AnCora: Multilingual and Multilevel Annotated Corpora

Maria Antònia Martí, Mariona Taulé, Manu Bertran and **Lluís Màrquez**

CLiC-UB (Centre de Llenguatge i Computació, Universitat de Barcelona)
TALP, Software Department, Universitat Politècnica de Catalunya
{amarti, mtaule}@ub.edu, {mbertran,lluism}@lsi.upc.edu

Introduction

To have at our disposal linguistic resources with morphosyntactic and semantic information, either lexicons or tagged corpora, appears to be an obvious necessity for most –if not all– natural language processing (NLP) applications. Furthermore, annotated corpora also constitute a crucial resource to acquire or infer linguistic knowledge about how languages are used. In this line, it is widely accepted that linguistically annotated corpora are a very useful resource for computational and linguistic analysis of languages. Thus, annotated corpora become an essential reference for any computational tool, technique or application applied to unrestricted text analysis, and are especially necessary for machine learning systems. Obviously, the more explicit linguistic information they contain, the more interesting and useful they are. Moreover, it is important not only to provide corpora annotated at different levels of linguistic analysis (e.g. morphological, syntactic, semantic, pragmatic annotation) but it is also fundamental to guarantee the quality of these annotations. Ultimately, the annotation quality of the corpora determines to a large extent the effectiveness and quality of the NLP systems and techniques based on them (Màrquez et al., 2004). In order to guarantee the quality of the annotated corpora, it is necessary to establish an appropriate tagging methodology to achieve a systematic and consistent tagging process. Taking all these elements into consideration, we present two 500,000-word multilevel annotated corpora –one for Catalan (AnCora-CAT) and one for Spanish (AnCora-ESP)– developed in last years within the framework of several projects. [1]

In this paper we describe the methodology and the general criteria established to systematize the annotation process to build the AnCora corpora as well as the information contained at each linguistic level of analysis. The development of these basic resources for Catalan and Spanish constitutes a primary objective, since there is a lack of this kind of resources for these languages. The aim was to build reference corpora for both languages and define a consistent methodology that could be followed in further annotations.

The AnCora corpora were annotated at different levels of linguistic description: the whole Catalan corpus is annotated with morphological, syntactic, and semantic information; as for Spanish, the morphological and syntactic levels are already completed, while the semantic annotation covers 40% of the corpus (~200,000 words). The annotation process was carried sequentially from lower- to upper-level layers of linguistic description (i.e. first morphology, next different levels of syntactic description, and finally semantic annotation). The annotation was performed manually, semi-automatically, or fully automatically, depending on the corresponding linguistic information. First, both corpora were morphologically tagged and disambiguated using automatic linguistic tools (Civit 2003; Civit & Martí 2004a), and afterwards were manually checked throughout the syntactic annotation. An automatic shallow parser was also applied to recognize base constituents. Shallow parsing served as starting point for handling the annotation at the level of full syntax. The syntactic annotation involved labelling manually constituents and syntactic functions (Civit & Martí 2004b). The dependency relationships were also automatically derived from the constituent annotation to produce a dependency-based version of the Treebank (Civit et al., 2006). With respect to the semantic annotation, the corpora were annotated at different levels: 1) basic syntactic functions were tagged in a semiautomatic way with arguments and thematic roles (Taulé et al., 2006a) taking into account the semantic class related to the verbal predicate (Taulé et al., 2006b); 2) WordNet synsets were manually assigned for all nouns in the corpora; and 3) named entities were also manually annotated (Borrega et al., 2007a; 2007b).

This building process implied different checking steps at each tagging stage. The results of the automatic processes were manually checked in a subsequent annotation. This was the case of

[1] The development of these corpora was partially funded by the Catalan Government and the following projects: 3LB (FIT-150-500-2002-244), CESS-ECE (HUM2004-21127), Lang2World (TIN2006-15265-C06-06, http://gplsi.dlsi.ua.es/text-mess/index.php/Portada) and PRAXEM (HUM2006-27378-E).

morphological parsing and semiautomatic role labelling. Regarding manual annotation, some experiments related to inter-annotator agreement were carried out –especially at the syntactic and semantic levels– in order to evaluate the quality of the results. We have considered the score of inter-annotator agreement as a quality criterion. An annotation manual, where the criteria followed to avoid inconsistencies are specified, was developed for each level of annotation.

The current versions of the AnCora corpora were used in several international evaluation campaigns, at CoNLL-2006/2007 and SemEval-2007, concerning different syntactic and semantic NLP tasks. The corpora are freely available for research purposes and can be downloaded from the main website of the AnCora corpora (http://clic.ub.edu/ancora/)

This paper is organized in 6 sections. Firstly, Section 1 gives a description of both corpora. In section 2, the annotation methodology is described. The main principles of the three levels of linguistic annotation –morphological, syntactic and semantic (lexical and sentential)– are presented in sections 3, 4, and 5, respectively. Finally, section 6 concludes and outlines future work on the development of the AnCora corpora.

