EPiC Series in Language and Linguistics
Volume 1, 2016, Pages 181–190
CILC2016. 8th International
Conference on Corpus Linguistics

Sentence length and NP complexity of general and medical written academic and media texts. An analysis using a trained syntactic parser.

Carlos Herrero Zorita
Antonio Moreno Sandoval

Autonomous University of Madrid
carlos.herrero@uam.es, antonio.msandoval@uam.es

Abstract

The main objective of this work is to perform a comparative analysis of sentence and main noun phrases complexity in two different types of discourses, written media and academic prose, using a trained syntactic parser (Stanford PCFG Parser). For this purpose, we have selected three written sources: a general media corpus, a medical media subcorpus and a medical academic prose subcorpus. From a total of more than 160000 sentences, we have carefully selected as the study sample a total of 300, which have been morphologically and syntactically annotated. Influenced by other studies related to syntax and statistics, our hypothesis is that NPs from academic prose and written media will contain four or more words, and those belonging to academic prose will be larger than the latter. The NPs studied are those that perform the main functions of the clause: subject, object (direct and indirect), attribute and time expressions. The results show a confirmation of our hypothesis. The academic subcorpus has the longest sentences and more complex NPs than the other texts. On the other hand, written media corpora achieve smaller NPs but their results are quite similar.

1 Introduction

This study we will describe in the following pages belongs to a wider project: the creation of an automatic Spanish syntactic parser. This parser has been created through a training process; that is, from a newspaper Treebank corpus, a series of grammatical rules are automatically deducted, which can be used in the future to syntactically tag any text. In-between this process, we have performed a syntactic

A.Moreno Ortiz and C.Pérez-Hernández (eds.), CILC2016 (EPiC Series in Language and Linguistics, vol. 1), pp. 181–190

study regarding the complexity of the noun phrases of a group of sentences, which will be described below. The purpose of this study is to test the parser on new texts as well as performing a preliminary work that could outline a methodology for future, more complex studies based on syntax and text typology.

The main objective of this work is to perform a comparative analysis of noun phrases' (NPs from now on) complexity in two different types of texts: written media and written academic prose. This comparison will be performed automatically, using our trained syntactic parser for Spanish. Following Evert and Krenn's (2005) claim, we believe that the experiment should be performed in similar conditions as the measurements. That is, since the parser has been trained with newspaper sentences, an equal number of samples have been selected for the comparison from medical written media (taken from medical journals). The academic prose texts will be also used as a way of finding significant differences.

This study is heavily influenced by other studies related to syntax and quantitative statistics using corpora for purposes such as discourse analysis or text typology and identification. In our case, our study will focus complexity of NPs, the main conveyors of information in a sentence. The issue of *complexity* of an NP is, however, not easy to define. As (Berlage, 2014: 33) explains, there are many ways to measure NP complexity and many previously work on the matter. We can measure the number of words, syllables, dative and genitive alternation, phonology, etc. For this study we have selected three methods based on Berlage's work: (1) NP length –NP (b) is more **complex because it has more words–; (2) NP structure or tree height –NP (b) is more complex** because it has more phrasal nodes, syntactically speaking–; and (3) sentential NP level –NP (b) is more complex because it contains a clause.

a) **The man** came home
b) **The man I saw at the park** came home

Table 1 and Figure 1 represents an example of NP complexity measurements for the NP subject "The man I saw at the park", from sentence b, represented in Figure 1:

The man I saw at the park		
Length	7	7 words: The / man / I / saw / at / the / park
Height	7	7 node levels in total
Sentential	Yes	It contains a clause (CL)

