CAT3LB: a Treebank for Catalan with Word Sense Annotation

Montserrat Civit, Núria Bufí, M. Pilar Valverde
CLiC Centre de Llenguatge i Computació – Universitat de Barcelona
Adolf Florensa s/n (Torre Florensa) 08028 Barcelona
{civit, nuria, pilar}@thera-clic.com

1 Introduction

It is widely admitted that Treebanks constitute a crucial resource both to develop NLP applications and to acquire linguistic knowledge about how a language is used. As for minority languages, a new parameter has to be taken into account when we analyse their normalization degree: that of their presence or their absence in Internet and the number and kind of computational resources and tools it processes.

In this paper[1] we present the work done in the building of CAT3LB, a Treebank for Catalan with word sense annotation, which is part of the 3LB Project[2]. The paper is organised as follows: section 2 deals with previous processes (morphological annotation, tagging and chunking); section 3 deals with the syntactic annotation itself (constituents and functions); section 4 deals with word sense annotation; and section 5 presents some conclusions.

2 Previous processes

CLiC-UB [3] and TALP-UPC [4] groups have developed so far a framework for the automatic processing of Catalan and Spanish, based in a pipeline structure[7]. Firstly, the raw text is morphologically analysed with MACO (see section 2.1); secondly,

[1]We would like to thank the three reviewers for their useful and interesting comments about the abstract: they have helped us a lot in the writing of the full paper.

[2]This research has been partially funded by the Spanish Research Department: X-Tract2 (BFF2002-04226-C03-03) and PROFIT (FIT-150500-2002-244).

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

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

it is disambiguated with RELAX (see section 2.2); and finally, it is chunked with TACAT and a handwritten grammar for Catalan (see section 2.3).

2.1 Morphological Analysis

MACO is a Morphological Analyser for Catalan, Spanish and English that provides both lemma(s) and POS-tag(s) for each word and whose output has the following form:

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

The tagset for Catalan codifies 13 part-of-speech categories (noun, verb, adjective, adverb, pronoun, determiner, preposition, conjunction, interjection, dates, punctuation marks, numbers and abbreviations) as well as subcategories and morphological features, as it is proposed by Eagles [13]. The total amount of tags is 321[5].

2.2 Morphological Tagging

Once the text has been morphologically annotated, RELAX, a constraint-based probabilistic tagger [17], selects the best pair lemma-tag; Relax is trained from manually annotated texts and it allows the introduction of manually written constraints. The accuracy of the output varies between 94-96%, but it is increased up to 95-97% after the introduction of handwritten constraints [11].

2.3 Chunking

Finally, the chunking is done with TACAT [4] and a context free grammar for Catalan of about 1920 handwritten rules. Catalan has a rich inflexional morphology, so the concept of chunk can be extended (i.e. we use a larger conception of chunk than Abney's, [2], [3]): the grammar puts together words if, according to their form, one can be sure they go together. For instance, a noun phrase may include:

$$[(Determiner) + (adjective) + head + (noun|AdjP|PP_{de})]$$

that is: an (optional) determiner in a pre-head position; an (optional) adjective before the head; and another element after the noun: either another noun, or an adjective or a prepositional phrase headed by the preposition *de*[6].

[5] They can be found, with a large explanation, at http://clic.fil.ub.es/doc/categorias_eagles_cat03.htm

[6] Meaning 'of' or 'from'. This PP was included in the nominal chunk after a detailed analysis of about 300 examples: the PP was almost always depending on the immediately preceding noun.

This extended conception of chunk is similar to that proposed in [15] and largely reduces the annotation time, even if it is not error free (this chunking produces some errors in the analysis of the PP_{de} attachment (1-2%) but we consider that it is assumable if, in contrast, it largely reduces the annotator's work).

