Building Cat3LB: a Treebank for Catalan

Montserrat Civit, Núria Bufí and Ma. Pilar Valverde

CLiC, Centre de Llenguatge i Computació
Universitat de Barcelona
C/ Adolf Florensa s/n (Torre Florensa)
08028 Barcelona
{civit, nuria, pilar}@clic.fil.ub.es

Abstract

In this paper we present some specific issues related to the syntactic annotation of Cat3LB, a 100,000-word Catalan treebank (2,500 sentences). In the syntactic annotation we follow an incremental process with different levels of complexity: bracketing and labelling of constituents and functional tagging. Some automatic pre-processing steps have been done: morphological analysis, tagging and chunking. In this paper we will concentrate, however, on the syntactic annotation.

1. Introduction

Catalan is a 10-milion speaker language spoken in four different countries (Spain, France, Italy and Andorra), but it only has an official status in one of them (Andorra) which is the smallest one. One way to escape from the increasing pressure from the major languages and to avoid the demise of Catalan is to develop basic language resources. In this field, Catalan is not in a bad situation, since basic tools and resources have been developed so far (see section 2.). However, Catalan did not have one of the most important resources needed both to develop NLP applications and to acquire linguistic knowledge about how a language is used: a Treebank.

In this paper we present the development of a freely available treebank for Catalan: **Cat3LB**, built within the **3LB** project, whose goals are to build three treebanks: one for Catalan (**Cat3LB**), one for Spanish (**Cast3LB**) and finally another for Basque (**Eus3LB**). The **3LB** project is funded by the Spanish government[1].

Section 2. describes the previous processes and tools that have been used in the construction of the Treebank. Section 3. presents the main characteristics of the syntactic annotation carried out in the Cat3LB Treebank. Finally, section 4. comes to conclusions.

2. Previous Processes

CLiC-UB[2] and TALP-UPC[3] groups have developed so far a framework for the automatic processing of Catalan and Spanish, based in a pipeline structure shown in figure 1 (Atserias et al., 1998).

These tools include: a morphological analyser (MACO+), a morphosyntactic tagger (RELAX) and a chunker (TACAT). As the tagger is a probabilistic one, the **CLiC-TALP-CAT** corpus was developed in order to allow the tagger to learn the rules for the disambiguation. Then this corpus has been used to build the Treebank. The current CLiC-TALP-CAT corpus consists of 100,000

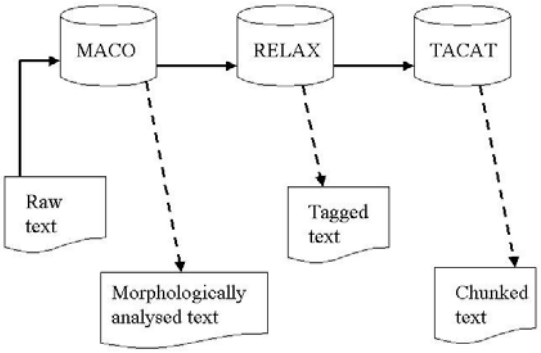

Figure 1: Pipeline of the NLP tools

words coming from, on the one hand, the EFE news agency (25,000 words) and, on the other, from the ACN -Catalan News Agency- (75,000 words). All texts were originally written in Catalan.

MACO+ is a Morphological Analyser for Catalan which provides both lemma(s) and POS-tag(s) for each word. Its output has the following format:

$$word \quad lemma_1 - tag_1 \quad ... \quad lemma_n - tag_n$$

Tags codify 13 part-of-speech categories (noun, verb, adjective, adverb, pronoun, determiner, preposition, conjunction, interjection, dates, punctuation, numbers and abbreviations) as well as subtypes and morphological features, as it is proposed by Eagles (Monachini and Calzolari, 1996). The total amount of tags is 321 and they can be found, with a large explanation, at http://clic.fil.ub.es/doc/categorias_eagles_cat03.htm

The morphological analysis and lemmatisation of the sentence *El secretari general d'ERC ha participat aquest migdia en un acte davant la Delegació d'Hisenda de Tarragona per reclamar un finançament just per Catalunya*[4] is as follows:

El el DA0MS0 ell PP3MSA00

[1]PROFIT project FIT-1505000-2002-244. The development of tools and resources presented here has also been funded by XTRACT-2 (BFF2002-04226-C03-03).

