Issues in the Syntactic Annotation of Cast3LB

Montserrat Civit†, Ma. Antònia Martí†, Borja Navarro‡,
Núria Bufí†, Belén Fernández‡, Raquel Marcos‡
†CLiC Centre de Llenguatge i Computació
Universitat de Barcelona
{civit,amarti,nuria@clic.fil.ub.es}
‡Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
{borja,bfernan,rmarcos@dlsi.ua.es}

Abstract

In this paper we present some specific problems concerning the annotation of **Cast3LB**, a 100,000-word Spanish treebank. We summarise the general annotation framework and discuss in more detail crucial problems found during the annotation task like ellipsis, punctuation marks, complex verb forms, comparative sentences and coordinated elements.

1 Introduction

In this paper we present some specific problems concerning the annotation of a Spanish treebank. It is our aim to build a 100,000-word Spanish treebank (**Cast3LB**) which will be enriched in the future with semantic as well as pragmatic information, and which will be free for research purposes. The construction of such a treebank is part of the **3LB** project[1], funded by the Spanish government, whose goals are to build three treebanks: one for Catalan (**Cat3LB**), one for Spanish (**Cast3LB**) and finally another for Basque (**Eus3LB**).

To start working with our corpus we need a series of NLP tools described in (Civit and Mart´ı, 2002): a morphological analyser, a tagger and a chunker for Spanish. In order to build the treebank the AGTK toolkit (Cotton and Bird, 2002) is being used, since it allows us to use our tagset and our chunker output.

[1]PROFIT (FIT-150500-2002-244);
URL: http://www.dlsi.ua.es/projectes/3lb/index_en.html.

There are four levels of annotation:
level 1: syntactic annotation: constituents;
level 2: syntactic annotation: functions;
level 3: semantic annotation: words senses;
level 4: pragmatic annotation: coreference.
This paper is centered on the first level: annotation of syntactic constituents.

In the establishment of our framework we have taken into account preceding works covering a wide range of languages in order to cope with different kinds of syntactic phenomena: (Abeill´e et al., 2002), (Bies et al., 1995), (Brants et al., 2002), (Hajic, 1998), (Montemagni et al., 2001), (Moreno et al., 2001)[2].

Our main goal has been to define consistent criteria for the syntactic annotation (Civit, 2002), especially in what concerns the treatment of coordination, comparative structures, complex verb forms, etc. and the establishment of general principles to handle with some specific problems such as punctuation marks that, so far, have neither received an homogeneous nor an unified treatment. These are the points we concentrate on in this paper.

2 General framework

Corpus Linguistics and more concretely the development of treebanks have a short background. That is why there is not a well-established methodology for developing such linguistic resources. There is a wide range of choices for researchers

[2]It should be mentioned that this Spanish treebank is not freely available and that the annotaions we propose are completely different.

when the definition of the annotation criteria is to be done, or the units of analysis and the representation system are to be chosen. In **Cast3LB** we follow an incremental process in the levels of annotation. We have taken decisions in order to make our system flexible and transportable to different romance languages and to new cases that may appear, but also consistent in all levels of annotation and with regard to linguistic data. This viewpoint justifies the methodological decisions presented in what follows.

We annotate explicit elements and only add new nodes for elliptical subjects of finite sentences. Since Spanish is a pro-drop language, the subject may not appear explicitly, although it is always recoverable from verbal inflexion because of the agreement in person and number with the verb. Information about the subject is necessary especially for further annotation of anaphora and coreference. Apart from that, we do not add new nodes to the tree (see section 3.1 for more details about this phenomenon).

As for concrete annotation of sentences, it should be noticed that we do not alter the surface word order of elements because it could mean a loss of pragmatic information. However, having taken this decision, we need to face different problems, all of them related to discontinuous constituents. We have adopted some conventions to mark the syntactic function of such elements (see section 3.3).

We decided to follow the constituency annotation scheme, instead of annotating dependencies. Constituency seems better for Spanish, which is a free constituent order language but has a quite fix word order within the constituents.

We do not only label constituents with tags such as *noun phrase, prepositional phrase,* etc., but also give them a functional tag *subject, object,* etc[3]. However, we do not deal with nominal complements. Our main goal now is to acquire knowledge of language performance in real text in order to infer a full and deep grammar for Spanish and to build a verbal lexicon with subcategorisation information.