Table 1: Example of NP complexity

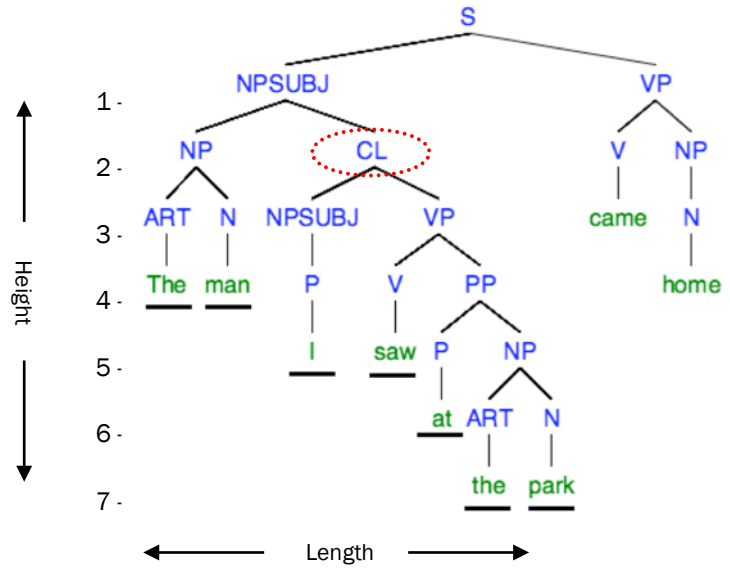

Figure 1: Graphical representation of Table 1

We have selected 100 sentences from the Treebank newspaper media corpus as reference group and compared them with 100 sentences from medical media and another 100 from medical academic subcorpora. Our assumption, based on Biber et al. (1999, 97), are the following:

- NPs from academic prose and written media will contain around four or more words due to the higher lexical density of this type of texts.

- NPs from the academic texts will be more complex than those from newspaper and medical media. The written medium allows information to be more extensive than in spoken discourse, especially in academic and fiction prose, where elements such as argumentation, evaluation, dialogue passages, etc. enlarges the text. Written news reportage also allows extensive writing. However since its aim is to convey as many information as possible in less space, we can assume media that NPs will be lighter than the academic texts.

- NPs in the newspaper and medical media will be more similar than those from academic texts, since they belong to the same domain, we can assume the NPs will be more similar between each other than those from the academic texts.

If the syntactic tagging process is able to detect these differences already spotted in previous research, we could assume the parser is working correctly.

2 Data

As stated above, the data and samples for this purpose have been extracted from three different written sources:

- **Newspaper media source**: *UAM Spanish Treebank* corpus (Moreno et al., 2003). It is a collection of 1600 syntactically analysed newspaper sentences, manually annotated by linguists at the Laboratorio de Lingüística Informática – UAM. This corpus follows the Penn Treebank style but adapted to the Spanish language: POS tags and syntactic features have been conformed to fit Spanish sentence structure and word functions (see Moreno et al 1999 for a detailed description). The corpus consists of nearly 25000 words and averages 15 words/sentence. Data were taken from the newspaper *El País* and the consumer association magazine *Compra Maestra*. This set will serve as the reference and for training the parser. Figures 2 and 3 show respectively the overall phrase distribution and the words per sentence distribution.

- **Medical media source**: *Otro Médico* / *OCU Salud* subcorpora (from *Multimedica* corpus (Moreno and Campillos, 2013). A collection of more than 330000 words from medical journals written for the general public was gathered by the same research team for the MultiMedica project. The register style is journalistic and informative, whose aim is to spread information related to Health and diseases, as well as nutrition. This source has been used in this experiment for written medical media evaluation.

- **Academic source**: *Harrison* subcorpus (also from *Multimedica* corpus). A collection of texts comprised of nearly 4 million words selected from the medical manual *Harrison*. This corpus includes professional and scientific documents written by medical doctors, instead of journalists. Here it has been used for the academic prose evaluation.

In summary, we have two different types of texts, the written media (news and medical) and the academic (medical) prose. The two questions to explore are a) whether the noun phrases have the same complexity among these domains and types; and b) which type has the most complex NPs? The previous studies of Biber and Conrad for English revealed that nouns are very common in both types but even more common in newspapers than in academic prose. Since Spanish has no noun pre-modification (excluding morphological compounding) we expect an extensive use of prepositional phrases after nouns, as well as relative clauses. In the *Longman Grammar of English*, news register shows longer NPs than the academic one, both in premodification and postmodification. On the other hand, prepositional phrases are extremely common in academic prose, and relative clauses are more frequent in news than academic texts. Biber (2006:224) argues that "'ideational' functions (using language to convey propositional information) are important for explaining the dense use of nouns, adjectives, prepositional phrases, and complex noun phrases in the written university registers.

3 Methodology

The procedure for this study has followed several steps, which will be explained in this chapter: (1) Creation of the parser; (2) Sentence length study and selection of samples from the corpora; (3) Tagging of samples; (4) NP study.

First of all, the *Spanish Treebank* was used to create and train a syntactic parser using the *PCFG Stanford Parser* (Klein and Manning, 2013). Even still in a beta phase (the training has only been done with roughly 1600 sentences) it already achieves positive results that allow us to perform preliminary syntactic studies.

