Chapter 9

DEVELOPING A SYNTACTIC ANNOTATION SCHEME AND TOOLS FOR A SPANISH TREEBANK

Antonio Moreno, Susana López, Fernando Sánchez
Universidad Autónoma de Madrid, Spain
{sandoval;susana;fernando}@maria.lllf.uam.es

Ralph Grishman
New York University, USA
grishman@cs.nyu.edu

Abstract This chapter will describe our experience developing specifications and tools for building a Syntactically Annotated Corpus (SAC) for Spanish newspaper texts. The initial corpus consists of 1,500 sentences extracted from *El País Digital* and *Compra Maestra*, with a total of 22,695 words. The paper will address several of the relevant topics for any SAC project, namely methodology, data selection, annotation scheme, tools, and experiments.

Keywords: Spanish, Annotation Scheme, Tools.

1. INTRODUCTION

We are developing the specifications and tools for building a Syntactically Annotated Corpus (SAC) of Spanish newspaper texts. The initial phase of the project started in December 1997 and currently the corpus consists of 1,500 syntactically annotated sentences extracted from newspapers. The team consists of two linguists (Moreno and López), who are responsible for the selection and annotation of the sentences, as well as for the decisions regarding the format and encoding strategies; and of two computational linguists (Sánchez and Grishman) who have developed the tools both for annotation and for validation. The starting point was a series of previous documents on syntactic annotation,

concretely (Marcus et al., 1993; Bies et al., 1995; EAGLES, 1996; Skut et al., 1997). First of all, we wanted to develop our own specifications incorporating the relevant features for Spanish into the general mainstream of corpus annotation. We treat this topic in section 3. Secondly, we wanted to base our specifications on the experience gained in annotating real text. The first 500 sentences were taken as a sample of the complexity and variety that we could find in newspaper texts. We took the sentences from two different sources (see next section) and we were seeking variety in the phenomena. We subsequently produced three different versions of the specifications, reflecting our experience with those sentences. The specifications are open to new additions and changes; we could even incorporate significant modifications to the current guidelines. Finally, in this preliminary stage of the corpus, we also wanted to explore and develop some tools for helping human annotators with the coding and debugging tasks. This experience will be explained in 4.

2. DATA SELECTION

The current corpus consists of 1,500 sentences taken from:

- **El País Digital**, the newspaper on-line edition. Sentences 1-50, 101-150 and 201-1500. Each one is an isolated sentence, without context.
- **Organización Consumidores y Usuarios (OCU)** magazines. Sentences 51-100 and 151-200. The sentences are integrated in a context of several paragraphs.

The fifty-fifty division of the first 200 sentences is intended for testing annotation within a pure syntactic environment and annotation with a discourse perspective. The average sentence length is 15.13 words/sentence. The sample has been taken selectively from the different sections of the sources, reflecting different styles. The variety and complexity is more evident in the first 500 sentences than in the last ones, since in the former case the goal was to test and refine the annotation scheme, while in the latter the objective was to enlarge the corpus. For this reason, the last 1,000 sentences tend to be less complex, that is, there is a smaller number of subordinate and embedded clauses, ambiguities, etc., and also those sentences usually are shorter. The selection has been made according to the linguists' criteria, seeking variety, but this decision may have caused the corpus to be biased to certain types of constructions. In order to solve this problem, in future phases the sentences will be chosen randomly[1]. The sources were chosen based on availability and legal permit. The main purpose of these 1,500 sentences was to develop the specifications and the tools; in subsequent phases of the treebank, new sources will be added. In the same fashion, future plans are to include other types of texts.

3. ANNOTATION SCHEME

This section will deal with three points: criteria for the identification of the linguistic units, the actual notation used in the treebank, and some strategies followed in the application of the scheme.

3.1 Identification of the linguistic units

We will discuss briefly some of the most relevant linguistic issues for any SAC in any language. Firstly, we will address how to handle complex constituents. Then, we will show how null elements are analyzed in our scheme.