Due to this reason, we do not follow any concrete theoretical framework. Instead, we want **Cast3LB** to be useful for as many people as possible, linguists as well as computer scientists or anyone interested in the Spanish language.

Verbal phrases pose a very in-depth problem regarding the relationship between the verb and its arguments and between the subject and the verbal phrase. Theoretically speaking, the latter contains the verb, its arguments and some of the adjuncts, so this situation should be reflected in the syntactic annotation. We found several reasons to justify the decision of not taking into consideration the verbal phrase node and to adopt a more neutral solution in which all the main constituents of the sentence (subject, verb, arguments and adjuncts) are daugthers of the root node. Firstly, language use shows that in many cases it is not easy to differentiate between arguments and adjuncts (as stated in (Marcus et al., 1994)). Secondly, there is a long discussion about whether the subject is an argument or not. Thirdly, there is a lack of a wide coverage lexicon for Spanish with rich information about argument structure of verbs. Finally, since, in Spanish adjuncts can appear at any place of the sentence, we should alter the surface word order to reflect this relationship. All in all, as said above, our decision was not to take into consideration the verbal phrase and give a quite flat representation of sentences[4],[5]:

```
(S
  (sn
    (espec.fs
      (di0fs0 Una))
    (grup.nom.fs
      (ncfs000 información)
      (s.a.fs
        (aq0fs0 periodística))))
  (sn
    (pp3csd00 le))
  (gv
    (vmis3s0 desmontó))
  (sn
    (espec.fs
      (da0fs0 la))
    (grup.nom.fs
      (ncfs000 estratagema)))
  (Fp .)))
```

[3] See appendix A for the **Cast3LB** tagset.

[4] This is the real representation of trees. However, in order to simplify the examples, in what follows we provide simpler forms of analysed sentences. Moreover, tags for syntactic functions appear only when functions are involved; if not, we only reproduce the constituent labeling.

[5] 'A piece of journalistic information dismantled his stratagem'.

Last, but not least, we only give one functional tag to each constituent. This means that we do not deal with control structures for the moment[6].

3 Issues in the annotation process

In this paper we would like to concentrate on concrete issues on annotation, those that showed a high degree of variation among the different **Cast3LB** annotators. We discuss here problems concerning the general structure of the sentence: ellipsis, punctuation marks, complex verb forms and some kinds of structures such as comparative clauses and coordinated elements. In next subsections these problems are discussed and the adopted solution is given.

3.1 Ellipsis

Ellipsis is a phenomenon that may appear in all the sentence constituents. Only subject ellipsis is recovered in our framework (see section 2). In the other cases, no node is added to the trees. The main reasons for such a decision are that we do not wish to alter linguistic data from corpus; that recovering elliptical elements requires subjective interpretation and results may differ among annotators; that there is a huge amount of such elements and one may find more elliptical elements than explicit ones; and, finally, that one should follow one grammatical theory. However, when the elliptical element is the verb, we mark the sentence node with a * symbol. The rest of the cases are left unspecified.

We can distinguish four main cases in verbal ellipsis. The first one happens in coordinated clauses sharing the same verb. In this case the second clause has no verbal form (example-1). A second case of elliptical verb appears in coordinated sentences having a complex verb form: an auxiliary finite verb and a main verb in a non-finite form. In this case, the auxiliary verb may be elided in the second sentence (example-2). A third kind of verbal ellipsis occurs in comparative sentences, because they share the verb with the main clause too (example-3). Finally, in general, this phenomenon may occur in any kind of sentence, as shown in examples 4 and 5[7].

Example-1:

```
S.co_[
    S_[ Zarrabeitia puso la rebeldía , ]
    coord_[ y ]
    S*_[ Delgado la gallardía ] . ]
```

Example-2:

```
S.co_[
    S_[ ... habría alcanzado a Camargo ]
    coord_[ y ]
    S*_[ obtenido un botín de... ] . ]
```

Example-3:

```
S_[ ...
    sn_[ nos ]
    gv_[ han señalado ]
    sn_[
       sn_[ más faltas ]
       S.F.AComp* [ que a nuestros
                    rivales] ] . ]
```

Example-4:

```
S*_[
    sa_[
      s.a.mp_[ Indignos
             sp_[ de la civilización
                  que les cobija]]] . ]
```

Example-5:

```
S*_[
   sp_[ Por qué ]]
```

The main idea behind this annotation is to be able to retrieve all sentences with elliptical verbs in order to start a linguistic study of this phenomenon.

