out as part of the grammar checking project *CorrecTools*, in which errors are detected and corrected using linguistic-based NLP techniques. We use information from comprehensive and student grammars as well as the results of linguistic research in order to identify a set of constraints on the formation of N+N structures. We use these constraints to propose a classification of the N+N errors found in our corpus and design detection and correction strategies for each error type.

Keywords: NLP; grammar checking; N+N structures; error analysis

1. Introduction

N+N structures are composite nominals in which the head and the attributive dependent(s) are nouns (Payne & Huddleston, 2002). In other words, they are NPs in which one or more nouns are used in modifier or complement function, e.g. *university library, peanut butter cookie*. They are syntactic constructions, and as such they differ from compound nouns (Payne & Huddleston, 2002), although this distinction is still controversial (Giegerich, 2004) and the two constructions are often conflated in SLA research and NLP research on multiword expressions.

N+N structures are common in English but are not a regular feature of French. Moreover, since the use of such structures is linked to a high level of proficiency in an L2 (Olshtain, 1987) and there is a lack of straightforward semantic rules for their formation, they are seldom taught in language courses. As a result, the use of such structures leads to errors of different types (see section 2).

This paper presents the preliminary steps to the implementation of strategies for the automatic detection and correction of errors linked to the use of N+N structures. It is part of a larger grammar checking project (*CorrecTools*) targeted at French-speaking advanced users of English in professional and personal contexts. We focus on errors which are not detected by existing grammar checkers. Errors are automatically detected and corrected with linguistic-based NLP techniques using <TextCoop> (Saint-Dizier, 2010), a logic-based platform for language processing implemented in Prolog.

[*] Contact author. Tel.: +33 (0)5 61 55 62 44
E-mail address: garnier@irit.fr

The error types that are the focus of the project have been selected through the manual analysis of a corpus of productions by French-speaking advanced users of English (i.e. *CEFRL* levels B2-C1) in professional and personal contexts (Albert, Garnier, Rykner, & Saint-Dizier, 2009). It includes scientific papers, learner productions (French section of the *International Corpus of Learner English* (Granger, Dagneaux, Meunier, & Paquot, 2009)), personal and professional emails, and technical reports.

2. Identification of errors

2.1. Previous work

The body of research in linguistics on N+N structures is too broad to be fully reviewed here; we should however point out there is no consensus on the list of semantic relations that can exist between the head and its nominal dependent(s) (Benczes, 2006). Research in SLA has investigated the acquisition and use of noun compounds. NLP research on N+N structures has been focused on disambiguation, accurate bracketing and the automatic recognition of semantic relationships (e.g. Barker & Szpakowicz, 1998; Buckeridge & Sutcliffe, 2002). The detection and correction of N+N errors produced by users of English has not yet been the subject of research in the grammar checking branch of NLP. N+N structures have not been a major focus of CALL either: to the best of our knowledge, there exists only one project centered on the teaching of English compounds (Boucher, Danna, & Sébillot, 1993).

2.2. Presence of N+N errors in the corpus

In our corpus, the use of N+N structures is the second most important cause for errors in the Noun Phrase after determination, and are especially prominent in the subcorpus of scientific papers, which accounts for 71.83% of all N+N errors. N+N structures tend indeed to be more frequently found in technical documents (Buckeridge & Sutcliffe, 2002). We ran a tailored search for N+N structures in the entire French section of the *ICLE* (226,922 words) and found that 5.32% of relevant N+N segments carried an error.

2.3. Classification of N+N errors

We used a comprehensive descriptive grammar of English (Payne & Huddleston, 2002), an explicative grammar of English for French-speaking students (Larreya & Rivière, 2005) as well as the results from linguistic research in order to identify constraints on the formation of N+N structures. We were able to identify a set of four constraints:

- the noun that functions as a modifier (N^1) is usually 'number-neutral', as are adjectives, and is therefore in the singular form;
- the noun that functions as a modifier (N^1) has generic reference rather than definite reference;
- N+N structures of more than two nouns are ambiguous and may be hard to understand;
- although there is no definite list of possible semantic relationships between the head noun and its noun modifier(s), not all semantic relationships can be represented with N+N structures.

We found four error types that correspond to these constraints. In addition, there is an overuse of the definite article *the* with N+N structures in our corpus, and a high number of N+N errors in the learner corpus is due to the use of an erroneous N+N structure where another expression exists in the lexicon (e.g. *consumption society*). Six main error types were thus identified. The following table presents the distribution of errors according to these types in the subcorpus of scientific papers (46,140 words) and in the entire French section of the *ICLE*:

Type	Example	% in scientific papers	% in learner prod.
Types of N+N errors linked to semantics			
Definite N^1	**the domain specificities*	14.49	0
Unusual semantic relationship	**his love obsession*	30.43	13.33
Other types of N+N errors			
Morphology	*the *ghettos sickness*	23.19	22.22
Stacking	**information extraction technology results*	14.49	0
Lexicon	*our *consumption society*	0	42.22
Article use	**the music detection*	17.39	22.22

Table 1. Distribution of N+N errors according to type in 2 relevant subcorpora.

Erroneous segments may carry several errors, e.g. **semantic links classes hierarchy*. This is the case for 35.3% of errors in the subcorpus of scientific papers. For feasibility reasons, we set aside lexical errors and focus on N+N errors linked to the other five main error types.

3. Methods and strategies

3.1. Challenges

The detection and correction of N+N errors entails several challenges:

- accurately analyze segments as Noun Phrases and as errors using the resources available in <TextCoop>;
- organize rules to avoid conflicts;
- deal with segments that contain several N+N errors;
- deal with the addition of accurate prepositions and necessary determiners, as well as with the scope of adjectives that might be present in the original structure.