3 Syntactic annotation

Catalan is a romance, pro-drop language in which the constituent order is quite free and, within the constituents, the word order is quite fixed (i.e. in the noun phrase, for instance, adjectives can precede or follow the noun, but the most frequent case is the postposition; the determiner is always the first word in such a phrase; it is extremely rare to find two adjectives preceding the noun head; all relative clauses and prepositional phrases follow the head; etc.). Linguistic tradition deals with Catalan in terms of constituents ([14], [5]) and we do so; moreover, in the sentence, movement phenomena are basically related to constituents.

In order to do the syntactic annotation, we used the AGTK interface [12] developed at the University of Pennsylvania. It has been slightly modified in order to allow both the processing of special characters and the processing of xml text format[7].

3.1 Constituents

In a first step, 25.000 words[8] were syntactically annotated in parallel 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 [9]. 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*[10].

One of the main sources of disagreement was whether to consider as a single word certain complex structures, like *posar èmfasi*[11]. As annotators adopted different criteria, the length of the final sentence was different from one annotation to the other. Since our agreement 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 results. This issue was accurately analysed and very strict criteria were established in the guidelines to deal

[7] Trees are stored in two formats: bracketed text, like the Penn TreeBank and xml format.

[8] This corresponds to 640 sentences and the average number of words per sentence is 39.

[9] [19] is the last version of the guidelines for the constituent annotation.

[10] As they are used in Parseval.

[11] To emphasise.

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

Table 1: Annotators' agreement (1)

with multiwords. In order to evaluate the annotators' agreement for those first sentences, we also carried out the evaluation taking into account only the sentences having the same length.

As for concrete aspects of the annotation, we would like to point out some of the most significant issues: types of sentences, coordinated structures and discontinuity. As the Cat3LB treebank has been developed within a larger project (**3LB**) and after the development of the Spanish Treebank (Cast3LB) [9], [10], we have take advantage of the previous annotation process.

The root node of sentences is always **S**, standing for *sentence*. If the sentence has no verb, then the tag is **S***. Regarding clause types, we distinguish finite (**S.F.**) and non-finite clauses (**S.NF.**), on the one hand, and, on the other, completive (**S.F.C, S.NF.C**), relative (**S.F.R**) and adverbial ones (**S.F.A, S.NF.A**). Finite adverbial clauses, moreover, are splitted into three groups: those considered as being a verbal adjunct (namely those meaning time, place, cause, purpose or manner **S.F.A**), those considered to be adjuncts of the predicate (conditional, concessive and consecutive ones **S.F.ACond, S.F.AConc, S.F.ACons**) and those being adjuncts of a noun or an adjective, the comparative clauses (**S.F.AComp**).

We pay special attention to the treatment of coordinated structures: we consider coordinated elements to be equivalent in the syntactic structure[12], 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.

Cases of discontinuity have been dealt with in two different ways: some of them at the constituent level, and the others at the function one (see section 3.2). Discontinuity dealt with in the first level is mostly related to the noun phrase and

[12]The solution in [1] is completely different, since they consider the first element to be the head in the coordinated nodes.

involves a noun complement which is separated from the head by a (verbal) complement; in this case, the separated complement formally depends, in the representation on the nearest S node, but the **.1** index marks where it must be interpretated. An example of this situation appears in the sentence *en detectar-se la presència d'un brot infecciós a principis del mes de maig que va afectar 12 malalts*[13] in which the relative clause *que va afectar 12 malalts* depends on the noun *brot* but is separated from it by a verbal complement *a principis del mes de maig*. In this case, we add an index **.1** to both elements involved in the discontinuity, so that the resulting analysis is as it appears in figure 1.

