Syntactic, semantic and pragmatic annotation in Cast3LB

Borja Navarro*, Montserrat Civit**, Mª Antonia Martí**,
Raquel Marcos*, Belén Fernández*.

*Departamento de Lenguajes y Sistemas Informáticos.
Universidad de Alicante.
Alicante, Spain
{borja,rmarcos,bfernan}@dlsi.ua.es

**CLIC, Centre de Llenguatge i Computació.
Parc Científic de Barcelona
Barcelona, Spain
civit@clic.fil.ub.es
amarti@fil.ub.es

Abstract

The objective of this paper is to present the linguistic annotation of the Cast3LB project, the Spanish part of the general project 3LB[1]. The objective of this general project is to build three linguistic annotated corpora: one for Catalan (called "Cat3LB"), one for Basque (called "Eus3LB") and one for Spanish (called "Cast3LB"). Both the Catalan and the Spanish corpora include 100,000 words each, and the Basque corpus 50,000. The aim is to annotate all these words with linguistic information at three levels: syntactic, semantic and pragmatic.
In this paper we present the proposal for the annotation of each linguistic level in the Spanish corpus (Cast3LB). At the syntactic level we mark constituents and functions; at the semantic level, the unambiguous senses of the words –based on EuroWordNet—, and finally, at the pragmatic level, the coreference between some noun phrases, pronouns and elliptical elements. We describe the tools and formats used in our annotation framework and provide some samples of the results.

1 Introduction

The construction of a linguistically annotated corpus is not an easy task, but it is essential for the development of the research in Natural Language Processing (NLP) and Computational Linguistics (CL). From a computational point of view, for example, a linguistically annotated corpus provides correctly analyzed samples of real language, which are necessary to train or evaluate computational systems. Linguistically speaking, the corpus is an essential source of information for the study of languages, and it is the basis for the training of empirical methods.
Up to now, only some corpora (mainly built for English) have been annotated at the syntactic level (Treebanks). For example, one of the best known treebanks is the PennTreeBank (Marcus *et al.* 1993), with 7 million words annotated at the morpho-syntactic level. In the last few years, however, treebanks for other languages have been created, as, for example, French (Abeille *et al.* 2002), Italian (Montemagni *et al.* 2002), Polish (Marciniak *et al.* 2002), German (Brant *et al.* 2002), Bulgarian (Simov *et al.* 2002), etc.
As for Spanish, the Universidad Autónoma de Madrid has developed a Spanish treebank with 1500 sentences (Moreno *et al.* 2002), although this resource is not available to the general public. At the University of Santiago de Compostela, a syntactic data base of Spanish has been built (Rojo 1990), which allows to make linguistical queries through a web navigator. Moreover, the Spanish Royal Academy (RAE) has introduced the corpus CREA in Internet[2], that can be consulted by web navigators too.
However, these Spanish corpora or treebanks do not satisfy our interest because data are not accessible for NLP purposes and they only contain syntactic information, without any reference to other annotations, like semantic and pragmatic ones.
We want to remark that only few corpora had been annotated with semantic or pragmatic information. Two of the most important corpora with semantic annotation are the SemCor corpus (Miller *et al.*

[1] Project funded by the Spanish government: PROFIT Program (FIT-150500-2002-244). Groups involved in the 3LB project are: CLiC, Centre de Llenguatge i Computació (Universitat de Barcelona); TALP Research Center (Universitat Politècnica de Catalunya); Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Alicante; Departamentode Lenguajes y Sistemas Informáticos de la Universidad de Valencia and IXA Group, University of the Basque Country, Computer Science Faculty.
[2] http://corpus.rae.es/creanet.html