3.2. Detecting and correcting N+N errors

As a preliminary step, NPs are detected using a small hand-coded grammar of the NP. N+N errors are then detected and corrected using patterns, rewriting rules and additional heuristics. We use several types of resources, i.e. lexicons present in <TextCoop>, additional tailored lexicons, and web resources. The two error types linked to semantics are treated together: the segments that carry these types of errors have the same surface structure and can be corrected in the same way (see below). Since these strategies are designed to be included in a grammar checking/tutorial system with corrective feedback (Garnier, 2011), they rely on collaboration between user and system in order to complete corrections.

Prepositions are selected using lexical information available with nouns, especially deverbal nouns; *of* is inserted by default (e.g. **his love obsession* → *his obsession* ***with*** *love* (to be obsessed **with**); **the domain specificities* → *the specificities* ***of*** *the domain*). If the N+N structure is developed into [Noun + Prepositional Phrase], *the* is inserted by default in front of the head noun if it is in the singular (e.g. **queries treatment* → ***the*** *treatment of queries*).

The following list presents the treatment of each error type. They are presented in the order in which they are run, which has been designed to minimize the number of interventions on each NP while detecting as many errors as possible:

1. **Stacking errors:**
   - **information extraction technology results* → [N^1+N^2+N^3+N^4] = error?
   - automatic query to Google Scholar to evaluate the frequency of use of the segment
   - low frequency = rewrite segment:
     - [Det. + N^4 + Prep. + (Det.) + (N^1+N^2+N^3)]: **the results of information extraction technology*
     - [Det. + (N^3+N^4) + Prep. + (Det.) + (N^1+N^2)], etc.
   - intervention of the user requested in order to select an adequate correction.

2. **Morphology errors:**
   - *the *ghettos sickness* → [(Det.)+N^1(plural)+N^2] = error
   - rewrite segment:
     - [N^1(singular)+N^2]: *the ghetto sickness*
     - [Det. + N^2 + Prep. + ((Det.)+N^1)]: *the sickness of the ghettos*
     - [(Det.) + N^1 + Gen. + N^2]: *the ghettos' sickness.*
3. **Semantic errors:**
   - *his *love obsession* → [(Det.)+N^1+N^2] = error?
   - segment matched with a list of noun-noun compounds
   - not found = automatic query in Google N-gram corpus
   - low frequency = rewrite segment:
     - [(Det.) + N2 + Prep. + (Det.) + N1]: *his obsession with love.*
4. **Article use errors:**
   - **the music detection* → [Art.+N^1+N^2] = error?
   - segment matched with a list of noun-noun compounds
   - not found = automatic query in Google N-gram corpus
   - low frequency = rewrite segment:
     - [N^1 + N^2]: *music detection*
     - [Art. + N2 + Prep. + (Det.) + N1]: *the detection of music*

4. Future work

The work presented in this paper constitutes the preliminary step to the implementation of these strategies in <TextCoop>. The correction of N+N errors presents a number of important difficulties, and a several issues need to be investigated further before implementation is possible, e.g. the choice of preposition and the prediction of article insertion. Once implemented, the strategies will be submitted to testing and evaluation. The project also includes the generation of adaptive corrective feedback.

5. References

Albert, C., Garnier, M., Rykner, A. & Saint-Dizier, P. (2009a). Elements de stratégies de correction automatique de textes : le cas des francophones s'exprimant en anglais. In I. Biskri & A. Jebali (Eds.), *Multilinguisme et traitement des langues naturelles* (pp. 55-70). Québec: Presses de l'Université du Québec.

Barker, K., & Szpakowicz, S. (1998). Semi-automatic recognition of noun modifier relationships. *Proceedings of the 36th annual meeting on Association for Computational Linguistics* - (Vol. 1, pp. 96-102). Morristown, NJ, USA: Association for Computational Linguistics.

Benczes, R. (2006). *Creative Compounding in English*. Philadelphia, Amsterdam: John Benjamins.

Boucher, P., Danna, F., & Sébillot, P. (1993). Compounds: an Intelligent Tutoring System for Learning to Use Compounds in English. *Computer-Assisted Language Learning, 6*(3), 249-272.

Buckeridge, A. M., & Sutcliffe, R. F. E. (2002). Disambiguating noun compounds with latent semantic indexing. *COLING-02 on COMPUTERM 2002 14*, 1-7. Morristown, NJ, USA: Association for Computational Linguistics.

Garnier, M. (2011). Explanation and corrective feedback in grammar checking systems. *Proceedings of the 6th International ExACt workshop at IJCAI'11*, 81-90.

Giegerich, H. J. (2004). Compound or phrase? English noun-plus-noun constructions and the stress criterion. *English Language and Linguistics*, *8*(1), 1-24.

Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. (Eds). (2009). *International Corpus of Learner English* v.2. Louvain La Neuve: Presses Universitaires de Louvain.

Larreya, P. & Rivière, C. (2005). *Grammaire explicative de l'anglais*, 3ème ed. Paris: Pearson Longman.

Olshtain, E. (1987). The acquisition of new word formation processes in second language acquisition. *Studies in Second Language Acquisition*, *9*, 221-231.

Payne, J., & Huddleston, R. (2002). Nouns and noun phrases. In R. Huddleston & G. K. Pullum (Eds), *The Cambridge Grammar of the English Language* (pp 323-524). Cambridge: Cambridge University Press.

Saint-Dizier, P. (2011). <TextCoop>, un analyseur de discours base sur des grammaires logiques. *Proceedings of TALN'11*.