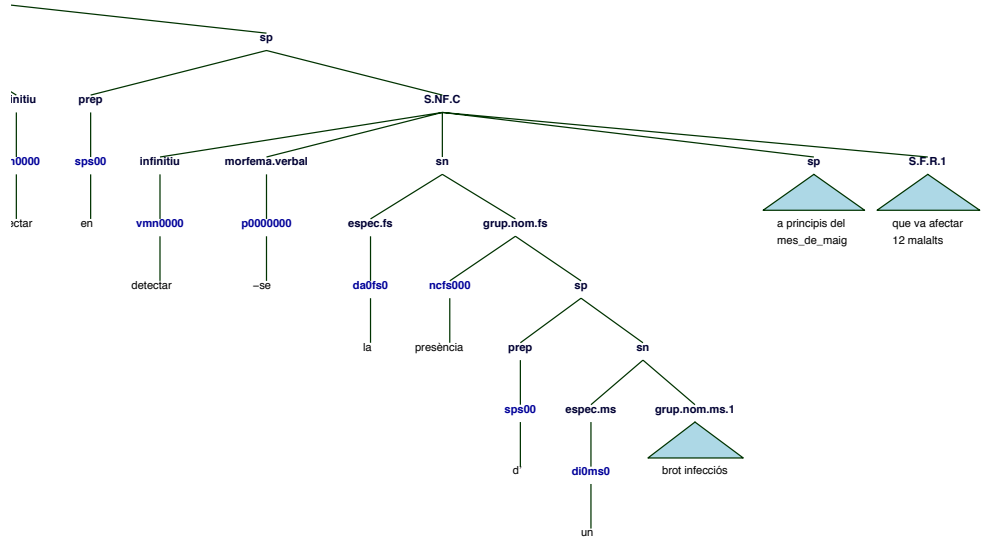

Figure 1: Constituent Discontinuity (1)

3.2 Functions

We have extended our constituency-based scheme with the annotation of grammatical functions (as it was done in the Susanne Corpus [18] or in the Penn Treebank II [16]), that is, by adding functional tags to the phrase structure annotation[14].

[13] When detecting the presence of an infectious outbreak at the beginning of May that infected 12 sick people.

[14] [8] is the guidelines for the functional annotation

A quantitative analysis of the annotators' agreement was done, as it was for the constituent annotation. The results are shown in table 2. In this case, we only consider the labelled precision, because the annotators worked over the previous constituent annotation, so the bracketing was the same. This evaluation was done in two different times: at the first one, the guidelines were not yet complete, while at the second they were.

First phase: 698 sentences	
LP	0.9009
Second phase: 45 sentences	
LP	0.94915254

Table 2: Annotators' agreement (2)

Most of the discrepancies were due to errors in the annotation (i.e. one annotator forgot to put the functional tag or did not apply correctly what was said in the guidelines).

Only daughter nodes of sentences and clauses are given a functional tag (i.e. we do not deal with noun complements). We have established a set of 14 basic tags (see table 3), 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 58. Basic tags are shown in table 3.

Tag	**Gloss**	**Tag**	**Gloss**
-SUJ	subject of a finite verb form	-AO	Sentence Adjunct
-CD	Direct Object	-ET	Textual Element
-CI	Indirect Object	-MOD	Modifier
-ATR	Attribute	-PASS	Passive Mark
-CPRED	Predicative Complement	-IMPERS	Impersonal Mark
-CREG	Prepositional Object	-VOC	Vocative
-CAG	Agent		
-CC	Circumstance		

Table 3: Basic Functional Tagset

The remaining cases of discontinuity are dealt with in the functional tagging. There are two of such cases. The first case is related to clitics and the second is related to raising movement.

When the direct object (with ergative verbs it may happen with the subject too) is an undetermined noun phrase (i.e. a noun phrase with an indefinite article or a quantifier), the substitution by the clitic is partial, and only the noun is replaced by the clitic, but not the determiner; so the direct object is splitted into two elements, one before and the other after the verb. For such cases, we have created a special tag suffix (**.d**), which appears in both the two elements. Figure 2 shows one of these cases. The sentence is *dels quals només se'n conserven dos*[15]