1993), and the DSO Corpus of Sense-Tagged English (Ng and Lee 1996). The language of both corpora is English, and they use WordNet as a lexical resource for the sense annotation.
Regarding pragmatic annotation, the two most largest corpora with anaphoric and coreferential annotation are the one developed at the University of Wolverhampton (Mitkov *et al.* 2000) for English –that contains 60 000 words—, and the one developed at the University of Stendahl, Grenoble and Xerox Research Centre Europe (Tutin *et al.* 2000) for French. Their objective is to annotate one million words. We will discuss later with more details the semantic and coreferential tagged corpora.
At the moment, there is not a large corpus with semantic or coreferential annotation in Spanish. Regarding the semantic annotation, only the SENSEVAL forum[3] has developed a brief corpus with some semantically annotated Spanish words (18 nouns, 13 verbs and 9 adjectives). As for as the anaphoric or coreferential annotation, some brief corpora have been annotated in Spanish like, for example, a fragment of the LexEsp (Sebastián *et al.* 2000) developed at the University of Alicante (Palomar *et al.* 2001). In any case, there is not corpus for Spanish with such a large semantic or coreferential annotations as there is for English or French. Moreover, there is not one single corpus in Spanish which includes syntactic, semantic and pragmatic annotation, as the one we are developing now.
The project 3LB is at the moment in process of development. We have define the proposal of annotation for each linguistic level – that we present in this paper – and, currently, we are annotating at the syntactic level. The semantic annotation is going to begin early. The finally step is the coreferential annotation, that will begin when the syntactic annotation achieve more development. In principle, the project lasts until the end of 2003.
After this introduction, we present the Cast3LB annotation process. In the next section, we explain the annotation we propose for each one of the linguistic levels. Finally, we include some samples of the corpus and the conclusions.

2 CAST3LB corpus

For Spanish, the selected corpus is a part of the CLIC-TALP corpus, which is made up of 100.000 words from LexEsp (Sebastián *et al.* 2000) plus 25.000 words coming from the EFE Spanish Corpus, given by the *Agencia EFE* (the official news agency) for research purposes. The EFE corpus is comparable to the other corpora in the project (Catalan and Basque).
We have selected this corpus because it contains a large variety of Spanish texts (newspapers, novels, scientific papers, and so on), both from Spain and South-America, so it is a good representation of the present state of the Spanish language, moreover the automatic morphological annotation of this corpus has been manually checked (Civit 2003).

2.2 An overview of the annotation process in Cast3LB

The spirit of the annotation scheme is to build a flexible system eable to be transported to different romance languages and to new cases that might appear, but with consistency at all levels of annotation and in respect to annotation data.
At the syntactic level we follow the constituency annotation scheme. The main principles of syntactic annotation are the following (Civit and Martí 2002, Civit *et al.* 2003): firstly, only the explicit elements are annotated (except for elliptical subjects); secondly, we do not alter the surface word order of the elements; thirdly, we do not follow any theoretical framework; fourthly, we do not take into account the verbal phrase, rather, the main constituents of the sentence become the daughters of the root node; finally, this syntactic information is enriched by the functional information of the main phrases, but we have not taken into account the possibility of double functions.
At the semantic level, we annotate the sense of the nouns, verbs and some adjectives. We assign the specific sense (or senses) of each one by means of the EuroWordNet offset number (Vossen 1999). Also, due to the fact that some words are not available in EuroWordNet or do not have the suitable sense, we have created two new tags to mark this circumstance.
At the pragmatic level, we mark the coreference of nominal phrases and some elliptical elements. The coreference expressions taken into account are personal pronouns, clitics, elliptical subjects and some elliptical adjectives. The definite descriptions are not marked. The possible antecedents considered are the nominal phrases or other coreferential expressions.

[3] http://www.senseval.org/

2.3 Formats of encoding and tools

In Figure 1 we show the relationship between the three annotation levels and the formats of each one of them. The three levels are related through the XML format. At present, many annotated corpora use XML format as a mark-up language (see, for example, Hinrichs and Simov (eds.)), because it is a stable standard that facilitates the relationship between different levels of annotation and allows data to be easily transported; besides, it is suitable for the exploitation of the corpus in the web.
However, in the syntactic annotation we use the PennTreeBank format, because it is the standard most common used in encoding syntactic information. To solve the standardisation of the format, a XML translator has been developed, which not only translates the PenTreebank format of the syntactic level to the general XLM format, but it can translate the XML format of whole corpus into Peentreebank format (Fig. 1). This solution allows us to take advantage of the two standards.