1. AnCora – ANnotated CORporA

AnCora are two multilingual corpora of 500,000 words each: a Catalan corpus (AnCora-CAT) and a Spanish (AnCora-ESP) one, built in an incrementally way from the previous 3LB corpora (Civit & Martí 2004b). 3LB-CAT and 3LB-ESP corpora are two treebanks of 100,000 words each, corresponding to 4,000 sentences for Spanish and 2,600 for Catalan. Both corpora were tagged automatically with morphosyntactic information (PoS tags) and manually checked. They have been widely used as training corpora for both rule- and learning-based PoS tagging systems. 3LB treebanks were syntactically tagged with constituents and functions in a manual way. 3LB-ESP was created with 75,000 words from *Lexesp* –a Spanish balanced corpus of six million words (Sebastián et al., 2000)– and with 25,000 words from the Spanish EFE news agency. 3LB-CAT consists of 75,000 words from the EFE news agency and 25,000 words from the ACN Catalan news agency.

AnCora is the result of enlarging the 3LB-CAT/ESP corpora up to 500,000 words and enriching them with semantic information at different levels: argument structures, thematic roles, semantic classes, named entities (NE) and noun senses. In this way, 400,000 words were added to each corpus coming from different press sources. 200,000 words from the Spanish EFE news agency[2] and 200,000 words from the '*El Periódico*' newspaper were added to AnCora-ESP. On the other hand, 200,000 words from the Catalan ACN news agency[3] and 200,000 words from the Catalan version of the '*El Periódico*'[4] newspaper were added to AnCora-CAT. This information is summarized in Table 1 for Spanish and in Table 2 for Catalan.

Spanish	Amount	Sources	Anotation levels	Procedure
3LB-ESP	100,000	EFE (25,000) Lexesp (75,000)	PoS	Automatic and manual validation
			Chunking	Automatic
			Syntax	Manual
AnCora-ESP	500,000	EFE (225,000)	PoS	Automatic

[2] http://www.efe.es

[3] http://www.acn.cat

[4] The 200,000-word subset coming from '*El Periódico*' corresponds to the same news articles in Catalan and Spanish spanning from January to December 2000.

		Lexesp (75,000) El Periódico (200,000)	Chunking	Automatic
			Syntax	Manual
			Thematic Roles[5]	Semi-automatic
			Noun senses	Manual

***Table 1**:* The AnCora-ESP corpus in figures

Catalan	Amount	Sources	Annotation levels	Procedure
3LB-CAT	100,000	EFE (75,000) ACN (25,000)	PoS	Automatic and manual validation
			Chunking	Automatic
			Syntax	Manual
AnCora-CAT	500,000	EFE (75,000) ACN (225,000) El Periódico: (200,000)	PoS	Automatic
			Chunking	Automatic
			Syntax	Manual
			Thematic Roles	Semi-Automatic
			Noun senses	Manual

***Table 2**:* The AnCora-CAT corpus in figures

Next sections (2-6) describe the methodology applied to develop both corpora as well as to the kind of annotation that was incorporated.

2. **Methodology**

This section presents the methodology that was followed throughout the process of corpora annotation, the inter-annotator agreement tests as well as the conversion of the constituent treebank into a dependency one.

In the process of corpora annotation we opted for a step-by-step procedure, revising at each step the results of the previous stage. In the case of manual annotation, we computed inter-annotator agreement rates whenever it was possible (e.g. in constituent annotation) in order to assess the quality of the annotation and, indirectly, the appropriateness of the annotation guidelines. With respect to the annotation tools, we used specific tools in-house designed (i.e. MACO, TACAT, TreeTrans and 3LB-SAT, which will be later described) by the involved research groups as well as tools developed by other research groups and that were adapted to meet our needs. Regarding automatic processes, these were revised in the subsequent tagging process or manually checked.

2.1. Annotation processes

The annotation process was carried out sequentially, from the most basic levels of analysis –that is, PoS and chunking– to the most complex ones –namely, full syntax and semantics. Each level of annotation implied checking and completing the previous levels in order to guarantee high quality and minimize the error rate. Each layer of annotation was considered independent from the others. Regarding the degree of automation, we can distinguish three kinds of annotation processes: full automatic, semiautomatic and manual.

[5] Up to now only 200,000 words have been tagged with thematic roles and WordNet synsets.

2.1.1. **Automatic processes**

Automatic processes were used to produce morphological tagging and shallow parsing. With this aim, we employed accurate and high-coverage morphological analyzers, POS taggers –*MACO* (Carmona et al., 1998)– and chunkers –*TACAT* (Atserias et al., 1998)– for Catalan and Spanish. The morphological analyzer, which integrates POS tagging, shows error rates in a range of 2-4% depending on the corpus.

Chunking was an intermediate process with the sole purpose of handling the full syntactic annotation. The chunking process gave as output a flat parse tree with partial constituent analysis. At this level we also solved the analysis of morphological units –such as complex verbal forms– that had not been handled in the previous morphological analysis. We gave preference to a shallow analysis rather than an in-depth analysis so as to assure correctness. We preserved the final quality of these automatic processes by manually checking during the subsequent annotation process –the syntactic one– which was manually performed.