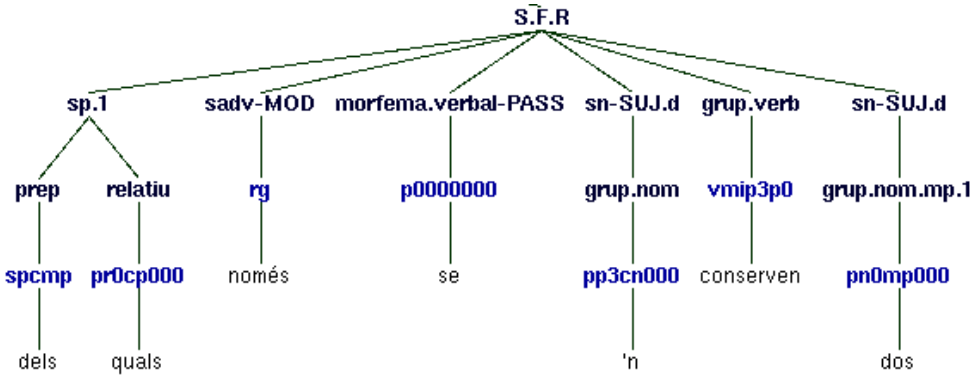

Figure 2: Constituent Discontinuity (2)

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: *dels pagesos que hi vulguin anar* (figure 3)[16], 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*.

In Catalan it is possible for a complement to appear twice in the sentence. It usually happens with direct an indirect objects (but also with other verb complements), when the phrase goes before the verb and it has to be repeated by a clitic[17]. This is related to the inversion of constituents in the sentence: the most usual word

[15]Literal: from which only [passive mark] [clitic] survive two
Translation: from which only two survive.

[16]Lit: from farmers who [locative-clitic] want to go; translation: from farmers who want to go there

[17]If there is no repetition the sentence is considered to be ungrammatical.

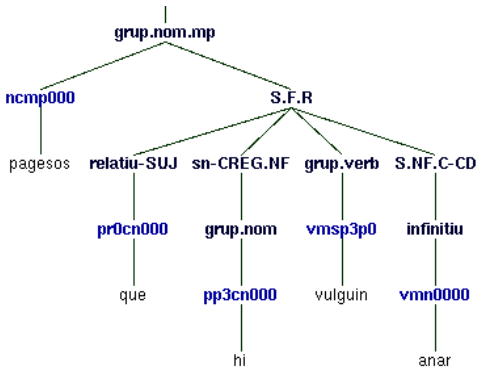

Figure 3: Constituent Discontinuity (3)

order in Catalan is SVO, and when it is inverted (OVS) we need to mark the inversion, so the complement is repeated by a clitic. In these cases we add a suffix **.r** to the function tag. An example of such phenomenon is shown in the sentence *El rànquing l' encapçala la final de la Champions_League*[18], in which the direct object appears twice at the beginning of the sentence (see figure 4).

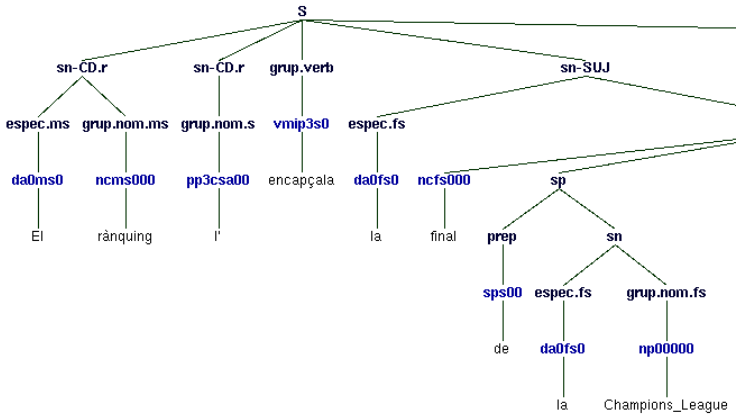

Figure 4: Doubled functions

One of the most controversial points related to functional tagging has been the

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

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. This clearly appears when the answer to the question *Anirem al cine demà?*[19] has to contain the locative clitic *hi*: *hi anirem*[20]. Bearing in mind the state of the art about this point, we decided to give the adverbial tag (**-CC**) to those elements being optional, while the function tag **-CREG**, standing for 'selected PP' is used for the mandatory complements, no matter whether they are locative or not.