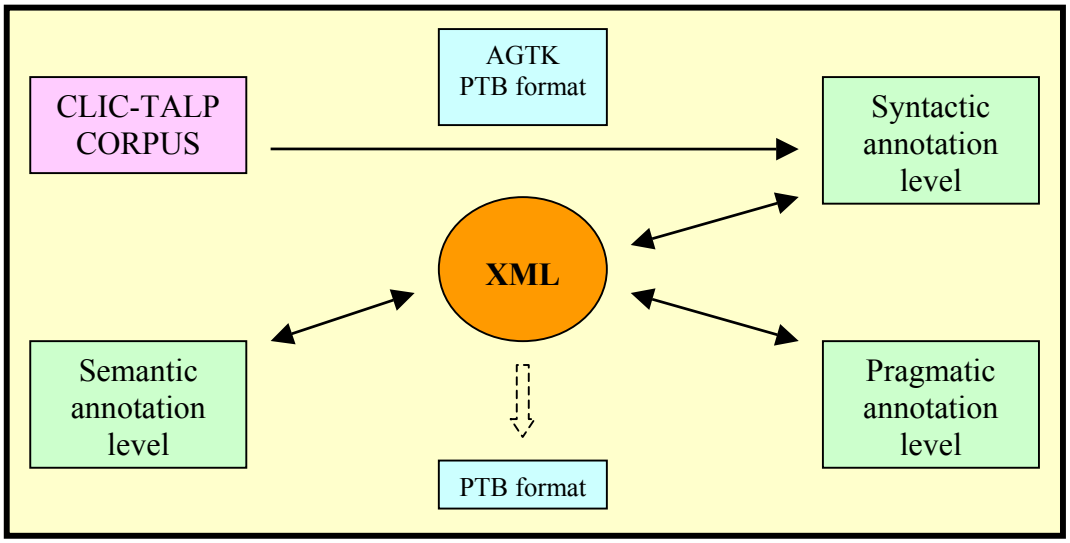

Figure 1. Formats

From a general point of view, we follow a semiautomatic approach in the annotation process. At each level of annotation, an automatic system –based on NLP techniques—analyzes the corpus and proposes the possible tags. The annotator, provided with an editor, checks the proposal and corrects any mistakes. The tools we use at each level are the following:

- Syntactic level:
  - We use the syntactic chunker TACAT (Atserias *et al.* 1998), that locates the basic phrases of the sentences (NP, PP, AjP, AvP), but does not propose a possible syntactic tree.
  - The AGTK toolkit: a visual editor for syntactic annotation, developed at Linguistic Data Consortium. This tool allows the annotator to establish the syntactic tree of each sentence, to correct possible mistakes of the chunker and to introduce new tags if it is necessary (as, for example, the syntactic function).

- Semantic level:
  - At this level, we use EuroWordNet (Vossen 1999) as a lexical database. The sense of each word is marked with the sense number of the Spanish WordNet (Alonge *et al.* 1998).
  - To look for the sense number, a WordNet browser integrated in AGTK has been developed. This browser displays the different senses of any given word, and the annotator selects the correct one (or ones).

- Pragmatic level:
  - At this level we use an anaphora resolution system based on Palomar *et al.* 2001, that propose possible anaphoric elements in the corpus and a list of their possible antecedents.
  - With this, through a newly developed web editor, the human tagger checks whether the anaphoric element is correct and selects its antecedent.

Each level is annotated independently. However, the anaphora resolution system takes advantage of the syntactic information annotated in the first step (syntactic annotation) in order to propose possible antecedents. The syntactic chunker and the WordNet browser use the Part of Speech information of the CLIC-TALP corpus.

3 General Proposal of Annotation

3.1 Syntactic annotation

The syntactic annotation process follows two main phases and each phase is organized into several steps. In the first phase, all syntactic constituents, the main syntactic functions and some elliptical elements are labelled. The main principles for the syntactic annotation process are described below.

Implicit versus explicit information. Spanish is a pro-drop language, so the subject is usually omitted. In the Cast3LB, only elliptical subjects are added for more details. Other elliptical data will be treated in the second phase, when anaphoric and coreferential phenomena are involved.

Constituency versus dependency annotation. There is an open discussion about the annotation scheme to be assumed when building a treebank. On the one hand, some papers claim that dependency annotation is more suitable for free word order languages (Boguslavsky *et al.* 2002, Brants *et al.* 2001, Oflazer *et al.* 2001), while others make their choice on the basis of the application required (Rambow *et al.* 2002). Finally, in some cases, the annotation system follows the linguistic tradition (Bemova *et al.* 1999).
On the other hand, constituency is usually employed to annotate languages like English in which there is a fixed constituent order. Moreover, in this case, there is an almost exact matching between constituents and functions, that is, the position of a given constituent corresponds to one concrete syntactic function. For instance, in canonical declarative sentences, any noun phrase immediately preceding a verb is usually the subject.
Spanish is a free constituent order language, although the word order cannot be altered within a constituent. Three ways can be used to say "John came this morning": *Juan ha venido esta mañana; Esta mañana Juan ha venido* and *Esta mañana ha venido Juan*, which are not exactly equivalent in their meaning. The focused element varies in each sentence, thus, the pragmatic meaning is different. Furthermore, there are two noun phrases and both can precede the verb, so that it is impossible to know which the subject is (unless semantic information is available). At this stage, constituent annotation is convenient for Spanish as a previous step for the annotation of syntactic functions.
There are other reasons for this choice such as the software previously available. It is a common idea to take the maximun advantage of the tools developed so far. As the input of the treebank is the result of a chunker, then the most recommended solution is the phrase structure annotation.