2.1.2. Manual processes

A manual process was adopted for a deep syntactic annotation (constituents and functions), for strong and weak NE classification, and for WordNet *synsets* (senses) assignment to nouns. In order to ensure the quality in the annotation results, a very strict methodology was followed in all manual processes, including annotators' agreement tests. In brief, the basic annotation criteria would firstly be proposed in the guidelines (version 0.1). A minimum of three coders would then annotate the same corpus span following these guidelines. The resulting annotation would be checked and disagreements discussed. The suggested solutions, as overt and exemplified as possible, would then be included in the guidelines in order to guarantee the coherence and consistency of the data contained in the treebank. This process would be repeated with different sets of linguistic units until the annotators' disagreements would decrease to a very low rate, thus ensuring that each layer of annotation had at least 95% inter-annotator agreement. Once this inter-annotator agreement was reached, inter-coder agreement tests stopped, and the annotation process was completed on an individual basis.

The manual checking process was applied for syntactic functions, strong and weak NEs, and for WordNet synsets assignment. In the case of constituents, a direct comparison is not so easy, since no specific measures for the quantitative comparison of the annotators' agreement exist and it is not possible to manually compare the whole constituent structure. For this reason we decided to use what might be considered the basic objective measures, namely the ones defined in the *Parseval* workshops (Black et al., 1991), originally oriented to evaluate wide-coverage syntactic analysers for English, in order to compare the similarity of their results with the reference parse trees (*gold standard*).

Given that we had no gold standard, the comparison process in our case was carried out comparing the results of each annotator against the others. We adapted the *Parseval* measures in order to obtain precision measures on bracketing, labelling, and overlapping. The quantitative evaluation of the agreement (two annotators) was performed in five steps and alongside some of the disagreement problems were solved (see Civit et al., 2003 for further details).

Firstly, once the basic annotation principles had been established (Bufí et al. 2007[6]), 100 sentences were annotated by two coders and the criteria were revised and extended. Secondly, 220 more sentences were annotated, and more details about the adopted system resulted from the

[6] Last version of the guidelines.

discussions on the annotation schema. Thirdly, the previous annotations were revised and compared so as to check both whether the guidelines did not contain ambiguities and whether the annotators were already familiar with the working system. Fourthly, 670 more sentences were annotated (key test). Finally, the fifth step concerned the results of the evaluation over the last 30 sentences. Table 3 shows the results of the quantitative analysis, where LP stands for "labelled precision rate", BP for "bracketed precision rate", and CB for "consistent brackets recall rate".

	LP	**BP**	**CB**
Step 1	0.63359	0.72611	0.81072
Step 2	0.71166	0.80454	0.87124
Step 3	0.76537	0.84762	0.90487
Step 4	0.79222	0.85979	0.90821
Step 5	0.86927	0.90889	0.94958

Table 3: Quantitative evaluation of the constituent annotation

The degree of confidence achieved makes these treebanks well suited to be used as gold-standard given that the degree of consistency is high enough, as can be seen in Table 3. The guidelines cover all the cases that can be found in our corpora and represent the basis for further developments of a full grammar of Spanish and Catalan.

2.1.3. **Semiautomatic processes**

The semantic annotation of verbal predicates was done semiautomatically**.** Firstly, we automatically associated thematic roles with the syntactic functions of each verb in the treebank taking two verbal lexicons as source of information –AnCora-Verb-Cat and AnCora-Verb-Esp (see 2.2 in this section)– where the mapping between syntax and semantics had been previously established by hand. A set of manually written rules automatically mapped part of the information declared in these verbal lexicons onto the syntactic structure, that is, tagging the treebanks with thematic roles and semantic classes. We defined three different types of rules taking into account the kind of information they are based on:

a) Rules based on a specific function or morphosyntactic property. For example, if the predicate has associated the verbal morpheme 'PASS' (passive voice), then its subject has the argument position Arg1 and the thematic role *patient* (SUJ-Arg1-PAT).
b) Rules based on the semantic properties of the predicates. For instance, when predicates are monosemic, the mapping between syntactic function and argument and thematic role as well as the assignment of the semantic class is directly realized. In the case of polysemic verbs, the mapping can be partial because it is only automatically assigned the unambiguous information.
c) Rules based on the type of adverb or prepositional multiword appearing in a specific constituent. For instance, if the prepositional multiword '*a_causa_de*' (*because_of*) or the adverb '*aún*' (*still, yet*) in Spanish, appears in an adverbial complement (function = CC), then it is automatically assigned the argument and thematic role ArgM-CAU (an adjunct argument with the thematic role *cause*) as well as ArgM-TMP (an adjunct argument with the thematic role *temporal*) respectively.

We applied these rules following a decreasing heuristics according to the degree of generality, that is, we applied first the more general rules of type a), secondly the type c) rules and, finally, the type b) rules.