Asymmetric constituents. Our scheme distinguishes the orthographic string from the lexical unit. Problems arise when the relation between them is asymmetric:

- **Multiwords**: many orthographic words to one lexeme.
- **Amalgams**: one orthographic word to many lexemes.

Multiterm constituents (i.e. several words forming a single unit) are problematic for a linear phrase structure representation. We have developed a specific treatment for the following phenomena in Spanish:

- *Verb periphrases*, which are a combination of TENSED VERB + PARTICLE [2] (a PREP or a C) + UNTENSED VERB (see Figure 9.1).
- *Lexicalization* (see example below for "hombres de paja", *men of straw*).

```
(N ``<hombres de paja>'' ``hombre_de_paja'' MASC PL))
```

- *Locutions* (the Spanish term for multiword) that can be adverbs (eg. "en efecto", *effectively*) prepositions (eg. "antes de", *before*) or conjunctions (eg. "tan pronto como", *as soon as*).
- *Time expressions*, *idioms*, and other typical constructions in newspaper texts (see 3.3).

In an amalgam more than one constituent is represented in a unique orthographic word. There are two typical cases in Spanish: the so-called *portmanteau* words ("del" = de + el, "al" = a + el), and the post-clitics ("dármelo" = dar + me + lo). In both cases, we want to express the fact that several units are involved and that on the surface they appear as one word.

```
(S
  (NP SUBJ MASC SG P3
    (N ``<Manuel>'' ``Manuel'' PROPER))
  (VP TENSED PRES MODAL SG P3
    (V ``<tiene que ir>'' ``ir'' TENSED PRES MODAL SG P3
      (AUX ``tener_que'' TENSED PRES SG P3)
      (V ``ir'' UNTENSED INFINITE))
    (PP A LOCATIVE
      (PREP ``<al>'' ``a'')
      (NP
        (ART ``el'' DEF MASC SG)
        (N ``<dentista>'' ``dentista'' MASC SG)))
    (NP TIME
      (ART ``<el>'' ``el'' DEF MASC SG)
      (N ``<viernes>'' ``viernes'' MASC SG))))
```

Figure 9.1. Periphrasis: "Manuel tiene que ir al dentista." *Manuel has to go to the dentist.*

In all these cases, our strategy is the same: to use an element of the feature description for the surface string, and another for the more abstract interpretation. This is the sample for a portmanteau word:

```
(PP
  (PREP ``<del>'' ``de'')
  (NP
    (ART ``el'' DEF MASC SG)
    (N ``<libro>'' ``libro'' MASC SG)))
```

where ``<del>'' represents the actual string, and ``de'' and ``el'' are the lexemes. Clitics in Spanish can appear before (preclitics) or after (postclitics) the verb. Preclitics are always separate words (*se lo dio*), and postclitics are always joined to the verb (*dárselo*). Annotating preclitics is not problematic, since they are pronouns, and therefore NPs. For postclitics we could split the input string into parts: "dar" "se" "lo". But we want to show that both, verb and clitic(s), constitute a compound. We mark this using the following format (some non-relevant features for the example have been ommited):

```
(VP UNTENSED INFINITE
  (V ``<drselo>'' ``dar''...  #CLITIC ID-1
    (NP
      (P ``se'' PERS P3 SG DISCONTINUOUS REF-1))
    (NP OBJ1
      (P ``lo'' PERS P3 SG DISCONTINUOUS REF-1))))
```

Null elements. We will follow basically the Penn Treebank scheme for empty elements, but we will only annotate **null subjects** (*) and **ellipsed material** (*?*) such as VP, PP, CL, etc. in coordinations, and also required objects. We do not mark traces (*T* in the Penn Treebank). Null subjects can be found in two different constructions:

- As a pro-drop language, Spanish usually omits implicit subjects, which are nevertheless recoverable from the verb agreement (see Figure 9.2).
- In the infinitive, participle and gerund clauses, which normally are referred to as *raising* and *control sentences*.

```
(S
  (NP * SUBJ PL P1)
  (VP TENSED PRES IND PL P1
    (V ``<Regresamos>'' ``regresar'' ...  PAST IND PL P1)
    (ADVP TIME
      (ADV ``<ayer>'' ``ayer''))))
```