Maintaining the surface word order. According to the previous points, no word order alterations are made during the annotation process. The strategy now is quite conservative. However, we are not ruling out this possibility in further developments of the treebank. Generally speaking, sentence structure can be easily represented. The problem appears with discontinuous elements: comparative clauses, in which the two elements usually occur separately or with movement phenomena in interrogative sentences, especially when a wh-word comes from an embedded clause like *qué te gustaría ser* ('what would you like to be'). In comparative structures, the clause is adjoined to the node containing the adverb of comparison. In the second case, the constituent at the beginning of the sentence will be given a special tag when annotating syntactic functions.

Being theory-neutral. Linguistic theories give solutions to some specific problems but lack coverage, that is, they work with a hypothetical model of language that does not face problems arising from corpora. Besides, theory deals with very specific (even rare) phenomena which hardly ever appear in corpus (see Sampson 1987). In the literature about treebanks, two positions about theory foundations arise: treebanks which are theoretically founded and treebanks that are theory-neutral. Among treebanks that are annotated according to one theory, two cases should be mentioned: treebanks annotated following the GB framework, like the PennTreeBank, and those annotated according to the HPSG theory. The PennTreeBank (Marcus *et al.* 1993 and Taylor *et al.* 2001) is annotated with the principles of the X-bar theory, even though there is not a full application of all the theoretical issues. Some difficulties arise, for instance, with the need of distinguishing between arguments and adjuncts,

and with the PP-attachment, as stated in Taylor *et al.* 2001 and Marcus *et al.* 1994. Marciniak *et al.* 2001 and Simov *et al.* 2002 follow the HPSG theory. The former justifies the choice on the premise that it facilitates the evaluation of an HPSG grammar; provides a uniform way to represent different types of linguistic information; and is widely used in computational linguistics. The latter claims that HPSG allows to simultaneously represent constituents as well as dependency relationships; that this theory permits a consistent description of linguistic facts; that it enables translation to other formalisms; and, finally, that it can be used to support annotators' work. It is worth noticing that these treebanks are constituted by chosen sentences instead of by large texts more or less randomly selected. Therefore, even if the number of sentences is high, they do not deal with what is largely understood by real text, that is, text reflecting any kind of linguistic phenomena.
As for annotation systems which do not follow any specific theory, it should be said (as in Abeillé *et al.* 2001) that this option permits to adopt solutions equally profitable for linguists, computer scientists, psycholinguists, etc. Following this proposal, we do not wish an application of one or another linguistic theory, but to fix a standard of constituency and functional annotation, neutral enough to be used in any research on Spanish and easy to translate into other formalisms.
We think that more neutral the annotation scheme is, more suitable it will be for NLP purposes and for linguistic research. In fact, nowadays there is no theory about language use, and in order to build one, it seems necessary to know previous relevant facts about language use. Neutral, shallow annotations give 100 % coverage, even though they imply a loss in depth. Simpler annotation seems a better starting point because it is always possible to add new finer-grained annotation levels over a first shallow one.