[2]URL:http://clic.fil.ub.es/index_en.shtml

[3]URL:http://www.talp.upc.es/TALPAngles/index.html

[4]*The secretary general of ERC has taken part at noon in an act in front of the tax office in Tarragona in order to claim a fair funding for Catalonia.*

```
secretari secretari_general LI-NCMS000 secretari NCMS000
general secretari_general LF-NCMS000 general AQ0CS0 general NCCS000 general NCMS000
d' de SPS00
ERC ERC NP00000
ha ha I Ha NCMS000 haver VAIP3S0 hectàrea NCMN000
participat participar VMP00SM participat AQ0MSP
aquest aquest DD0MS0 aquest PD0MS000
migdia migdia NCMS000
en en DA0MS0 en PP3CN000 en SPS00
un 1 Z un DI0MS0 un DN0MS0 un PI0MS000 un PN0MS000
acte acte NCMS000
davant davant NCMS000 davant RG davant SPS00
la el DA0FS0 ell PP3FSA00 la I la NCMS000 La NCMS000
Delegació_d'_Hisenda_de_Tarragona Delegació_d'_Hisenda_de_Tarragona NP00000
per per NCMS000 per SPS00
reclamar reclamar VMN0000
un 1 Z un DI0MS0 un DN0MS0 un PI0MS000 un PN0MS000
finançament finançament NCMS000
just just AQ0MS0 just NCMS000 just RG
per per NCMS000 per SPS00
Catalunya Catalunya NP00000
```

After the morphological analysis, a constraint-based probabilistic tagger (RELAX) selects the most correct lemma-tag pair according to the near context (it mainly uses the preceding and the following word). Constraints have been automatically infered from the CLiC-TALP-CAT corpus, whose desambiguation was manually validated, and, in order to improve the accuracy, some handwritten rules have been added, mainly refering to the lemma ambiguity as well as to the subcategory one. Once the sentence has been disambiguated, the tagged text is as follows:

```
El el DA0MS0
secretari_general secretari_general NCMS000
d' de SPS00
ERC ERC NP00000
ha haver VAIP3S0
participat participar VMP00SM
aquest aquest DD0MS0
migdia migdia NCMS000
en en SPS00
un un DI0MS0
acte acte NCMS000
davant davant SPS00
la el DA0FS0
Delegació_d'_Hisenda_de_Tarragona Delegació_d'_Hisenda_de_Tarragona NP00000
per per SPS00
reclamar reclamar VMN0000
un un DI0MS0
finançament finançament NCMS000
just just AQ0MS0
per per SPS00
Catalunya Catalunya NP00000
```

Finally, the chunking with TACAT and a context free grammar for Catalan of about 1920 handwritten rules provides us with the following analysis:

```
S_[
sn_[ espec-ms_[ El_da0ms0 ]
     grup-nom-ms_[ secretari_general_ncms000
                   sp-de_[ prep_[ d'_sps00 ]
                           sn_[ grup-nom-fp_[ ERC_np00000 ] ] ] ] ]
grup-verb_[ ha_vaip3s0 participat_vmp00sm ]
sn_[ espec-ms_[ aquest_dd0ms0 ]
     grup-nom-ms_[ migdia_ncms000 ] ]
     sp_[ prep_[ en_sps00 ]
          sn_[ espec-ms_[ un_di0ms0 ] grup-nom-ms_[ acte_ncms000 ] ] ]
     sp_[ prep_[ davant_sps00 ]
     sn_[ espec-fs_[ la_da0fs0 ]
          grup-nom-fs_[ Delegació_d'_Hisenda_de_Tarragona_NP00000 ] ] ]
     sp_[ prep_[ per_sps00 ] S-NF-C_[ infinitiu_[ reclamar_vmn0000 ] ] ]
     sn_[ espec-ms_[ un_di0ms0 ] grup-nom-ms_[ finançament_ncms000
                                    s-a-ms_[ just_aq0ms0 ] ] ]
     sp_[ prep_[ per_sps00 ] sn_[ grup-nom-fp_[ Catalunya_np00000 ] ] ] ]
```

The construction of the treebank itself has consisted of the manual embedding of the chunks as well as of the functional tagging.