Figure 9.2. "Regresamos ayer." *We came back yesterday*

In both cases we will use the same format:

```
(NP * FUNCTION FEATURE1 FEATURE2 ...)
```

For null subjects in sentences with a tensed main verb, the agreement features (i.e. number and person) are encoded. For null subjects in untensed sentences, we need to specify a **reference index (REF)** in the empty NP, and an identity index (ID) in the co-referenced NP (see Figure 9.3). The co-referenced index is not always easy to assign. In those cases where there is not a clear candidate for the ID feature, then it is not specified. We do not annotate NP movement in passives, as the Penn Treebank does, nor do we mark any trace (for motivation, see 3.3).

On the other hand, we use the Penn Treebank tag, ***?***, for marking any ellipsed element other than the empty subjects. However, we do not follow the Penn Treebank strategy of annotating ellipsis in comparatives, but only in coordinations. In addition, we include the REF feature in the null element, and the ID feature in the co-referenced one, as we do with null subjects in untensed sentences.

3.2 Notation

The Penn Treebank style has been chosen as a model for the format. The main reason was that that format can be easily handled by Treebank-trained parsers such as the Apple Pie Parser developed by the Proteus project at NYU[3]. The only significant innovation with respect to the Penn Treebank scheme is the addition of features. Thus, the annotation scheme is a combination of typical category (POS and phrase) tag with feature values that specify the syntactic information for each (terminal and non-terminal) element. We annotate different layers of information: syntactic categories (i.e., parts-of-speech such as NOUNS, ADJECTIVES, ...), syntactic functions (e.g. SUBJ, OBJ1 (direct object), OBJ2 (indirect object)...), morpho-syntactic features (i.e. number, gender, tense, etc.) and some semantic features (HUMAN, TIME, etc.). The actual

```
(S
  (NP SUBJ ID-1 SG P3
    (N ``<Juan>'' ``Juan'' PROPER SG P3))
  (VP TENSED PRES IND SG P3
    (V ``<quiere>'' ``querer'' TENSED PRES IND SG P3)
    (CL INFINITIVE OBJ1
      (NP * SUBJ REF-1)
      (VP UNTENSED INFINITE
        (V ``<leer>'' ``leer'' UNTENSED INFINITE)
        (NP OBJ1
          (ART ``<un>'' ``uno'' INDEF MASC SG)
          (N ``<libro>'' ``libro'' MASC SG)
          (PP DE
            (PREP ``<de>'' ``de'')
              (NP
                (N ``<Chejov>'' ``Chejov'' PROPER)))))))
```

Figure 9.3. "Juan quiere leer un libro de Chejov." *John wants to read a book by Chekhov.*

string is presented between ``<...>'', and its lexeme is between ``...''. All this information about each category is expressed in a unified format.[4] Lexical categories are represented as:

```
(CAT ``<string>'' ``lexeme'' FEATURE1 FEATURE2 ...  )
```

For the actual catalogue of allowed categories and features, we have developed an "Inventory of units", ordered by types. This typed inventory is described in (Moreno et al., 1999)[5]. In order to label non-terminal constituents, we follow a simple, vertical, indented format:

```
(CAT1 ...
  (CAT2 ...)
  (  ...   )
  (CATn ...))
```

3.3 Application strategies

In order to apply the annotation scheme, we have made some decisions:

- **Eliminate redundancy as much as possible**: To avoid duplication, the information is usually specified only once in the relevant layer. For instance, internal agreement features between a noun and its modifiers are not percolated up to the NP if they are not be used in further stages (for instance, the NP inside a PP). This strategy leads to **underspecification**, since not all syntactic details are included[6].

- **When there is an uncertainty about how to assign a given feature, prefer not to do it**. This attitude produces ambivalence, but we prefer that, better than making arbitrary decisions.