In the automatic annotation process we obtained either full annotations –containing information about the arguments and the thematic roles– or partial annotations with only arguments or thematic roles. Afterwards the results of the automatic annotation were analyzed analytically.

The syntactic functions receiving semantic annotation are: subject (SUJ), direct object (CD), indirect object (CI), prepositional complement (CREG), attribute (ATR), predicative complement (CPRED), and adverbial complement (CC). Sentential adjuncts (AO), vocative (VOC), textual elements (ET) and negation, impersonal and passive marks did not receive any semantic information. The recall of the semantic automatic tagging was in the 56-60% interval, depending on the language (Martí et al., 2007). Afterwards we manually completed the thematic role annotation.

2.2. The AnCora lexicons

The AnCora-VERB lexicons (AnCora-Verb-Cat for Catalan and AnCora-Verb-Esp for Spanish) were obtained deriving all the syntactic schemata in which a predicate appears in the treebank for each sense of each verb. This information allowed declaring manually the mapping from syntactic functions to thematic roles, as well as the corresponding argument position. Figure 1 shows the full information associated with the entry ‘*mejorar’* (sense 01) (‘to improve’) in the AnCora-Verb-Esp lexicon.

mejorar - 01
LSS1.1 (A1)
SUJ Arg0##CAU
CD Arg1##TEM
CC ArgM##TMP/#ADV

EJ: "obligará a mejorar la calidad del ataque"
EJ: "que han mejorado las relaciones laborales"

+ANTICAUSATIVA
LSS2.2 (B2)
SUJ Arg1##TEM
CC ArgM##ADV/para#FIN

EJ: "Por una parte, las técnicas de diseminación han mejorado mucho"
EJ: "el mencionado proyecto de ley sea mejorado para permitir nombres así"

Figure 1*:* Lexical entry of ‘mejorar’ (to_improve) in AnCora-Verb-Esp

In the AnCora-Verb lexicons, each predicate is also related to one or more semantic classes (Lexical Semantic Structures, see section 5.1.1), depending on its senses (LSS1.1 and LSS.2.2 in the example of Figure 1).

The development of the AnCora-Verb lexicons was carried out following the same quality control as the one followed throughout the manual annotation of the corpus: after a first proposal of verb classes and their corresponding theta-roles (see Sections 5.1.2 and 5.1.3), a group of seven trained linguists elaborated a subset of 30 verbal entries. The resulting entries were compared, the disagreements discussed and the verb classes modified when necessary. Disagreements were mainly due to differences in class or theta-roles assignment. This process was applied over several subsets of 30 verbs until no relevant disagreements arose.

2.3. From constituents to dependencies

It is commonly accepted that constituent annotation is richer than that of dependencies, since it contains different descriptive levels with a wide range of variability in the internal structure of constituents. Furthermore, in this kind of annotation, the head of each constituent can be easily inferred from the information contained in the constituents. In turn, dependencies provide an immediate description, without intermediate descriptive levels, because each tree node corresponds to a word. According to Lin (1998), dependency trees allow for more meaningful error measures and comparisons; and further works have confirmed this idea (Beil et al., 2002). Therefore, it is easier to go from a constituent structure to a dependency one: heads can be easily obtained and intermediate levels can be avoided so as to obtain a complete standard dependency representation. In contrast, when going from a dependency structure to a constituent one, the result is a quite flat constituent structure lacking intermediate description levels.

Bearing in main all these considerations, we decided to adopt a constituent annotation and, in a subsequent process, to convert constituent treebanks into dependencies with the hope of enlarging the research on NLP for the two concerned languages (Civit et al., 2006). The conversion of the AnCora treebanks from their original constituent format into dependencies was done automatically but we needed to write manually the *head* and *function tables*. The process was also used to improve the quality of the first annotation and to modify the annotation for further extensions of the treebanks.

On the one hand, the so-called ‘head table’ was created to indicate which of the daughter nodes of a constituent was the head. The goal of the head table was thus to associate each non-terminal tag with either another non-terminal tag or a PoS-tag (the head). Therefore, the subsequent dependencies were simple pairs of elements with no edge labels. Basic assumptions in the head table were that each non-terminal node is a head, and that heads are linguistically based. The format of the head table is as follows:

tag1 = (operator) tag2

where tag1 is the mother and tag2 the daughter. There are three operators in the head table: *rightmost*, *leftmost* and *only_one*. The first two select a tag2 according to its place: the rightmost (or the leftmost) element of a given sequence; while *only_one* works in the cases in which only one element of a given type[7] exists. The head selection is linguistically motivated, that is, the most linguistically-intuitive head was chosen. In addition, there is another crucial element in this table: the order in which daughters are selected as heads. Let us consider the following Catalan verbal forms for the verb ‘cantar’ (to sing):