3.2 Semantic annotation

The main purpose for semantic annotation is to have a useful resource for Word Sense Disambiguation (WSD) systems in Spanish. This sematically annotated corpus will be used as a training corpus for the development of unsupervised systems and as a reference in general evaluation tasks. Only nouns, verbs, adjectives and adverbs will be annotated. At the end of the project, we will have a large amount of words with an unambiguous sense tag in a real context.
In order to annotate the sense of the words, we need a lexical database containing the words of the language with all their possible senses. The lexical resource we use is the Spanish WordNet. We have decided to use WordNet for several reasons. First of all, because up to now WordNet is the lexical resource commonly used in Word Sense Disambiguation tasks. Secondly, it is one of the most complete lexical resources currently available for Spanish. Finally, as part of EuroWordNet, the lexical structure of Spanish and the lexical structure of Catalan and Basque are related, therefore, the annotated senses of all the three corpus of the project 3LB are related too.
Our proposal is based on the SemCor corpus (Miller *et al.* 1993). This corpus is formed by a portion of the Brown corpus and the novel *The Red Badge of Courage*. Altogether, it is formed by approximately 250.000 words, where nouns, verbs, adjectives and adverbs have been annotated manually with WordNet senses (Miller 1990). Another corpus with semantic annotation based on WordNet is the DSO corpus (Ng and Lee 1996)[4]. In this corpus the most frequent English ambiguous nouns and verbs had been annotated with the correct sense (121 nouns and 70 verbs). The corpus is formed by 192.800 sentences from the Brown Corpus and the *Wall Street Journal*, and it has also been annotated by hand. Finally, the SENSEVAL forum has developed a few sense annotated corpora for the evaluation of the WSD systems (Kilgarrif and Palmer 2000), some of which also use WordNet as a lexical resource.
The tag used to mark the sense of a word is its offset number, that is, the identification number of the sense (synset) in the InterLingua Index of EuroWordNet.
We can distinguish two methods for annotate a corpus semantically. The first one is linear (or "textual" method: (Kilgarriff 1998)), where the human tagger annotates the sentences token by token up to the end of the corpus. This way the tagger must read and analyze the sense of each word every time it appears in the corpus. The second method is transversal annotation (or "lexical" method: (Kilgarriff 1998)), where the human tagger annotates word-type by word-type, all the occurrences of a word in the corpus first, then all the occurrences of an other word, and so on. With this method, the tagger must read and analyze all the senses of a word only once[5].
We have followed the transversal process. The main advantage of this method is that we can focus our attention on the sense structure of one word and deal with its specific semantic problems: the main sense or senses of this word, the specific senses, and so on. Then we check the context of the single

[4] Available from the Linguistic Data Consortium (http://www.ldc.upenn.edu)
[5] This transversal o lexical method has been used in other projects as, for example, the HECTOR project.

word each time it appears and select the corresponding sense. Through this approach, the semantic characteristics of each word is taken into consideration only once, and the whole corpus achieves greater consistency. Through the linear process, however, the annotator must remember the sense structure of each word and their specific problems each time the word appears in the corpus, making the annotation process much more complex, and increasing the possibilities of low consistency and of disagreement between the annotators.
On the other hand, the transversal method is at a disadvantage when we have to annotate a large corpus, because no fragment of the corpus is available until we have annotated the whole corpus completely. To avoid this, we have selected a fragment of the whole corpus and have annotate it by means of the linear process.
Everybody agrees that semantic annotation is a tedious and difficult task. From a general point of view, the main problem in the semantic annotation is the subjectivity of the human tagger when it comes to the selection of the correct sense, because there is usually more than one sense to a word, and, due to the granularity of WorNet, more than one could be right for a given word. When it occur, we annotate all the correct senses.
Another important problem in the semantic annotation is the poor agreement between the different annotators, due to the ambiguity and/or vagueness of many words. Unfortunately, at this first step of the semantic annotation process, we do not have enough data about the agreement between annotators. This is still an open project and we will extract this kind of data in the coming months.
Finally, not all nouns, verbs, adjectives and adverbs are annotated, because EuroWordNet is a limited resource and does not contain all words or senses. As far as we can see it lacks (i) the synset, (ii) the word, (iii) the synset and the word, and (iv) the link between the synset and the word.
In order to deal with these cases we have defined two more tags in EuroWordNet:

- C1S: for cases where the word is located, but not its correct sense (due to a lack of sense, or because there is no link between the word and the synset).
- C2S: for cases where the word is not located (because it is not there, or because both the word and the synset is missing).

3.3 Linguistic coreference annotation

At the pragmatic level, we focus on the coreference and anaphora annotation, which are "indispensable, albeit time-consuming, preliminary to anaphora resolution, since the data they provide are critical to the development, optimisation and evaluation of new approaches" (Mitkov, 2002: 130).
We agreed to annotate the coreferential elements and their antecedents. These coreferential elements are the coreferential ellipsis, the pronominal anaphora and the coreferential chains.
Specifically, in each one, we mark:

Coreferential Ellipsis:
- The elliptical subject, made explicit in the syntactic annotation step. Being a noun phrase, it could also be an antecedent too.
- The coreferential adjective, that is, noun phrases with an adjective complement where the noun head is elliptic. The antecedent is usually a previous noun phrase with a similar structure.