3. Syntactic Annotation

The task of manually building a treebank requires a tool for facilitating annotators' work. After looking for different freely available interfaces, we decided to use the **AGTK** toolkit set up by the Pennsylvania University (Cotton and Bird, 2000). The main advantage was that it could easily accept our chunker output as well as our large tagset. Figure 2 shows this interface. The main utilities of such an interface are that it allows to move, remove, adjoin and add nodes and tags; it also lets us split and merge sentences and words. Crossing branches is not allowed, so discontinuous constituents are given special tags, as we will comment further on.

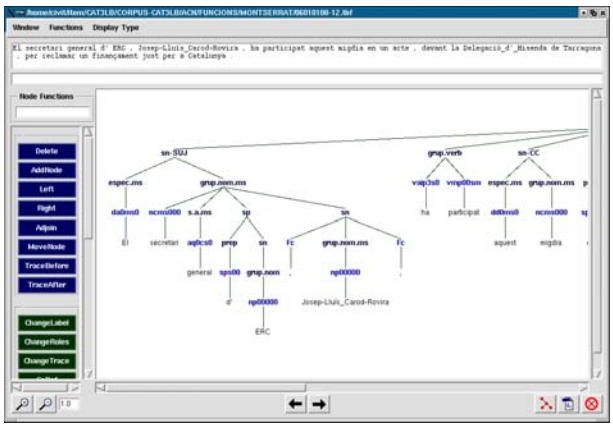

Figure 2: AGTK Interface

3.1. Annotation process

In a first step, 25.000 words were syntactically annotated in parallel (constituent annotation) by two linguists. This first annotation was then used, on the one hand, to refine the annotation criteria and, on the other, to enlarge the annotation guidelines previously established[5]. The comparison between the two annotations gave the results shown in table 1, in which **LP** stands for *labelled precision*; **BP** for *bracketed precision* and **CB** for *consistent bracketing*[6].

LP	0.876478
BP	0.90953004
CB	0.943214
same-lenght Sentences	
LP	0.9198125
BP	0.93964505
CB	0.96512s0

Table 1: Annotators' agreement

One of the main sources of disagreement among the annotators was whether to consider as a single word certain complex structures such as *des que* ('since'), *donar lloc a* ('give rise to') and so on. Annotators adopted different criteria when labelling and bracketing these units. This affected the lenght of the sentences: if such expressions were taken as multi-words, there were fewer terminal elements (words) in the sentence than if they were taken separately. Since our measures take into account the starting and finishing points of each constituent in the sentence, the fact that the length of the sentence varied implied a substantial decrease of the agreement measures. This issue has been accurately analysed and very strict criteria to deal with multiwords were established in the guidelines. However, our aim has been to evaluate also the agreement figures obtained only for those sentences whose lengths coincided. After the correction of these discrepancies, results improved significantly, even though a full agreement seems impossible.

[5](Valverde et al., 2004) is the last version of the guidelines for the constituent annotation.

[6]As they are used in Parseval.

The remaining corpus has been then annotated by only one linguist.

Once the constituent annotation was finished, we started the functional tagging. The process was the same: annotation in parallel of 10,000 words by two linguists; comparision of the results[7]; discussion of the differences; enlargement of the guidelines[8]; annotation by one linguist on the remaining sentences.

Up to now, the full corpus has been annotated at the constituent level, and half

of it at the functional one. At the first level, annotators spent one hour to annotate ten sentences; while at the second one, they needed the same time to annotate 25 sentences. The average number of words per sentence is 39.

3.2. Constituent annotation

General principles have been defined to be the main guidelines in the constituent annotation process. Firstly, we only mark explicit elements, although we add nodes for elliptical subjects, as in the initial purpose of the project it was settled that anaphoric and coreferential information would be annotated. As for elliptical finite verbs, instead of adding a new node, we mark sentence nodes with a suffix *; that usually happens in coordinated structures, for which the second one is verbless.

Secondly, we follow the constituency annotation scheme instead of the dependency framework, since this is the framework that best matches Catalan morphosyntactic features.