3.2 Punctuation marks

In the tradition of Corpus Linguistics, punctuation marks are considered one among other elements of the text and are usually treated as another class of words (i.e. they receive a specific tag). In the **Cast3LB** framework, they maintain their *POS*-tag and no more labels are added to them at the syntactic annotation level.

Up to now, the solution they have received has been limited to the place they fill in the tree. In the literature (see (Bies et al., 1995), (Abeill´e et al., 2002) and (Moreno et al., 1999)) different solutions are proposed, all of them being different. It is true that strong marks (full stop, question and

[6]See also (Civit and Mart´ı, 2002) for a more detailed discussion about the general framework.

[7]Ex-1: 'Zarrabeitia put the rebelliousness and Delgado the courage'; Ex-2: '... he would have caught up Camargo and obtained a booty of ... '; Ex-3: '... there have been more fouls given against us than against our rivals'; Ex-4: 'Unworthy of the civilization that shelters them'; Ex-5: 'Why?'.

exclamation marks[8] and three dots) usually are the last element of the sentence; but, as for medium marks (colons, semicolons and commas), given solutions differ extensively.

If we concentrate on colons, in the PennTreeBank guidelines (Bies et al., 1995) only two 'possibilities' are presented: colons in appositions and *colorful environments*. In the Spanish treebank (Moreno et al., 1999), colons are treated the same way than other punctuation marks: *they will be the sister or the daugther of their closer constituent, depending on their position inside the sentence.* Finally, in the *Corpus Le Monde* (Abeill´e et al., 2002) it is stated that colons, and punctuation marks in general, are usually not included inside any constituent. However, some commas belong to the subordinate clauses.

Handling with punctuation marks is not a trivial matter, and if one wants to provide a consistent annotation, a critical analysis about this issue is required. We have endeavoured to relate their place in the tree with the function they have in the sentence. In some cases, punctuation marks are used for delimiting and identifying a specific structure. In other cases, they provide some kind of semantics to the understanding of the whole sentence. In the first case, our general rule is to include the punctuation mark in the node it parenthesises: at the beginning or at the end of subordinated clauses; at the beginning and at the end of parenthetical elements; etc. In coordinations without conjunctions, punctuation marks are given the same place than conjunctions; in the other cases, they are siblings of the coordinated nodes.

The second case refers mostly to colons, which we interpret differently taking into account their semantics. According to the correct use stated by the Spanish Academy of Language the main uses they have in Spanish are the following: before or after an enumeration; as introductory elements of direct speech; as a delimiter of an example from the rest of the sentence; and, in general, as connective elements to link clauses, which may express different relationships: cause, consequence, conclusion, summary, explanation of what precedes, etc.

[8]Notice than in Spanish these marks appear at the beginning and at the end of the sentence.

In order to systematise the analysis of colons, we extracted 35 sentences containing colons out of the corpus. After a detailed analysis we get this typology:

1. Colons introduce direct speech. Then, they are treated as if they were a subordinating conjunction and so are the first element of a completive clause.

   La gran cuestión era: ¿se tendrá Romario que ir a Río?[9]

```
S_[
   sn_[ La gran cuestión ]
   gv_[ era ]
   S.F.C_[
         :
         >
         morf.pron_[ se ]
         gv_[ tendrá
              sn_[ Romario ]
              que ir ]
         sp_[ a Río ]
         ? ] ]
```

2. Colons are the beginning (sometimes the end) of an enumeration. This is the so-called *shopping-list model*, and is, by far, the most frequent use of colons. Here, they are the first element of the series (example a) or the last, if the series precedes the head (example b).

   a.- Amenaza con ganarlo todo: la general, la montaña, la clasificación por puntos[10]

```
S_[
   sn.e_[ *0* ]
   gv_[ amenaza ]
   sp_[
     con
     S.NF.C_[
      infinitiu_[ ganarlo ]
      sn_[ todo
         sn.co_[
             :
             sn_[ la general ]
             ,
             sn_[ la montaña ]
             ,
             sn_[ la clasifi-
             cación por puntos]
             ]]]]
```

   b.- A su lado, el apoyo de un amigo, compadre, compañero leal y alter_ego: Stoichkov[11]

[9]'The great question was: will Romario have to go to Rio?'