Secondly, we began with an analysis of the length of the sentences of the three corpora. These first results, presented in Table 1, already reveals differences between the media and the academic texts: the

most frequent sentences –those belonging to 1σ of the Standard Deviation– are roughly situated in the range of 8-30 words in the media corpora, whereas the academic texts have sentences between 12-36 words long. From these ranges, we randomly selected 300 sentences, 100 from each source, as samples for the study.

From the 1σ of the Standard Deviation ranges, we randomly selected 300 sentences, 100 from each source, as samples for the study. It is remarkable that both in news and academic texts the range covers the 75% of the total population of sentences. In contrast, the medical media corpus presents a more dispersed sentence length distribution. Figure 2 shows the different distributions.

	Newspaper media	Medical Media	Academic prose
Corpus	UAM Treebank	OCU-Salud / Otro Médico	Harrison
Total N (of sentences in corpus)	1520	16812	149615
Overall mean (length)	15,12	19,82	24,73
Length range of sentences selected	**8-21**	**8-32**	**12-36**
N sentences in range (1σ of SD)	**1149**	**11749**	**112078**
Percentage of 1σ from total N	75,59	69,88	74,91

Table 2: Sentence length

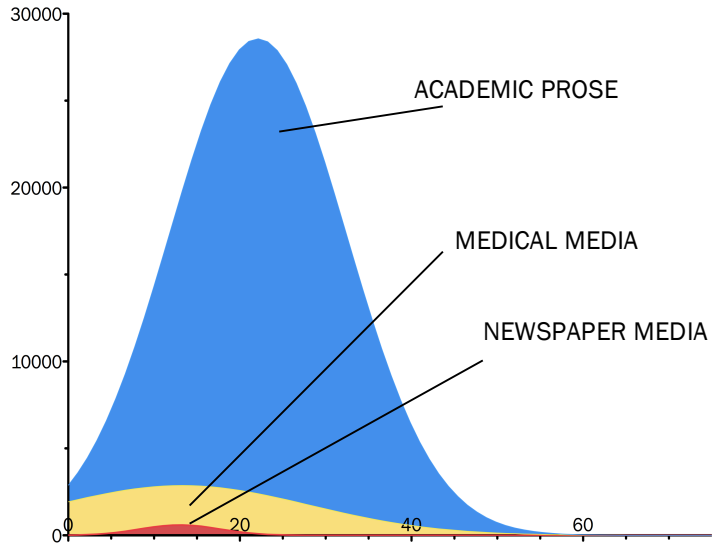

Figure 2: Sentence distribution comparison

Next, the 300 sentences selected for the study were tagged using *Grampal* morphological analyser (Example 1), and then syntactically parsed using our trained parser (Example 2 shows the *Lisp* format output, represented as a tree in Figure 3).

La/ART prueba/N de/PREP sensibilidad/N está/AUX diseñada/V para/PREP calcular/V la/ART sensibilidad/N de/PREP una/ART cepa/N clínica/ADJ bacteriana/ADJ

a/PREP determinado/ADJ antimicrobiano/N en/PREP circunstancias/N estandarizadas/ADJ ./PUNCT

Example 1: Example of an input sentence, tagged with *Grampal*.

```
(S
 (NPSUBJ (ART La) (N prueba)
      (PP (PREP de)
            (NP (N sensibilidad))))
 (VP
      (V (AUX está) (V diseñada))
      (PP (PREP para)
            (CL
                  (VP (V calcular)
                        (NPOBJ1 (ART la) (N sensibilidad)
                              (PP (PREP de)
                              (NP (ART una) (N cepa)
                                    (ADJP (ADJ clínica))
                                    (ADJP (ADJ bacteriana)))))))))
 (PUNCT .))
```

Example 2: Example of an output syntactically parsed sentence in Lisp format.