4 Semantic annotation

In a last step, a subset of the corpus of 10,000 words has been annotated with CatalanWordNet. Only nouns, verbs, and adjectives receive a semantic tag. In the annotation process words were annotated throughout the whole corpus, so as to ensure the consistency of their annotation.

The total amount of nouns in the subset of the corpus is 841 lemmas (some of them appearing 33, 30, 28 times; others appearing only once), 380 adjectives and 403 verbs. The most frequent nouns were *grup, govern, any, empresa, president*[21]; the most frequent adjectives were *català, nou, passat, polític, socialista*[22]; finally the most common verbs were *tenir, estar, presentar, poder, fer, donar*[23]. Some verbs were not annotated when they were the auxiliary for the compound tenses o complex verbal forms.

In order to do the annotation, we took a version of the EuroWordNet 1.5, that of December 2002, and built an interface to help the annotators: 3LB-SAT [6]. CatalanWordNet has 28,575 synsets (20,260 for nouns, 4,415 for adjectives and 3,900 for verbs). The ambiguity average is 1.790 senses per lemma if we considers all variants, and 3.182 if we consider only ambiguous lemmas (i.e. lemmas appearing in two or more synsets).

The main problem in the semantic annotation was that CatalanWordNet is incomplete and has not been extensively revised. For instance, for the word *president*, which referred in most of the sentences to the president of the Catalan Parliament or the president of a football team, it was impossible to assign a sense, because in CatalanWordNet there are only presidents for Republics, companies, meetings or the United States. Figure 5 shows the interface with the possibilities

[19] Will we go to the cinema tomorrow?
[20] We [clitic] will.
[21] Group, government, year, enterprise/business, president
[22] Catalan, new/nine, last/passed, political, socialist
[23] to have/to own, to be, to present/to introduce, to be able to/can, to do/to make, to give

for the word *president* displayed.

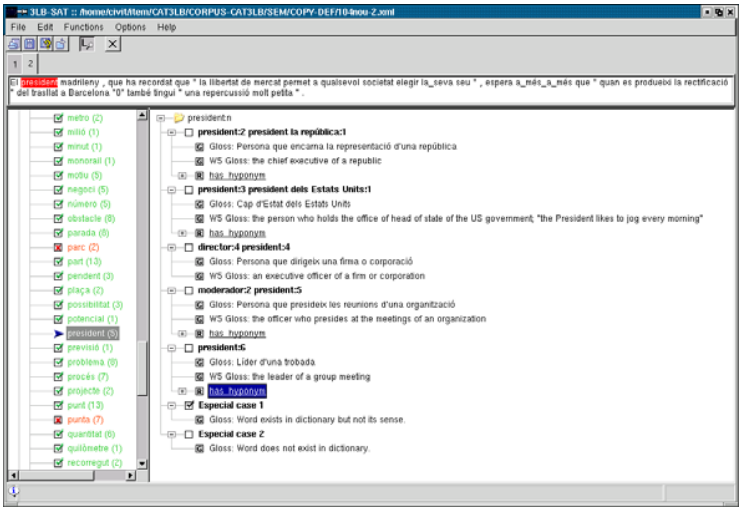

Figure 5: Semantic Annotation Tool

We added two more possibilities to the word sense annotation: **EC1** and **EC2**. EC1 stands for those cases in which the word appears in CatalaWordNet but not its sense; EC2 was thought to mark those words that did not appear in the CatalanWordNet.

The word sense annotation was done semiautomatically; on the one hand, the tag EC2 was assigned automatically if the word did not appear in CatalanWordNet; on the other hand, words being monosemous in CatalaWordNet were given the sense in an automatic way, but this assignation was manually checked because the given sense could not be the right one. The rest of the annotation was done manually. It was not possible to do any preliminary automatic annotation because there is no information in CatalanWordNet about the most frequent sense of the words.