[10]'He threatens to win everything: the overall standings, the climber and the points overall standings'.

[11]'By his side, the support of a friend, buddy, loyal companion and alter ego: Stoichkov'.

```
S*_[
   sp_[ A su lado , ]
   sn_[ el apoyo
      sp_[ de
         sn [
            grup.nom.co [
          un amigo, compadre,
          compañero leal y
          alter_ego : ]
            Stoichkov ] ] ]
   . ]
```

3. Colons introduce a complex element of the sentence, usually a clause, whose reference element is the whole preceding text. In this case, colons are attached to the highest right node and the whole constituent is adjoined to the element it refers.

   3.- Entonces decidió cambiar la perspectiva: desprestigió a la heroína devolviéndola a su hogar y la convirtió en una mujer común, salvo en su soberana belleza.[12]

```
S_[
  sadv_[ Entonces ]
  gv_[ decidió ]
  S.NF.C_[
   S.NF.C_[ cambiar la perspectiva ]
   S.co_[
     :
     S_[ desprestigió ... hogar ]
     coord_[ y ]
     S_[ la convirtió ...
                belleza ]]]
  . ]
```

Four of the found examples did not match exactly any of the described situations, but they showed similar characteristics.

We consider all of them as belonging to the *shopping-list* kind, but the noun they referred to did not immediately precede them. For instance,

Los demás ya son conocidos: la composición del resto del podio, la probabilidad de que Toni_Rominger se apunte también la montaña y la regularidad.[13]

In these cases, we adopted the adjunction model representation, even though, conceptually speaking, they should be sorted out in the second group.

[12]'Then he decided to change the perspective: he stripped the heroine of all her outstanding qualities returning her to her home and transforming her into a common woman, except for her breathtaking beauty '.

[13]'The rest of them are already known: the composition of the rest of the podium, the probability that Toni_Rominger succeeds too in winning the climber and the regularity overall standings'.

3.3 Complex verb forms

The Spanish language has an important number of so-called *periphrastic* verb forms. These are a combination of a finite verb form, optionally followed by a preposition or a conjunction, plus a non-finite verb form. The finite form gives the inflexion and adds some modal meaning; the non-finite one not only brings the meaning of the compound but also selects the complements of the whole form.

The grammar used by the chunker deals with the main periphrastic forms in Spanish. A set of 35 different complex verb forms was established there. However, periphrastic phenomenon is more a continuum than a set of clear-cut cases. Two of the main tests to check whether a given sequence of verbs is a periphrasis are: (1) it is not possible to substitute the non-finite verb form by a pronoun or a nominal expression; (2) clitics might appear before or after the whole compound. Let's consider the sentence

desea comprar un libro[14].

It is possible to replace the infinitive form (*comprar*) and its complements by a demonstrative pronoun, even by a clitic: *desea esto / lo desea*[15]. Instead, in

puede comprar un libro [16]

the substitution of the infinitive is impossible, and, however, it is possible to substitute *el libro* by a clitic, which may appear after or before the compound: *puede comprar***lo** / **lo** *puede comprar*. Such substitution is also possible in the first example, but the clitic has to appear after the compound: *desea comprar***lo** / ***lo** *desea comprar*.

Although almost all the examples conform to the rules, there are some cases in which one of the two rules does not work. This happens with the sequence *querer* + infinitive[17], like in the sentence

quiere comprar un libro [18].

On the one hand, the clitic accepts the two positions: **lo** *quiere comprar–quiere comprar***lo**; but on

[14]He/she wishes to buy a book.

[15]He/she wishes this / it.

[16]He/she can buy a book.

[17]'*to want to* + infiniti ve'.

[18]'He/she wants to buy a book'.

the other, the infinitive admits to be replaced by a pronoun: *quiere esto*[19].

There are some other middle-case periphrases like this one. As a general criterion we decided to consider complex verb forms only those following the two criteria, so *querer* + infinitive will not be considered as a periphrasis.