	Form	Translation	Grammar rule
1	*cantes*	(you) sing	verb
2	*ha cantat*	has sung	aux. + participle
3	*vol cantar*	wants to sing	verb + infinitive
4	*ha de cantar*	has to sing	aux. + preposition + infinitive
5	*està cantant*	is singing	verb + gerund

Figure 2: Simple and complex Catalan verbal forms, their translation, and the grammatical rules they convey

and the head rules:

[7] Other conventions in the head table are: < stands for the beginning of a pos-tag; < > for the whole pos-tag; { stands for the beginning of a constituent tag, while full constituent tags are written straightforward.

r1 *grup.verb* = *rightmost* infinitiu
r2 *grup.verb* = *rightmost* gerundi
r3 *grup.verb* = *rightmost* <vmp
r4 *grup.verb* = *rightmost* <vsp
r5 *grup.verb* = *only_one* <v
r6 *grup.verb* = *rightmost* <vap
r7 *grup.verb* = *rightmost* <vmi

***Figure 3*:** Head rules for obtaining the head of a **grup.verb**[8] constituent

By rule number 1 the first element selected as the head of the verbal node is the infinitive (*infinitiu*). This means that for cases 3 and 4 in Figure 2, a head has already been selected. According to the second rule, the next selected head is the gerund (*gerundi*), which means that example 5 in Figure 2 is given a head. The third rule selects as head the verbal form whose pos-tag starts with *vmp*, which corresponds to participles; and example 2 is given its head. For the previous examples, rule number 4 does not apply. Rule 5 states that if there is only one verbal form, it is the head, so example 1 is finally given its head.

The key open question concerning dependency representation is related to coordination. The fundamental difference between coordination and subordination is that while in the latter there is a dependent element and a head; in the former, the two (or more) concerned elements are equivalent. This equivalence relationship cannot be represented by means of a dependency tree, since the basic relation here is the head-modifier one. Different solutions can be found, but generally speaking, the head is either the coordinating conjunction, like in the *Prague Dependeny Treebank* analytical level (Hajik, 1999), or one of the coordinated elements, which is the solution we adopted.

Coordination is not only an open question by itself, but also in related phenomena. For instance, how to deal with complements depending on two (or more) coordinated elements. In the *Tiger* project (Brants et al., 2003) there are secondary edges, and in the *Danish Dependency Treebank* there are secondary governors to represent this phenomenon (Kromann, 2003). In our case, we decided to connect these complements only with the head of the coordinated element. For instance, the head for a coordinated nominal group is the leftmost nominal group (no matter its type), any complement of the coordinated structure will be thus connected with this element.

On the other hand, an additional conversion table ('function table') was created for all the nodes that are not daughters of sentence structures since in these cases nodes are only labelled with the constituent format. Verbs which were the head of a sentence were given the function 'root'. The format of the 'function table' is as follows:

tag1 < tag2 = function_tag

where tag1 is the daughter, tag2 the mother, and function_tag the function of the daughter with respect to the mother, i.e. the edge label. Some examples of the function table are:

[8] It corresponds to the verbal node in the treebank. Under this node all possible verbal structures are represented.

r1 espec.fs < sn = DETER
r2 sn.co < grup.nom.fp = APOS
r3 sn < sp = CPREP
r4 s.a.ms < sn = CN
r5 sadv < sa = CADJ

Figure 4**:** The additional function table for constituents without syntactic function in the constituent treebank

The first rule in Figure 4 establishes the edge label DETER (determiner) for any *espec.fs* (feminine singular specifier) depending on a *sn* (noun phrase). The second sets the edge label APOS (apposition) for any *sn.co* (coordinated noun phrase) depending on a *grup.nom.fp* (feminine plural nominal group). CPREP (complement of a preposition) is the edge label for any *sn* (noun phrase) depending on a *sp* (prepositional phrase). CN (complement of a noun) is the label for any *s.a.ms* (masculine singular adjectival phrase) depending on a *sn*. Finally, CADJ is the edge label for any *sadv* (adverbial phrase) depending on a *sa* (adjectival phrase).

Function tags coming from the treebank used for the conversion appear in Table 4 together with a gloss of their meaning, and in Table 5 appear those coming from the function table.

Tag	Gloss	Tag	Gloss
ATR	Attribute	CREG	Prepositional complement
CAG	Agent complement	ET	Textual element
CC	Adjunct	IMPERS	Impersonal mark
CD	Direct object	MOD	Verb modifier
CD.Q	Quantitative direct object	PASS	Passive mark
CI	Indirect object	SUJ	Subject
CPRED	Predicative complement	VOC	Vocative
CPRED.CD	CD predicative complement		

Table 4: Tags coming from the treebank

Tag	Gloss	Tag	Gloss
ADJUNCT	Adjoined element	CPREP	Complement of a preposition
AO	Sentence adjunct	DETER	Head determiner
APOS	Apposition	ESPEC	Non-head determiner
AUX	Auxiliary verb	INSERT	Inserted element
CADJ	Complement of an adjective	INTJ	Interjection
CADV	Complement of adverb	MORF	Verbal morpheme
CN	Complement of a noun	NEG	Negative element
CNEG	Complement of a negation	PUNC	Punctuation mark
CO	Coordinating element	ROOT	Sentence head
CONJUNCT	Coordinated element	SUBORD	Subordinating element