- **Reflect the surface syntax**: We are very cautious about empty categories, and how to annotate them. Only two kinds of null elements are labeled: null subjects and ellipsed material.

Some examples. An annotation scheme for a particular language must include some syntactic constructions that are unique for that language or for a cognate. However, those constructions are very frequent in corpora, as in the case of Spanish the so-called "se"-constructions. We will describe them here as an example of the annotation scheme coverage.

"Se"-constructions. One of the major problematic issues in Spanish is the one that involves the word "se". The actual categorial status of this word is rather problematic. Here we are going to treat it as a pronoun in the majority of cases, except when it functions as an intransitive marker or as an impersonal marker.

We have established five different annotations involving "se-constructions" depending on the kind of "se" that appears. These five annotations are the following ones:

- **"Se" that corresponds to "le"**: In this case, "se" is a pronoun that substitutes for "le" due to phonetic reasons. Example:"Se lo dio" (*He/she gave it to him/her*).

- **"Se" in a reflexive or reciprocal construction** This "se" is also treated as a pronoun, and as such it changes according to person and number ("me" SG P1, "te" SG P2, "se" SG P3, "nos" PL P1, "os" PL P2, "se" PL P3). When this pronoun appears, the subject and the patient of the action are the same. The difference between reflexive and reciprocal is a semantic one, so we are going to consider both equally. Example: "Yo me lavo" (*I wash myself*).

- **Pronominal or intrinsic "se"** This type of "se" appears in constructions where "se" is attached to the verb, that is, "se" is part of the verb; both appear together in the lexicon, for example "dormirse", "ponerse", etc. This type of "se" is going to be annotated as a compound word; it can precede or follow the verb, but we will treat it as a part of the verb either way. When the "se" precedes the verb, we will treat it as a pronoun with its own features plus the discontinuous and reference marks, to maintain its autonomy as a pronoun that forms part of the verb. If it follows the verb we will leave it that way, adding the feature #CLITIC.

```
(CL INFINITIVE
  (NP * SUBJ REF-1)
  (VP UNTENSED INFINITE
    (V ``<hacerse>'' ``hacerse'' UNTENSED #CLITIC)
    (PP CON OBL
      (PREP ``<con>'' ``con'')
      (NP
        (ART ``<el>'' ``el'' DEF MASC SG)
        (N ``<control>'' ``control'' MASC SG)))))
```

Figure 9.4. Pronominal "se": "... hacerse con el control..." *...to get control...*

```
(S
  (NP SUBJ MASC SG P3
    (ART ``<El>'' ``el'' DEF MASC SG)
    (N ``<libro>'' ``libro'' MASC SG))
  (VP TENSED PAST IND INTRANSITIVE SG P3
    (SE-MARK ``<se>'' ``se'' INTRANSITIVE)
    (V ``<rompi>'' ``romper'' TENSED PAST IND SG P3)))
```

Figure 9.5. Intransitive marker: "El libro se rompió"

.

```
(S IMPERSONAL
  (VP TENSED PRES IND SG P3
    (SE-MARK ``<se>'' ``se'' IMPERSONAL))
    (V ``<vende>'' ``vender'' TENSED PRES IND SG P3)
    (NP OBJ1
      (N ``<piso>'' ``piso'' MASC SG))))
```

Figure 9.6. Impersonal "se": "Se vende piso"

Example: "Tratan de hacerse con el control de la calle" (*They try to get control of the streets*). See Figure 9.4.

- **"Se" as an "intransitive-marker"** Here, what we have are three different constructions (that have been called "passive se", "inaccusative se" and "middle se") which have in common the loss of an argument. They are distinguishable by semantic means, but, as we are only concerned with syntactic criteria, they are to be annotated all the same. In this annotation, "se" is not treated as a pronoun but as a mark used to make intransitive constructions. Example: "El libro se rompió" (Figure 9.5).