In this case, we will find sequences such as *clitic + inflected verb + infinitive*[20] where the clitic depends on the infinitive. Given our flat representation of sentences, we face here a discontinuous constituent: the clitic is the direct object of the infinitive, which is the direct object of the finite verb form. The solution taken in the **Cast3LB** framework is to adopt a convention to mark this relation in the functional tag (**CD.NF**):
lo quiere comprar:

```
S_[
   sn.e-SUBJ_[ *0* ]
   sn-CD.NF_[ lo ]
   gv_[ quiere ]
   S.NF.C-CD_[ comprar ]
  ]
```

Periphrastic forms entail other problems. Even if they behave as compounds, it is quite usual in Spanish to find embedded adverbs or noun phrases in the compound:

puede quizá comprar / puede alguien comprar[21].

As we do not alter the word order, these atypical verbal groups having a non-verbal element in the middle are analysed as follows:

```
gv_[ puede
     sadv-CC_[ quizá ]
     comprar ]

gv_[ puede
     sn-SUBJ_[ alguien ]
     comprar ]
```

where the embedded element appears between the auxiliary verb form and the infinitive.

3.4 Comparative sentences

Comparative structures epitomize discontinuous constituents. In Spanish, more than in English, comparison is a syntactic phenomenon[22]. In the basic case, adverbs *más / menos*[23] depend on an adjective and then the conjunction *que*[24] introduces the comparative clause:

se muestren más frágiles que la moral del suizo[25]

Moreover, the clause has only one element, usually a noun phrase and the rest of the structure has to be inferred from the first part of the sentence. Spanish linguistic tradition considers comparative clauses to be sentential adjuncts (together with conditionals, concessive and consecutive clauses). However, in our framework, comparative elements are adjoined to the element containing the comparative adverb. The reason why we do so is because we think that we cannot split both elements (the adverb and the conjunction) in two different constituents; so the analysis for the sentence is:

```
S_[
   sn.e_[ *0* ]
   morf.pron_[ se ]
   gv_[ muestren ]
   sa_[
      sa_[ más frágiles ]
      S.F.AComp*_[ que la moral
                   del suizo ]]]
```

As shown in the example, the comparative clause belongs to the same constituent than the adjective in an adjoined structure.

3.5 Coordinated elements

We consider coordinated elements to be equivalent in the syntactic structure, so they are always represented as siblings (i.e. there is no head in coordinated structures). The tag of the mother node is the same than that for the daughters with the suffix **.co**. However, sometimes it happens that siblings do not belong to the same grammatical category. In this case we give the mother the less marked tag of her daughters according to their syntactic function. For instance, if coordinated nodes are an adverbial and a prepositional phrase (**sadv** and **sp**), the mother node is **sadv.co**, if the structure is a verb complement; instead, if it is a noun complement the tag is **sp.co**. If coordinated nodes are an adverbial sentence with a finite verb form (**S.F.A**) and an adverbial sentence with a non-finite verb form (**S.NF.A**), the tag for the mother is **S.F.A.co**.

[19] 'He/she wants that'.

[20] Sometimes + gerund, too.

[21] Literaly: *He / she can maybe buy. / Can somebody buy.*

[22] There are only half a dozen morphological comparatives: *mejor, peor, mayor, menor, superior, inferior* ('better, worse, bigger/older, smaller/younger, higher, lower').

[23] 'more / less'.

[24] 'than'.

[25] ' They seem weaker than the Swiss's morale '.

Sometimes, in coordinated sentences, there is a shared complement: *compra y vende casas*[26]. As in our framework we have as many sentences as verbal forms, the only way to show that *casas* depends on the two verbs is to include the noun phrase as a daughter of the coordinated node:

```
S.co_[
   S_[
     sn.e-SUBJ_[ *0* ]
     gv_[ compra] ]
   coord_[ y ]
   S_[
     sn.e-SUBJ_[ *0* ]
     gv_[ vende ]]
   sn-CD_[ casas ]
   . ]
```

Sometimes, however, the two coordinated verbs have different structures, and one of them requires a preposition for the complement while the other does not: ... *minando o acabando* **con** *sus ilusiones*[27]. In this case, the solution has been to attach the prepositional phrase only to the second verb, even though, semantically speaking, the complement is related to both verbal forms:

```
...
S.NF.A.co_[
   S.NF.A_[
     gerundi_[ minando ] ]
   coord_[ o ]
   S.NF.A_[
     gerundi_[ acabando ]
     sp_[ con sus ilusiones ] ] ]
```

where the node **S.NF.A.co** includes the two coordinated sentences and the prepositional phrase (*con sus ilusiones*) is attached to the second clause.