Thirdly, we do not alter the surface word order so as to maintain pragmatic information, even though it implies to face the problem of discontinuous elements. We do not consider *verbal phrase* as a node containing the verb and its complements, but as a node containing only simple verbal forms (*considera*, '(he/she) considers') and complex ones (*ha participat*, '(he/she) has participated'; *ha estat considerat*, '(he/she/it) has been considered').

Cases of discontinuity have been dealt with in two different ways: some of them at the constituent level, and the others at the functional one. Discontinuity dealt with in the first level[9] is mostly related to the noun phrase and involves a noun complement which is separated from the head by a (verbal) complement, like in the sentence *Altae és l'aposta de banc de gestió de fortunes de Caja_Madrid en el qual l'entitat pretén guanyar-se una quota de mercat*[10] in which the relative clause *en el qual ...* depends on the noun *banc* but is separated from it by the complement *de Caja_Madrid*. In this case, we add an index **.1** to both elements involved in the discontinuity, so that the resulting analysis is as shown in figure 3.

As for concrete aspects of the annotation, we distinguish among different types of clauses: finite and non-finite

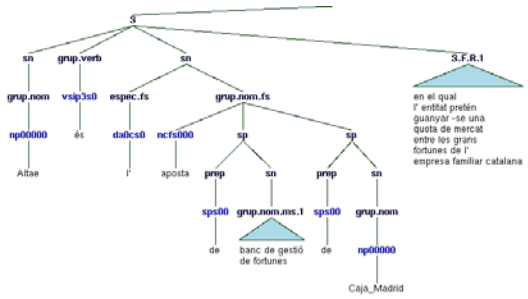

Figure 3: Constituent discontinuity-1

ones, on the one hand, and, on the other, completive, relative and adverbial ones. Adverbial clauses, moreover, are splitted into two groups: those considered as having a function as verbal complement (namely those meaning time, place, cause, purpose or manner) and those considered to be adjuncts of the verb (conditional, concessive, comparative and consecutive ones).

As it occurs in the PennTreeBank, we do use an equivalent to the PRN node for nodes that, generally speaking, do not belong to the sentence but to the discourse. In our case the tag is INC and mainly appears with reported speech.

We pay special attention to the treatment of coordination structures: we consider coordinated elements to be equivalent in the syntactic structure, so they are represented as siblings, which means that there is no head in such constructions. Shared complements are another issue related to coordination (i.e.: complements shared by two or more verbs); in these cases our solution is to adjoin the complement to the coordinated node.

Finally, we should point out that since punctuation marks are considered one among other elements of the sentence and receive a pos-tag, criteria to deal with them have to be clearly established, especially because they play a crucial role in the delimitation of the constituents (bracketing). We distinguish two main kinds of punctuation marks; those having a parenthetic role and those acting as a delimiter. The former appear as the first and the last element of the constituent, while the latter is the first element.

3.3. Functional tagging

Only daughter nodes of sentences and clauses are given a functional tag (i.e. we do not deal with noun complements). Given tags are subject, direct or indirect object, prepositional complement, adverbial complement, etc. and only surface functions are marked; that means that, for instance, we do not tag the subject of the infinitives depending on a control verb.

We have established a set of 15 basic tags, in order to cover all syntactic functions, and then, given specific marks (tag suffixes) to some of them in order to annotate specific cases of these functions. All in all, the total amount of tags at this level is 55.

The remaining case of discontinuity is dealt with in the functional tagging. An example of such a case is *només en queda un*[11] in which "en" and "un" are the two elements of the subject. In these cases, we add the suffix **.d** to the

[7] We do not have exact figures of the annotators' agreement, but they reached about 90% in the first sentences and more than 95% in the lastest ones.

[8] (Civit et al., 2004) is the guidelines for the functional tagging.

[9] See section 3.3. for discontinuity at the functional level.

[10] 'Altae is the bet for fortune managing of Caja_Madrid in which the company wants to gain market share'.

[11] Literally: 'only PRO there is one'; 'there is only one left'.

functional tag. This suffix must appear at least twice, one in each of the members of the splitted constituent.

Another case of discontinuity appears in relative or interrogative clauses, in which the relative (or interrogative) pronoun of a (non-)finite clause raises to the first position of the sentence, like in the sentence in figure 4: *dels pagesos que hi vulguin anar*[12], in which the selected locative complement (*hi*) belongs to the non-finite clause *anar* but appears before the main verb. For these cases, the functional tag has a suffix **.F** or **.NF** (depending on the type of the clause -finite or non-finite-) and the whole tag must be read as follows: *complement of the first finite or non-finite clause to the right*.