- **Impersonal "se"** These constructions have the characteristic that there is no syntactic subject. "Se" is a pronoun that marks the impersonality. Example: "Se vende piso", *Apartment for sale*, literally "someone sells an apartment". (See Figure 9.6). As in the previous case, we use the category label SE-MARK, but with the feature IMPERSONAL instead of INTRANSITIVE.

Other relevant linguistic topics in corpora. Another frequent phenomena that usually appear in corpora of many languages are idiomatic expressions, dates and hours, math operators, scores and measures, foreign words, etc. In contrast with their apparent lack of interest for theoretical linguistics, they are important in corpus annotation. In general, all these structures are handled as multiwords (see 3.1) since they are considered to be a "block".

4. TOOLS

For the construction of the tree bank, we make use of a combination of tools and resources either developed by the group or obtained from the public domain. These tools and resources can be divided into:

1 Annotation tools:

- **A morphosyntactic annotation system**. This system uses MTSEG, MULTEXT segmenter for texts[7]. This segmenter has been enhanced in a number of ways, including the addition of new resources, the identification of new text elements and better heuristics for sentence boundary identification. As a lexical resource, we use a 50,000 lemma lexicon accessed by a generating inflectional morphological component. Lexical analysis is reduced then to lexical look-up in the full form lexicon produced by the generator. This look-up strategy is complemented by a set of modules that handle complex phenomena (post-clitic pronouns, appreciative morphology, derivation...). Disambiguation is performed by means of a reductionist grammar. We are currently reusing the grammar developed within the *Constraint Grammar* formalism (Karlsson, 1995) as part of Ph.D. work described in (Fernando, 1997). This grammar (actually, the knowledge encoded in it) has been rewritten so as to be interpreted by a Perl program sequentially applying constraints found in the grammar. Constraints are distributed in five different grammars depending on reliability and complexity (locality v. long distance, mainly) of the rules. The approach to disambiguation favours recall rather than precision, so some disambiguation work is still left to the human posteditor, but the system rarely promotes an incorrect analysis.
- **A chunker** recognizing major phrases within the sentence. The grammar used by the chunker includes NPs, ADJPs, VPs, ADVPs, and PPs. These phrases can be recursively identified, so nesting of phrases is allowed up to a given maximum level. The chunker is applied after postediting the output from the annotation system.

Table 9.1. Error in feature assignment

	All categories	*ADJP*	*ADVP*	*NP*	*PP*	*VP*
Total number of cases	6364	592	262	2933	1503	1074
Total number of errors	672	51	70	457	35	59
Percentage of errors	10.5 %	8.6 %	26.7 %	15.6 %	2.3 %	5.4 %

Table 9.2. Types of errors

Types of errors	*All categories*	*ADJP*	*ADVP*	*NP*	*PP*	*VP*
Total	672	51	70	457	35	59
Missing features	422	22	29	333	25	13
Incorrect features	226	29	40	105	10	42
Unnecessary features	24	0	1	19	0	4

2 Debugging tools:

- **A graphical tree-drawer**. This tool, a public domain program called CLIG (Computational Linguistics Interactive Grapher[8]) allows the definition of clickable tree nodes which favors the drawing of nodes with just basic information so as to allow for rapid inspection of constituent structure, and then, by clicking on the relevant node(s) feature values can be checked. A program is used to automatically produce the forest of CLIG objects from treebank notation.
- **A feature checker** that controls the assignment of proper features for each category, described in the following section.
- **A phrase structure rule generator**, which is used to detect possible incorrect annotations. This is also described in the following section.