5 Conclusions

We have presented the development of Cat3LB, a Treebank for Catalan. We have shown the main issues in both the syntactic and semantic annotation processes. One of the open issues, now, is to start the revision of the CatalanWordnet: we plan to develop it in parallel with the corpus annotation in order to add the synsets that do not appear at present, but also in order to modify its structure by removing

unnecessary synsets or compact some others. We think that this parallel work is the best both for the corpora annotation process and the enrichment of the semantic net.

References

[1] A. Abeillé, F. Toussenel, and M. Chéradame. Corpus le Monde. Annotation en constituants. Guide pour les correcteurs. Technical report, LLF, UFRL, 2002. dernière mise à jour: 10-juillet-2002.

[2] S. Abney. Parsing by Chunks. In R. Berwick, S. Abney, and C. Tenny, editors, *Principle-Based Parsing*. Kluwer Academic, 1991. available: http://www.sfs.nphil.uni-tuebingen.de/ abney/.

[3] S. Abney. Partial Parsing via Finite-State Cascades. In *Proceedings of the ESSLLI'96 Robust Parsing Workshop*, 1996. availible: http://www.sfs.nphil.uni-tuebingen.de/ abney/.

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

[5] A. Bel. Les funcions sintàctiques. In J. Solà, M.R. Lloret, J. Mascaró, and M. Pérez, editors, *Gramàtica del català contemporani*, chapter S-2, pages 1075–1147. Empúries, 2002.

[6] E. Bisbal, A. Molina, L. Moreno, F. Pla, M. Saiz-Noeda, and E. Sanchís. 3LB-SAT: Una herramienta de anotación semántica. In *Procesamiento del Lenguaje Natural*, number 31, pages 193–199, Alcalá de Henares, 2003.

[7] J. Carmona, S. Cervell, L. Màrquez, M.A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé, and J. Turmo. An Environment for Morphosyntactic Processing of Unrestricted Spanish Text. In *Proceedings of the First Conference on Language Resources and Avaluation. LREC'98*, pages 915–922, Granada, 1998.

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

[9] M. Civit and M.A. Martí. Design Principles for a Spanish Treebank. In *Proceedings of the First Workshop on Treebanks and Linguistics Theories (TLT2002)*, pages 61–77, Sozopol, September 2002. available: http://clic.fil.ub.es/personal/civit.

[10] M. Civit, M.A. Martí, B.Navarro, N. Bufí, B. Fernández, and R. Marcos. Issues in the Syntactic Annotation of Cast3LB. In *Proceedings of the LINC03 Workshop*, Budapest, 2003. available: http://clic.fil.ub.es/personal/civit.

[11] L. Cots. Restriccions manuals de desambiguació en el corpus CLiC-TALP-CAT. Master's thesis, Universitat de Barcelona, Dpt. de Lingüística General, 2004.

[12] S. Cotton and S. Bird. An integrated Framework for Treebanks and Multilayer Annotations. In *Proceedings of the Second International Conference on Language and Evaluation LREC-2000*, Athens, Greece, 2000.

[13] EAGLES. Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A common Proposal and Applications to European Languages. EAG—CLWG—MORPHSYN/R, 1996. available: http://www.ilc.pi.cnr.it/EAGLES96/browse.html.

[14] M. L. Hernanz. L'oració. In J. Solà, M.R. Lloret, J. Mascaró, and M. Pérez, editors, *Gramàtica del català contemporani*, chapter S-1, pages 993–1073. Empúries, 2002.

[15] H. Kermes and S. Evert. Text analysis meets corpus linguistics. In *Proceedings of the Corpus Linguistics 2003*, Lancaster, UK, 2003.

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

[17] L. Padró. *A Hybrid Environment for Syntax-Semantic Tagging*. PhD thesis, Software Department (LSI). Technical University of Catalonia (UPC), 1998.

[18] G. Sampson. *English for the Computer. The SUSANNE corpus and Analytic Scheme*. Clarendon Press, Oxford, 1995.

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