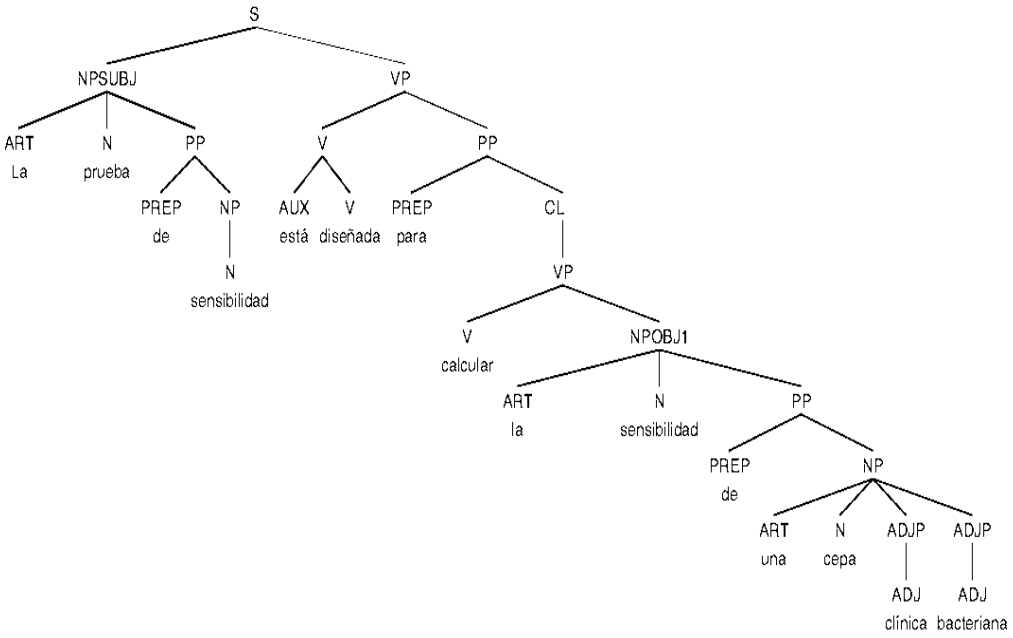

Figure 3: Tree representation of Example 2

During these processes, the syntactic tags have been simplified from the original tagset of the Treebank, as we believe a careful selected collection of tags and trained sentence will provide positive results (Rehbein, 2012). Also, the last step was a manual revision and correction of the results of the POS and syntactic tagging, before continuing with the statistical analysis. Figure 6 summarises the steps taken for the extraction of the syntactically analysed sample sentences that will be used for the study.

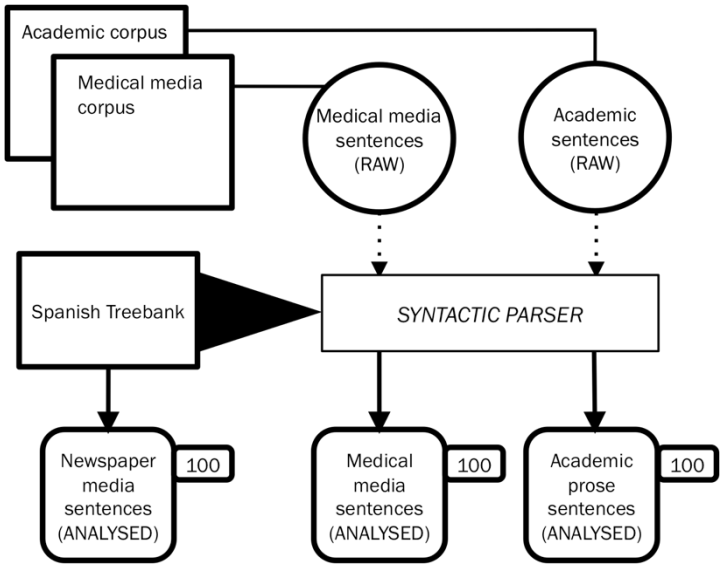

Figure 4: Extraction of sample sentences used for the study

An additional outcome of this study is the addition of the sample sentences to the training batch of the parser. Since they have been manually revised, they can be used for generating better grammar rules for the parser (*Active learning* process).

4 Results and discussion

The NPs studied are those that perform the main functions of the sentence: subject (NPSUBJ), object (direct and indirect) (NPOBJ), predicative NPs (NPPRED) and time expressions (NPTIME). The following pages will show the results of the study performed in the sample sentences: general frequencies of the NPs (Table 3) and NP complexity measures, length, height and sentential NPs (Figures 5, 6 and 7).

Although differences do not seem to be very wide, sentences in medical media appear to have an overall higher number of NPs. The most noticeable differences appear in the predicative NPs, with nearly 5 times more appearances in the medical domain. The frequencies of the TIME NPs are too low to perform a solid analysis, and will not be included in further calculations.