4 Conclusions and futher work

In this paper we have presented some of the crucial problems concerning the syntactic annotation of corpora in Spanish. So far, there are 2,300 annotated sentences with constituent and functional labels. Further work will consist, on the one hand, on the semantic tagging of nouns, verbs and adjectives using Spanish EuroWordNet (Alonge et al., 1998) and, on the other, on the annotation of anaphora and correference phenomena.

[26] 'He/she buys and sells houses'.

[27] '... undermining or finishing with his hopes'.

References

A. Abeill´e, F. Toussenel and M. Ch´eradame. 2002. *Corpus le Monde. Annotations en constituants. Guide pour les correcteurs* LLF, UFRL.

A. Alonge, N. Calzolari, P. Vossen, L. Bloksma, I. Castell´on, M.A. Mart´ı and W. Peters. 1998. The Linguistic Design of the EuroWordNet Database *EuroWordNet: A multilingual database with lexical semantic networks* Kluwer.

A. Bies, M. Ferguson, K. Katz and R. MacIntyre. 1995. *Bracketing Guidelines for Treebank II Style Penn Treebank Project* LDC.

S. Brants, S. Dipper, S. Hansen, W. Lezius and G. Smith 2002. The TIGER TreeBank *Proceedings of the First Workshop on Treebanks and Linguistics Theories (TLT2002)*, Sozopol, Bulgaria pp. 24–41.

M. Civit and M. A. Mart´ı. 2002. Design Principles for a Spanish Treebank *Proceedings of the First Workshop on Treebanks and Linguistics Theories (TLT2002)*, Sozopol, Bulgaria 61–77.

M. Civit. 2002. *Guía para la anotación sintáctica de Cast3LB: un corpus del español con anotación sintáctica, semántica y pragmática* 3LB WP-02-01; X-Tract-II WP-02-01.

S. Cotton and S. Bird. 2002. An integrated Framework for Treebanks and Multilayer Annotations *Proceedings of the Third Conference on Language Resources and Evaluation (LREC2002)* Las Palmas, Spain.

J. Hajic 1998. Building a Syntactically Annotated Corpus: the Prague dependency Treebank *Issues of Valency and Meaning* 1998.

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

S. Montemagni, F. Barsotti, M. Battista, N. Calzolari, O. Corazzari, A. Lenci, A. Zampolli, F. Fanciulli, M. Massetani, R. Raffaelli, R. Basili, M.T. Pazienza, D.Saracino, F. Zanzotto, N. Mana, F. Pianesi, R. Delmonte. 2001. Building the Italian Syntactic-Semantic Treebank, in Anne Abeill´e, editor, *Building and Using syntactically annotated corpora*, Kluwer, Language and Speech 2001.

A. Moreno, S. L´opez and F. S´anchez. 1999. *Spanish Tree Bank: Specifications (Version 5)* UAM.

A. Moreno, S. L´opez, F. S´anchez and R. Grishman. 2001. Developing a Spanish Treebank In Anne Abeill´e, editor, *Building and Using syntactically annotated corpora*, Kluwer, Language and Speech 2001.

A Syntactic tagset

Table 1 shows the tagset for sentences and clauses.

sentence	S
verbless sentence	S*
infinitive clause	S.NF.C
participle clause	S.NF.P
gerund clause	S.NF.A
completive clause	S.F.C
relative clause	S.F.R
adverbial clause	S.F.A
comparative clause	S.F.AComp
conditional clause	S.F.ACond
concessive clause	S.F.AConc
consecutive clause	S.F.ACons

Table 1: Sentences' tags

Table 2 shows the tagset for main constituents.

sn	noun phrase
sn.e	elliptical noun phrase
gv	verbal group
sp	prepositional phrase
sadv	adverbial phrase
sa	adjectival phrase
conj.subord	subordinating conjunction
coord	coordinating conjunction
interjeccio	interjection
neg	negative adverb
morfema.verbal	*se* in impersonal or passive clauses
morf.pron	*se* in pronominal uses

Table 2: Main constituents' tags

Table 3 shows the tagset used for syntactic functions.

-SUBJ	subject
-CD	direct object
-CI	indirect object
-CC	adverbial complement
-ATRIB	attribute
-CAG	agentive complement
-PRED	predicative complement
-CREG	prepositional complement

Table 3: Syntactic functions