Figure 4: Constituent discontinuity-2

Sometimes a functional element appears two times (is repeated in the sentence). It usually happens with direct an indirect objects, when the phrase goes before the verb and it has to be repeated by a clitic, like in *El rànquing l' encapçala la final de la Champions_League*[13], in which the direct object appears twice at the beginning of the sentence (see figure 5), and the tag for the direct object has the suffix **.r**.

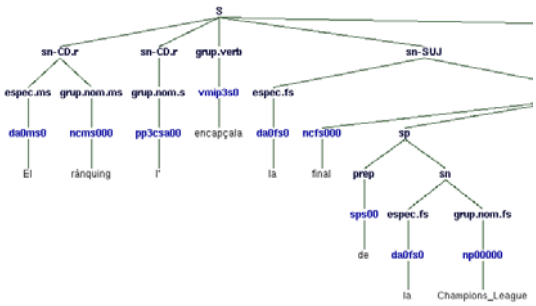

Figure 5: Double functions

One of the most controversial points related to functional tagging has been the distinction between prepositional complements selected or not by the verb. Linguistic criteria are not unanimous, especially those concerning the obligatoriness of the complement. It usually happens that locative complements are mandatory. Bearing in mind the state of the art about this point, we decided to give the adverbial tag (**-CC**) to those elements being optional, whilst the function tag **-CREG**, standing for 'selected PP' is used for the mandatory complements, no matter whether they are locative or not.

4. Conclusions and further work

We have presented different tools and resources developed so far for Catalan and paid special attention to the Catalan Treebank (**Cat3LB**): a morphological analyser (MACO+), a tagger (RELAX), a chunker (TACAT), a context free grammar and a manually validated corpus with morphosyntactic annotation (CLiC-TALP-CAT). All this resources are free for research purposes[14] and are intended to encourage linguistic and computational research on Catalan.

As further work and within the 3LB project, a subset of 10,000 words of the Treebank is being semantically annotated with Catalan-EuroWordNet (Benítez et al., 1998) and will also be freely available.

5. References

Atserias, J., J. Carmona, I. Castellón, M. Civit S. Cervell, L. Màrquez, M.A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé, and J. Turmo, 1998. Morphosyntactic analysis and parsing of unrestricted spanish text. In *Proceedings of the First Conference on Language Resources and Avaluation*. *LREC'98*. Granada.

Benítez, L., G. Escudero, M. López, G. Rigau, and M. Tailé, 1998. Methods and tools for building the catalan wordnet. In *Proceedings of the ELRA Workshop on Language Resources for European Minority Languages, at the First Conference on Language Resources and Avaluation*. *LREC'98*. Granada.

Civit, M., N. Bufí, and M.P. Valverde, 2004. Guia per a l'anotació de les funcions sintàctiques de cat3lb: un corpus del català amb anotació sintàctica, semàntica i pragmàtica. Technical report, CLiC. Available: http://www.clic.fil.ub.es/personal/civit/publicacions.html.

Cotton, S. and S. Bird, 2000. An integrated framework for treebanks and multilayer annotations. In *Proceedings of the Second International Conference on Language and Evaluation LREC-2000*. Athens, Greece.

Monachini, M. and N. Calzolari, 1996. Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. a common proposal and applications to european languages. Technical report, EAGLES. Available: http://www.ilc.pi.cnr.it/EAGLES96/browse.html.

Valverde, M.P., M. Civit, and N. Bufí, 2004. Guia per a l'anotació sintàctica de cat3lb: un corpus del català amb anotació sintàctica, semàntica i pragmàtica. Technical report, CLiC. Available: http://www.clic.fil.ub.es/personal/civit/publicacions.html.

[12] Lit: 'from farmers who there want to go'; translation: 'from farmers who want to go there'

[13] Cat:
El rànquing-CD l'-CD encapçala la final de la Champions_League
Lit.: 'The ranking-CD PRO-CD heads the final-SUBJ of the Champions League'
translation: 'the final of the Champions League heads the ranking'

[14] The mail contact is civit@clic.fil.ub.es