	Newspaper Media	Medical Media	Academic
NPs (total numbers)	174	223	211
SUBJ	**92**	**120**	**119**

OBJ	**76**	**85**	**75**
PRED	**3**	**15**	**16**
TIME	3	3	1

Table 3: NP frequencies from sample sentences

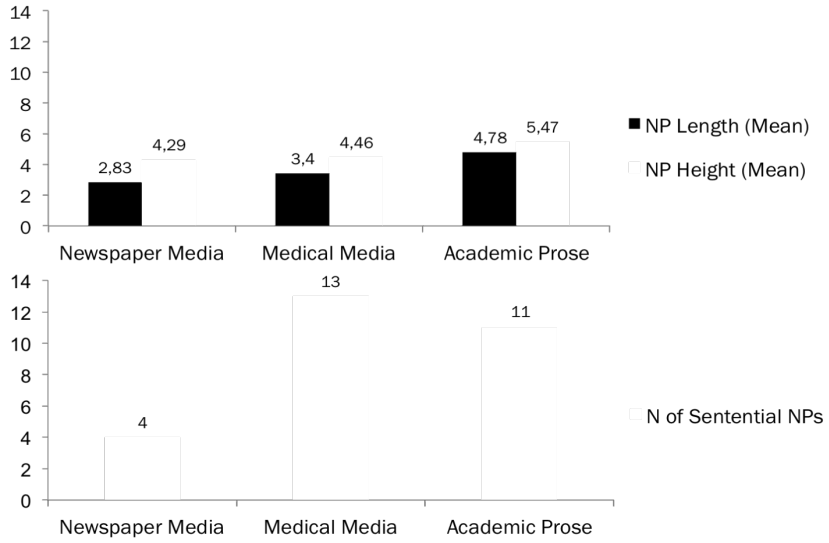

Figure 5: NP SUBJECT complexity results

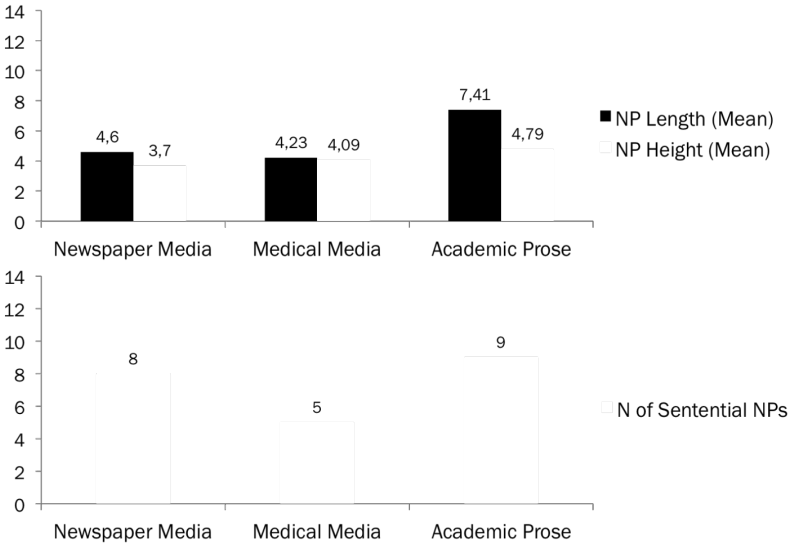

Figure 6: NP OBJECT complexity results

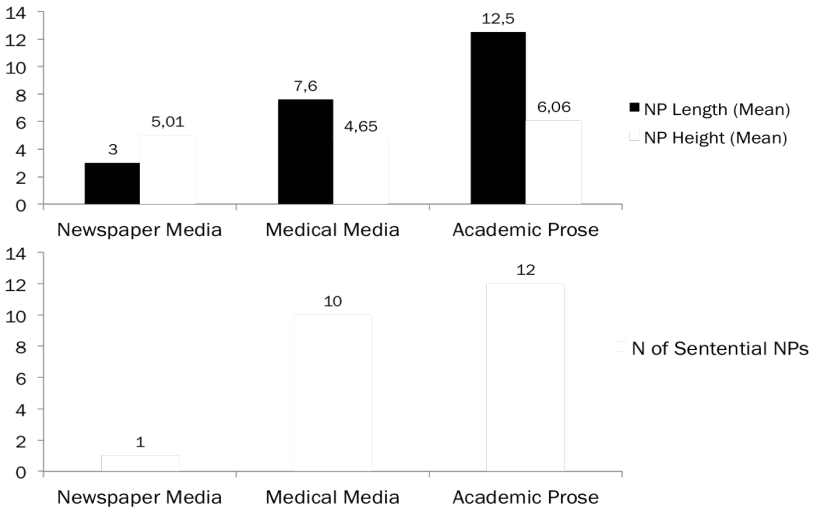

Figure 7: NP PREDICATIVE results

Even though the medical media sentences contain more NPs, in terms of their complexity we can draw three main conclusions:

Regarding NP length and height, the newspaper and medical texts have similar values. A series of additional t tests on the data confirm this assumption, as there is not a significant difference between both means (p values of 0,86 / 0,26 / 0,10 for subject, object and predicative lengths, respectively; values of 0,28 / 0,40 / 0,20 for their heights). The Academic prose texts, however, return higher numbers. Not only the overall mean values are higher, the p values also indicate a significant difference from the medical media sentences (0,004 / 0,0001 / 0,02 for length; 0,007 / 0,0001 / 0,09 for height, respectively). Also, it is worth mentioning the high frequencies and length of the predicative NPs in this domain, which suggests a high amount of copulativ9e sentences.

In terms of the number of sentential NPs, the numbers are too low to draw solid conclusions and inferential texts. The complexity of NPs in our texts appear to rely more on attaching phrases as modifiers of the main noun (adding length and height to the NP) than using VPs and clauses.

Therefore, we can classify our texts in two groups:

- Media texts (general newspapers and medical journals):
  - Similarity in sentence length (average of 15-19 words in total, 8-30 for samples' range)
  - Higher number of NPs in medical domain.
  - Similar NP complexity
  - Low number of sentential NPs

- Academic texts (medicine manual):
  - Sentences are longer than media texts (average of 24 words, 12-36 for samples' range)
  - Higher NP complexity than media texts
  - Significant high number of predicative NPs
  - Low number of sentential NPs

Confirming our hypotheses following Biber et al's claim (1999, 97), academic texts convey information in more space and more densely than media texts, which have less space available. Also,

due to the high number of predicative NPs and few sentential NPs, we could assume that the VPs in these are more simple and direct, semantic and syntactically speaking.

5 Conclusions

We believe this study of noun phrase complexity has been able to achieve several positive goals:

First of all, the syntactic parser appears to function correctly, as sentences from the same domain as the training set (media text) achieve similar trees, whereas the ones from the academic prose create bigger ones. This agrees with Biber et al's (1999: 97) research performed previously in the area and confirm our initial hypotheses, and will establish the beginning of a methodology that can be used more extensively in the future in studies such as text typology or studying the style of a literary author.

Secondly, even though we are aware the samples used for the study are only limited to 300 sentences, quality of the sampling precedes quantity, as the sentences have to be manually revised in order to be able to add them to the training set of the parser. However, since the sampling has been carefully prepared and inferential tests have been successfully applied, we assume the numbers to be valid.

Finally, we believe that the usage of automatic parsers is mandatory for similar studies made in this area, as they provide results impossible to achieve manually and faster and more precise. It is the duty of the linguist to make a previous study of the data in order to select meaningful samples, and make sure the output of the parsing is correct.

References

Berlage, E (2014). *Noun phrase complexity in English*. Cambridge: Cambridge University Press.

Biber, D., S. Johansson, G. Leech, S. Conrad, E. Finegan. (1999). *Longman Grammar of Spoken and Written English*. Edinburgh: Pearson Education Mimited.

Evert, S. and B. Krenn. (2005). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech & Language* 19(4), 450-466.

Klein, D. and C. D. Manning. (2003). Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430.

Moreno-Sandoval, A., López, S., Sánchez F., & Grishman, R. (2003). *Developing a syntactic annotation schema and tools for a Spanish treebank Treebanks: building and using parsed corpora*. Dordrecht: Kluwer.

Moreno Sandoval, A. & L. Campillos Llanos (2013) "Design and annotation of MultiMedica - a multilingual text corpus of the biomedical domain". In *Procedia - Social and Behavioral Sciences*, 95, pp. 33 - 39 (Actas seleccionadas del 5º Congreso Internacional de Lingüística de Corpus 2013, Universidad de Alicante, España. 14 - 16 de marzo del 2013). Berlin: Elsevier. ISSN: 1877-0428.

Rehbein, I., H. Hirschmann, A. Lüdeling and M. Reznicek. (2012). Better tags give better trees – or do they? *Linguistic Issues in Language Technology* 7 (10). 1-18.