Table 5**:** Tags coming from the function table

Once the conversion process was completed, the resulting treebank consisted of a tuple as follows:

position, stands for the word position in the sentence starting at 0
word, stands for the word form

lemma, stands for the lemma
pos, stands for the part-of-speech tag
head, stands for the head-position
function, stands for the syntactic function of the word

Figure 5**:** Fields in the dependency format of the treebank

The resulting dependency format is shown in Figure 6:

P	W	L	PoS	H	F
1	Per_tant	per_tant	rg[9]	13	ET
2	,	,	Fc	1	PUNC
3	les	el	da	4	DETER
4	escoles	escola	nc	13	SUJ
5	de	de	sp	4	CN
6	música	música	nc	5	CPREP
7	de	de	sp	4	CN
8	Reus	Reus	np	7	CPREP
9	i	i	cc	8	CO
10	Tortosa	Tortosa	np	8	CONJUNCT
11	passaran	passar	vm	13	AUX
12	a	a	sp	13	PREP
13	ser	ser	vs	0	S
14	conservatoris	conservatori	nc	13	ATR

Figure 6 The AnCora corpus in dependency format

The systematic conversion of AnCora-Cat and AnCora-Esp into a dependency structure was also a way of improving the quality of the original treebank, since the conversion process was very useful to check the quality of the annotation. These dependency treebanks were used as training and test corpora in the CoNLL Shared Task 2006 (Spanish) and 2007 (Catalan).

3. Morphosyntactic annotation

Prior to the syntactic and semantic manual annotations, the AnCora corpora were automatically annotated with PoS, the basis for further annotations. Besides, a shallow parsing (*chunking*) was applied in order to handle manual syntactic annotation.

The morphosyntactic annotation includes PoS assignment, lemmatization, and chunking. Since morphology deals exclusively with units separated by blank characters, complex morphological units as well as basic syntactic groups are solved at the chunking level.

[9] It is a simplified PoS notation.

3.1. Morphological annotation

The morphological annotation was carried out automatically and it meant associating the lemma, category and morphological attributes to each word in the corpus. The morphological annotation implied analysing morphologically each token by giving all the possible tags it could receive (see figure 7). Later, in the disambiguation process only one tag was selected.

Word	**lemma1**	**tag$_{1,}$**	**lemma2**	**tag$_{2,}$**	**lemma3**	**tag$_{3,}$**	**lemma$_4$**	**tag$_{4,}$ …**
bajo	bajar	VM1SIP,	bajo	AQ0MS,	bajo	P000,	bajo	NCMS, …

Figure 7: Output of the MACO morphological analyzer

The Spanish word ‘*bajo*’ can be the 3rd person singular form of the verb *go_down* (bajar-VM1SIP0), the adjective *short* (bajo-AQ0MS), the preposition *under* (bajo-P000) and the musical instrument *bass* (bajo-NCMS).

The annotation tool MACO (Atserias et al., 1998) is a Morphological Analyzer for Catalan, Spanish and English that follows a pipeline process: after a tokenization process, a subsequent module identifies dates, proper nouns, numerical expressions and currencies. Finally, the morphological analyzer assigns all possible morphological interpretations to the remaining tokens. Tokens can be single or multiword, which is declared in the MACO database.

The annotation system is based in the EAGLES proposal (Monachini et al., 1996), and it aims at making corpus analysis compatible with the grammatical tradition in order to guarantee the acceptance of the results for both fields. With respect to the lemma assignment, we followed the standards of morphology: singular for nouns and pronouns, masculine singular for adjectives and determiners, and infinitive for verbs. The tagset for Catalan and Spanish 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, morphological features, and specific categories like abbreviations, numbers, dates, and punctuation marks. There is also a label for unknown elements.

Each label consists of a definite number of digits: each digit expresses a predefined slot of information. In the case of nouns, for instance, the first digit expresses the main category (N, noun), the subcategory is shown in the second (C or P, common or proper name), the third position indicates its gender (M, F or C, corresponding to masculine, feminine or non-specified). Finally, the fourth position shows its number (S or F, singular or plural). The label for the word ‘*niño*’ (*child*) is NCMS (Noun Common Masculine Singular), and *joven* (‘young’) is labelled as NCCS (Noun Common C-non-specified-gender Singular). Figure 8 shows an example of the output of the morphological analysis in a vertical format:

Word	lemma$_1$	PoS$_1$	lemma$_2$	PoS$_2$	lemma$_3$	PoS$_3$	lemma$_4$	PoS$_4$
Si	si	CS	si	NCMS000	si	RG		
trabajo	trabajar	VMIP1S0	trabajo	NCMS000				
bajo	bajar	VMIP1S0	bajo	AQ0MS0	bajo 0	CMS00	bajo	SPS00
presión	presión	NCFS000						
bajo	bajar	VMIP1S0	bajo	AQ0MS0	bajo	NCMS000	bajo	SPS00
la	la	DA0FS0						
atención	atención	NCFS000						
.	.	Fp						