5. DEBUGGING AND ERROR STATISTICS

Using a small Perl program, we have managed to check and revise the features of the first 500 annotated sentences for an evaluation of the feature assignment. The data we have obtained are shown in Table 9.1[9]. We have also differentiated the types of errors made within each phrase (see Table 9.2). We can infer from this data that the most common errors are lack of features

and replacement of features (that is, wrong features instead of correct features). This may be due to the various changes we have made during the elaboration of specifications for the project. What we learned from this is which phrases are the most prone to error with respect to our feature annotation scheme. Those are NPs and ADVPs, and the rest have low rates of error. Also, we have learned that there are some contradictory features such as PERFECT and IMPERFECT in the same verb head, which result from adding the features of an auxiliary verb and the features of the main verb, and percolating them up to the VP level. This evaluation allowed us to improve our performance in the next 500 sentences, since we concentrated on our previous errors. We have not conducted a proper evaluation for those new sentences yet, but our estimation is that the current percentage of error in assigning features is below 5%. This figure will be significantly reduced when we make use of a *feature checker*, based in the feature specifications. On the other hand, we use a phrase structure (PS) rule generator from the annotated sentences in order to detect "strange" combinations of constituents (for example, a clause made up of "ADVP V PP" instead of "C VP ADVP V PP"). This tool provides a different point of view to the coder, since it presents the results of the annotation. The PS rules generator has been useful for detecting some inconsistencies. Finally, we also use a graphical tree viewer called CLIG, which allows inspection of the annotated sentence. The tree viewer not only shows the branches and the categories, but also the features for each node. Like the rule generator, the tree viewer provides another way of approaching the corpus; both useful for the human annotator and the user.

We are currently involved in a complete evaluation of the 1,500 sentences. Some of the results suggest that changes in the guidelines will be needed (see next section).

6. CURRENT STATE AND FUTURE DEVELOPMENT

In this chapter, we have presented some of the basic problems that researchers encounter when developing an annotation scheme for a Spanish treebank. We have offered some possible solutions to these problems, but we have also detected some limitations in our approach. In particular, the scheme proposed is probably excessively fine-grained from a linguistic point of view. Producing a reasonable number of trees (over 10,000) will be an expensive task in terms of time and research staff. The advantages of our linguistically oriented approach (mainly the possibility of inferring a feature-based grammar and extracting syntactic structures by descriptive or theoretical linguistic principles) can only be envisaged as long-term goals. However, with the 1,500 annotated sentences we managed to run an experiment: to use the treebank to train a statistical parser, the Apple Pie Parser (Sekine, 1998). The experiment

is reported in (Moreno et al., 2000). A stochastic context-free grammar was derived from 1,460 sentences, setting aside forty sentences for testing. The evaluation results obtained were 73.6% in recall, and 74.1% in precision. The most relevant conclusion we extracted was that restricting ourselves to a simple context-free skeleton instead of using the richer feature structures provided by the tree bank did not seem to be a major limitation on the performance of the trained parser. This fact, along with the difficulties mentioned above, suggests that it could be interesting to try a new approach. In particular, we are planning to develop two layers of annotation: one with only categorial and lexical information[10], and the other with the rich feature annotation. The idea is to test which type of information is more relevant for rule induction in Spanish. At the same time, concentrating on categorial/lexical information can speed up the process of tree annotation. In its current state, the treebank is not publicly available, although the annotation guidelines can be obtained from the project web page.

Recently López and Sánchez have left the project. Manuel Alcántara has conducted a complete revision of 1,500 sentences, and added a hundred sentences more to the treebank. Fernando Ares has translated the 1,600 parsed sentences into an XML format.

We are starting a new phase with new tools: for morphosyntactic annotation we are using GRAMPAL (Moreno and Goñi 1995) reimplemented in a Perl program by J.M. Guirao. Since the format has been moved to XML, the feature and POS validation is now performed by a XML parser. Moreno and Ares have defined a DTD for such a task, introducing some modifications in the annotation scheme.

Plans are to put the treebank in the public domain when the corpus reaches an acceptable size.

Acknowledgments

The research of Susana López has been supported by New York University. We want to thank El País Digital and Compra Maestra for letting us use their texts. We want to thank Lingsoft and especially Prof. Fred Karlsson for their permission to use the CG package during previous phases of this and other projects within Laboratorio de Lingüística Informática at UAM. Finally, we want to thank Karsten Konrad for putting his CLIG program in the public domain.