Anaphora: Two kinds of pronouns
- The tonic personal pronouns in the third person. They can appear in subject function or in object function.
- The atonic pronouns, specifically the clitic pronouns that appear in the subcategorization frame of the main verb.

Finally, there are sets of anaphoric and elliptical units that corefer to the same entity. These units form coreferential chains. They must be marked in order to show the cohesion and coherence of the text. They are annotated by means of the identification of the same antecedent. Therefore, anaphoric pronouns and elliptical elements that share the same antecedent form a coreferential chain.
Definite descriptions are a special type of coreferential expression. They consist of nominal phrases that refer to an antecedent. We do not mark them because they pose specic problems that make this task very difficult: Firstly, there is not clear criteria that allows us to distinguish between coreferential and no coreferential nominal phrases; secondly, there is not a clear typology of definite descriptions; and finally, there is not a clear typology of relationships between the definite description and their

antecedents. These problems could further increase the time-consuming in the annotation process and widen the gap of disagreement between the human taggers.
This proposal of annotation scheme is based on the one used in the MUC (Message Understanding Conference) (Hirschman 1997) as well as in the works of Gaizauskas and Humphreys 1996 and Mitkov *et al.* 2000: this is the scheme mostly used in coreferential annotation (Mitkov 2002).
In the coreference annotation, two linguistic elements must be marked: the coreferential expression and its antecedent. In the antecedent we annotate the following information:

- A reference tag that shows the presence of an antecedent ("REF"),
- An identification number ("ID"),
- The minimum continuous substring that could be considerer correct ("MIN").

In the coreferential expression, we annotate:

- The presence of a coreferential expression ("COREF"),
- An identification number ("ID"),
- The type of coreferential expression: elliptical noun phrase, coreferential adjective, tonic pronoun or atonic pronoun ("TYPE"),
- The antecedent, through its identification number ("REF"),
- Finally, a status tag where the annotator shows his confidence in the annotation ("STATUS").

In the next section, we will show a sample of the corpus with the coreferential annotation.
As previously mentioned in this paper, the main problem in the coreferential annotation is the low agreement between human taggers. There is usually less agreement in coreference annotation than in syntactic annotation (Mitkov 2002: 141). In order to reduce this low agreement, we annotate only the most clear type of coreferential units (pronouns, elliptical subject and coreferential adjectives), and we introduce the lowest necessary information. Moreover, with the tag "STATUS", the human tagger can show his confidence in the coreferential unit annotated and in the antecedent marked. However, at the moment, as occurs in the semantic annotation, we do not have enough data on the agreement between annotators.

5 Conclusions

In this paper we have presented the annotation scheme of the corpus Cast3LB, that is focused on three linguistic levels: syntactic, semantic and pragmatic. At the syntactic level we mark constituents and functions; at the semantic level, the senses of the words and at the pragmatic level the coreference phenomena. These three levels are annotated independently, but they are related to each other through the XML format.

Semantic annotation XML format (with syntactic information)

```
<Annotation id="ejemplo1:EJ1:Annotation4" type="wrd"
start="ejemplo1:EJ1:Anchor2" end="ejemplo1:EJ1:Anchor3">
<Feature name="label">gato</Feature>
<Feature name="synset">01457160n </Feature>
<Feature name="synset">01458079n </Feature>
<Feature name="synset">06051878n </Feature>
<Feature name="parent">ejemplo1:EJ1:Annotation5</Feature>
</Annotation>

<Annotation id="ejemplo1:EJ1:Annotation5" type="pos"
start="ejemplo1:EJ1:Anchor2" end="ejemplo1:EJ1:Anchor3">
<Feature name="lema">gato</Feature>
<Feature name="label">ncms000</Feature>
<Feature name="parent">ejemplo1:EJ1:Annotation6</Feature>
</Annotation>
```

Coreferential annotation XML format

```
<REF ID:R1 MIN:Menardo Fraile> Medardo Fraile </REF> juega a
un cinismo fácil y divertido. No quiero decir que lo sea,
cínico o divertido, sino que ante un mazo de hojas grabadas
<COREF ID:R2 REF:R1 TYPE:SUBJ-ELLIP STATUS:CIERTO> (SN)
</COREF> coloca <REF ID:R3 MIN:un cristal> un cristal bien
tallado </REF> y <COREF ID:R4 REF:R3 TYPE:CLIT STATUS:
CIERTO> lo </COREF> hace girar para que el sol rompa contra
<COREF ID:R5 REF:R3 TYPE:PRON STATUS:CIERTO> él </COREF> sus
rayos.
```