Figure 8**:** Output of the morphological analysis of the sentence: ‘*Si trabajo bajo presión bajo la atención’* [‘If I work under pressure my atention decreases’]

3.2. Morphological tagging

Since morphological analysis assigns all the possible tags to each word, a subsequent disambiguation process needs to be applied to obtain a single tag-lemma pair for each word. Once the morphological disambiguation process is applied, we obtain the following output of the previous sentence (Figure 9):

Word	**lemma**	**PoS**
Si	si	CS
trabajo	trabajar	VMIP1S0
bajo	bajo	SPS00
presión	presión	NCFS000
bajo	bajar	VMIP1S0
el	el	DA0MS0
interés	interés	NCMS000
.	.	Fp

***Figure 9*: Output of the morphosyntactic tagging**

The morphological disambiguation tool we used was RELAX (Padró, 1998). RELAX is a constraint-based probabilistic tagger that selects the sequence of tags that best satisfy a set of constraints for the sequence of input words. To find the output sequence of tags, RELAX performs approximate search using the Relaxation Labelling algorithm. Used as a pure probabilistic tagger, RELAX estimates the constraints from a tagged corpus considering the sequences of *n*-grams, but it can also incorporate manually written constraints in a definition language similar to Constraint Grammars. The accuracy of the pure probabilistic version of the tagger varies between 94-96% depending on the corpus and language of application. By using manually defined constraints to deal with exceptions and difficult cases the accuracy increases up to 95-97% (Civit, 2003). While automatic-learning constraints refer only to main categories, handwritten rules account for the ambiguity found in lemmas, subcategories and inflexions.

Tagging errors were basically due to intercategorial ambiguity concerning distinctions between determiners and pronouns, and between nouns and adjectives. Another source of errors was intra-categorial ambiguity because the tagger does not deal with some ambiguities concerning gender and number features for nouns and adjectives, and person for verbs (‘*salía/sortia’*, 1st and 3rd person singular), which can only be solved at the manual checking stage. Finally, a brief mention of particularly ambiguous words. The pronoun ‘*se’* (*himself, herself, itself*), for instance, can be reflexive, pronominal or a passive mark. Likewise, the form ‘*que*’ (‘that/which’) can be interpreted as a conjunction or as a pronoun.

3.3. Chunking

After the morphosyntactic analysis, we applied a fully automatic chunking process to the corpus in order to obtain a flat syntactic analysis, which was the input for the manual syntactic annotation. The more complete the partial parsing was, the less effort for the manual syntactic annotation.

The chunking process was carried out with TACAT (Atserias et al., 1998) and a context free grammar for Catalan and Spanish -*GramCat* and *GramEsp* respectively (Civit & Martí, 2005)- of about 2,000 handwritten rules. TACAT is a chart parser that works left–to–right and bottom–up. It applies the grammar and produces the chunking of the text, taking as input the output of the tagger. The chunker rules only group together those nodes with a 100% of certainty of setting up a chunk, leaving the remaining ones as chunks of only one element, unless for PP-attachment to nouns, as it is explained below.

Catalan and Spanish have a rich flexive morphology, so we extended Abney's (1991; 1996) concept of chunk, conceived in terms of 'major heads'. According to Abney, a 'major head' is any content word that does not appear between a function word *f* and the content word selected by *f*, or a pronoun selected by a preposition. As an example, '*proud'* is a major head in '*a man proud of his son'*, where two chunks are considered ('*a man'* and '*proud of his son'*) but not in '*a proud man'*, where there is only one chunk.

The idea of chunks which was adopted in our annotation system differed from Abney's because of the syntactic characteristics of Spanish. (Civit & Martí, 2005). In our approach we allowed the grammar to put together words if, according to their form, one could be sure that they go together. For instance, a noun phrase may include:

(det) (Adj) **Noun** [(Noun)/ (Adj)/ (SP[de])]

That is: an (optional) determiner in a pre-head position; an (optional) adjective before the noun; and another element after the noun: either another noun, or an adjective or a prepositional phrase headed by the preposition *de*. It is noteworthy that adjectives usually appear after a noun, and not before. If we had taken Abney's proposals literally, we should have considered *orgulloso* ('proud') as major head in *un hombre orgulloso de su hijo* but not in *un hombre orgulloso*. However, we consider that adjectives following and preceding a noun always belong to the nominal chunk. On the other hand, as it is well known, PP-attachment is one of the main issues in NLP. In our framework they were considered a separate chunk. Thus, for the sentence *un hombre orgulloso de su hijo*, the chunker produces this segmentation: [*un hombre orgulloso*] $_{NP}$ [*de su hijo*] $_{PP}$.