Notes

1. Another question related to this is whether 1,500 sentences are sufficient to consolidate an annotation scheme.

2. The particle does not occur in every case, like in *estar* + GERUND.

3. Future plans are to migrate to a standard like SGML/XML, as has been recently proposed for syntactically annotated corpora (Mengel et Lezius, 2000; Ide et al., 2000)

4. For free-order languages like Spanish, it could be appropiate to use two separate syntactic representations, one for the categorial and phrasal information, and the other for functional information. We have not chosen that representation because of simplicity and economy in the annotation process.

5. This is an 88-page report, for internal use. The specifications are available at the project WWW page: `http://www.lllf.uam.es/~sandoval/UAMTreebank.html`

6. However, after the experience with the 1,500 sentences, we are considering exploring the opposite approach: to promote redundancy, that is, to percolate features. The idea is to test which approach renders better results in training a parser.

7. This and other products from the MULTEXT project may be found in `http://www.lpl.univ-aix.fr/projects/multext/`.

8. CLIG is a grapher for visualizing linguistic data structures developed by Karsten Konrad in the Department of Computational Linguistics in Saarbrücken, Germany. CLIG is free for scientific purposes. The home page is `http://www.ags.uni-sb.de/~konrad/clig.html`.

9. The categories that have been considered are ADJPs, NPs, ADVPs, VPs, and, PPs. For this evaluation, we have not revised the head categories (e.g. N, Prep, V, etc.), nor major categories (e.g. S, CL).

10. Maybe incorporating a very selective feature information, such as agreement features

References

Bies, A., Ferguson, M., Katz, K., Macintyre, R. (1995). *Bracketing Guidelines for Treebank II Style Penn Treebank Project*.

Brants, T., Skut, S., Uszkoreit, H. (2003). "Syntactic annotation of a German newspaper corpus", in this volume.

EAGLES (1996): *Preliminary Recommendations for the Syntactic Annotation of Corpora*.

Ide, N., Bonhomme, P., Romary, L. (2000) "XCES: An XML-based Encoding Standard for Linguistic Corpora", *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*, Athens, p. 831-835.

Karlsson, F. and Voutilainen, A. and Heikkilä, J. Anttila, A. (1995). *Constraint Grammar: a Language-Independent System for Parsing Unrestricted Text*. Berlin, Mouton de Gruyter.

Marcus, M, Santorini, B, Marcinkiewicz, M. A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank, *Computational Linguistics*, 19, 2, p. 313-330.

Mengel, A., Lezius, W. (2000). "An XML-based representation format for syntactically annotated corpora", in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*, Athens, p. 121-126.

Moreno, A., López, S., Sánchez, F. (1999). *Spanish Tree Bank: Specifications. Version 4. 30 April 1999*. Internal document, Laboratorio de Lingüística Informática, UAM.

Moreno, A., Grishman, R., Lpez, S., Sánchez, F., Sekine, S. (2000). A Treebank of Spanish and its Application to Parsing, *Proceedings of the Second*

International Conference on Language Resources and Evaluation (LREC), Athens, p. 107-111.

Moreno, A. Goñi, J.M. (1995). GRAMPAL: A morphological model and processor for Spanish implements in Prolog, *Proceedings of the Joint Conference on Declarative Programming* (GULP-PRODE'95), Marina de Vietri, Italy.

Sánchez, F. (1997). *Análisis morfosintáctico y desambiguación en castellano*. Ph.D. Dissertation, Department of Linguistics, Universidad Autnoma de Madrid.

Sánchez, F. Ramírez, Declerck, Th. (1999). Integrated set of tools for robust text processing. *Proceedings of the VEXTAL Conference*, Venice.

Sekine, S. (1998). *Corpus-based Parsing and Sublanguage Studies*. Ph.D. Dissertation, Department of Computer Science, New York University.

Skut, W., Krenn, B., Brants, T., Uszkoreit, H. (1997). An Annotation Scheme for Free Word Order Languages, *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP)*, Washington, D.C.

Appendix: Sample of trees