Syntactic annotation PTB format

```
(sadv
      (rg Claro))
    (S.F.C
      (conj.subord
        (cs que))
      (sn
        (espec.ms
          (da0ms0 el))
        (grup.nom.ms
          (ncms000 disparate)))
      (gv
        (vsip3s0 es))
      (S.NF.C
        (infinitiu
          (vmn0000 contemplar))
        (sn
          (espec.fs
            (dd0fs0 aquella))
          (grup.nom.fs
            (ncfs000 realidad)))
        (sp
          (prep
            (sps00 con))
          (sn
            (espec.fs
              (da0fs0 la))
            (grup.nom.fs
              (s.a.fs
                (aq0fs0 fría))
              (ncfs000 lógica)
              (sp
                (prep
                  (spcms del))
                (sn
                  (grup.nom.ms
                    (S.NF.P
                      (aq0msp adormecido))
                    (ncms000 burgués)))))))))
```

6 Bibliography

Abeillé A., Clément L., Kinyon A. 2001 Building a Treebank for French. In Abeillé A. (ed), *Treebanks: Building and Using Syntactically annotated corpora.* Dordrecht, Kluwer, pp 165-187.

Alongue A., Calzorali N., Vossen P., Blocksma I., Castellón I., Martí M.A., Peters W. 1998 The Linguistic Design of the EuroWordNet. In *EuroWordNet: A multilingual database with lexical semantic network*, Kluwer.

Atserias, J. and Rodríguez H. 1998 *TACAT: TAgged Corpus Text Analyzer* Technical Report.Software Department (LSI). Technical University of Catalonia (UPC).

Bemova, A., Hajic J., Hladka B. and Panevova J. 1999 Morphological and Syntactic Tagging of The Prague Dependency Treebank *Journées Atala, Corpus annotés pour la syntaxe*, Paris, June.

Boguslavsky, I., Chardin I., Grigorieva S., Grigoriev N., Iomdin L., Kreidlin L. and Frid N. 2002 Development of a Dependency Treebank for Russian and its possible Applications in NLP *Proceedings of the Third Conference on Language Resources and Evaluation* (LREC2002), May.

Brants, T., Skut W. and Uszkoreit H. 2001 Syntactic Annotation of a German Newspaper Corpus In Anne Abeillé, editor, *Building and Using syntactically annotated corpora*, Kluwer, Language and Speech.

Brants T., Skut W., Uszkoreit H. 2001 Syntactic Annotation of a German Newspapaer Corpus. In Abeillé A. (ed), *Treebanks: Building ans Using Syntactically annotated corpora.* Dordrecht, Kluwer, pp 73-88.

Civit M., Martí M.A. 2002 Design Principles for a Spanish treebank1st Workshop on Treebanks and Linguistic Theories (TLT02), Sozopol, Bulgaria.

Civit M., Martí M.A., Navarro B., Bufí N., Ferrández B., Marcos R. 2003 Issues on the Syntactic Annotation of Cast3LB, In *Proceedings of the 4th International Workshop onLinguistically Interpreted Corpora (LINC 2003). EACL Workshop*. Budapest (Hungary).

Civit 2003 *Criterios de etiquetación y desambiguación morfosintáctica de corpus en Español.* PhD Thesis. Universitat de Barcelona. Forthcoming.

Gaizauskas R. and Humphreys K. 1996 Quantitative evaluation of coreference algorithms in an information extraction system, In Botley S. P. and McEnery A. M. (eds.), *Corpus-based and Computational Approaches to Discourse Anaphora*, Amsterdam, John Benjamins, pp. 143-167.

Hinrichs E., Simov K. 2002 *Proceedings of Workshop on Treebanks and Linguistic Theories (TLT 2002)*, Sozopol, Bulgaria (http://www.bultreebank.org/Proceedings.html).

Hirschman, L. 1997 MUC-7 coreference task definition *Message Understanding Conference Proceedings* (http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/co_task.html)

Kilgarriff A. 1998 Gold Standard Datasets for Evaluating Word Sense Disambiguation Programs. In *Computer Speech and Language. Special Use on Evaluation* 12(4), pp 453-472.

Kilgarriff A. and Palmer M. (eds.) 2000 *Computer and the Humanities. Special Issue on SENSEVAL*, 34 (1-2).

Marciniak M., Mykowiecka A., Przepiórkowsky A., Kupsc A. 2001 An HPSG-Annotated Test Suite for Polish. In Abeillé A. (ed), *Treebanks: Building ans Using Syntactically annotated corpora.* Dordrecht, Kluwer, pp 129-146.

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

Marcus, M, Kim G., Marcinkiewicz M.A., MacIntyre R., Bies A., Ferguson M., Katz K. and Schasberger B. 1994 The Penn Treebank: Annotating Predicate Argument Structure *Proceedings of the ARPA Human Language Technology Workshop*, Princeton.

Miller G. A. 1990 WordNet: An on-line lexical database, In *Intenational Journal of Lexicography*, 3(4), pp 235-312.

Miller G. A., Leacock C., Randee T., Bunker R. 1993 A Semantic Concordance, In *Proceedings of the 3rd DARPA Workshop on Human Language Technology*, Plainsboro, New Jersey, pp 303-308.

Mitkov R., Evans R., Orasan C., Barbu C., Jones L., and Sotirova V. 2000 Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies, In *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC 2000)*, Lancaster, UK.

Mitkov R. 2002 *Anaphora resolution*, London, Pearson.

Montemagni S., Barsotti F., Batista M., Calzolari N., Corazzari O., Lenci A., Zampolli A., Fanciulli F., Massetani M., Raffaelli R., Basili R., Pazienzia M., Saracino D., Zanzotto F., Mana N., Pianesi F., Delmonte R. 2001 Building the Italian Syntactic-Semantic Treebank. In Abeillé A. (ed), *Treebanks: Building and Using Syntactically annotated corpora.* Dordrecht, Kluwer, pp 190-210.

Moreno A., López S., Sánchez F., Grishman R. 2001 Building a Spanish Treebank. In Abeillé A. (ed), *Treebanks: Building and Using Syntactically annotated corpora.* Dordrecht, Kluwer, pp 149-164.

Ng H. T., Lee H. B. 1996 Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics,* Santa Cruz, California, pp 40-47.

Oflazer, K., Say B., Hakkani-Tür D.Z. and Tür G. 2001 Building a Turkish Treebank In Anne Abeillé, editor, *Building and Using syntactically annotated corpora*, Kluwer, Language and Speech.

Palomar M., Ferrández A., Moreno L., Martínez-Barco P., Peral J., Saiz-Noeda M., Muñoz R. 2001 An algorithm for anaphora resolution in Spanish texts. In *Computational Linguistics* 27(4), pp. 545-567.

Rambow, O., Crecwell C., Szekely R., Taber H. and Walker M. 2002 A Dependency Treebank for English *Proceedings of the Third Conference on Language Resources and Evaluation (LREC2002)* May.

Rojo G. 1990 La explotación de la Base de datos sintácticos del español actual (BDS). In Kock J. (ed) *Gramática española: Enseñanza e investigación*. Salamanca, Universidad of Salamanca.

Sampson. G. 1987 Probabilistic models of analysis. In Garside, R., Leech G. and Sampson G. (eds), *The Computational Analysis of English* Chapter 1, Longman.

Sebastián N., Martí M.A., Carreiras M.F., Cuetos F. 2000 *LEXESP: Léxico Informatizado del Español*, Barcelona, Edicions de la Universitat de Barcelona.

Simov K., Osenova P., Slavcheva M., Kolkovska S., Balabanova E., Doikoff D., Ivanova K., Simov A., Kouylekov M. 2002 Building a Linguistically Interpreted Corpus of Bulgarian: the BulTreeBank. In *Proceedings of LREC 2002*, Canary Islands, Spain, pp 1729-1736.

Taylor, A., Marcus M. and Santorini B. 2001 The PennTreeBank: an overview. In Abeillé A. (ed), *Building and Using syntactically annotated corpora*, Kluwer, Language and Speech.

Tutin A., Trouilleux F., Clouzot C., Gaussier E., Zaenen A., Rayot S., Antoniadis G. 2000 Anotating a large corpus with anaphoric links, In *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC 2000)*, Lancaster, UK.

Vossen, P. (ed.) 1991 *EuroWordNet General Document*, EuroWordNet (LE2-4003, LE4-8328), Part A, Final Document (http://www.illc.uva.nl/EuroWordNet/docs.html).