Consequently, noun phrases do not contain prepositional phrases, except for one case: when the PP is introduced by the preposition *de* ('of', 'from') immediately following the noun. Taking this decision involved an experiment that consisted in retrieving [noun +_ *de*_] sequences from the corpus. The hypothesis of departure was that '*de*'-PPs were attached to the previous noun. Out of 237 examples extracted from 210 sentences, 230 proved the hypothesis; 3 cases had ambiguous attachments and 4 were modifiers of other elements. Therefore, de-PPs were included in a NP if, and only if, the preposition *de* was next to the noun. This extended conception of chunk reduced the annotation time, although it was not error free. This chunking produced some errors in the analysis of the attachment (1-2%) but we considered that it was admissible if it reduced largely the annotator's work.

With respect to verbal phrases, the grammar recognized verbal groups, namely simple and complex verbal forms such as *es* ('is'), *ha sido* ('has been'), *debería haber sido* ('should have been'), *tiene que ser* ('has to be') in all their forms. Clitics and other particles (such as negative adverbs) were not included in the verbal group. Other elements recognized by the chunking grammar, such as relative pronouns, subordinating conjunctions, clitics, etc. were left as unary nodes in the tree.

In relation to coordination, this grammar only dealt with the simplest case: a coordinated chunk was made if, and only if, coordination occured between two (or more) single lexical items. For instance: a coordinated nominal chunk was built for *una lección de poderío y clase* ('one lesson of might and class') but it was not for *la debilidad sentimental, la resignación y el miedo* ('the emotional weakness, the resignation and the fear') due to the existence of articles and adjectives.

Finally, notice that during the chunking process, no clausal analysis was performed. It was possible to identify where they start, because the complementizer is mandatory in Spanish, but it was not possible to know where they finish. Therefore, they were manually built in the subsequent full syntactic annotation process. The main idea was to produce a 'correct' -even though incomplete - analysis because we relied on the correctness of the parser output to start the manual syntactic annotation.

4. Syntactic annotation

The full syntactic annotation process followed two main phases and each phase was organized into several steps. In the first phase, all syntactic constituents, and some elliptical elements were labelled. In the second phase, we took under consideration the main syntactic functions.

In order to build the AnCora treebank, the annotation criteria of the most significant existing corpora for different languages were consulted: *Susanne* and *Christine* corpus[10] (Sampson, 1995), *PennTreeBank-PTB* (Taylor et al., 2003; Marcus et al., 1993) and (Rambow et al., 2002) for English; the *Danish Dependency Treebank-DDT* (Kromann et al., 2003) and *Arboretum* (Bick, 2003)[11] for Danish; *Negra*[12] corpus (Brants et. al 2003), *Tüba*-D/Z treebank (Telljohann et al. 2006), and *Tiger* [13]corpus (Brants et al., 2002) for German; *Floresta* corpus (Afonso et al. 2002) for Portuguese; *BulTreeBank* (Simov et al., 2002) for Bulgarian; the *Prague Dependency Treebank-PDT* (Hajic, 1999; Böhmova et al., 1999) for Czek; for French (Abeillé et al., 2002); the *Hungarian National Corpus* (Varadi, 2002); the *Croatian National Corpus* (Tadic, 2002). For Polish we consulted (Marciniak et al., 2003); (Oflazer et al., 2001) for Turkish and (Bosco et al., 2000 and Montemagni et al., 2001) for Italian. Bearing in mind all these works, we defined a set of parameters to establish the main theoretical and methodological principles for building the Treebank

4.1. Basic assumptions

This section presents and discusses the major assumptions that were established to conduct the full syntactic annotation.

Implicit versus explicit information. Spanish is a pro-drop language, so the subject is usually omitted. In AnCora, only elliptical subjects were added since they are easily identifiable. We avoided the problem of recovering all the remaining elliptical constituents, because this implied an in-depth study that was beyond our purposes. In a further version of the treebank this and other questions such as the internal analysis of noun phrases will be considered.

Constituency versus dependency annotation. There is an open discussion about the annotation scheme to follow when building a Treebank. On the one hand, some papers claim that dependency annotation is more suitable for free word order languages (Brants et al., 2001; Oflazer et al., 2001; Boguslavsky et al., 2002), while others make their choice on the basis of the required application (Rambow et al., 2002). Finally, in some cases, the annotation system follows the linguistic tradition (Böhmova et al., 2003). 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

[10] http://www.grsampson.net/RSue.html; http://www.grsampson.net/ChrisDoc.html
[11] http://corp.hum.sdu.dk/arboretum.html / http://visl.sdu.dk/visl/da/info/
[12] http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/negra-corpus.html
[13] http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/

element varies in each sentence, thus, the pragmatic meaning varies. Furthermore, there are two noun phrases and both can precede the verb, which makes it impossible to know which one is the subject (unless semantic information is available). As a result, given that the order parameter concerns constituents (instead of words), the constituency framework was adopted. This is also why our previous tools were developed to prepare such an annotation scheme. Moreover, within the constituency framework we chose to annotate elements in a shallow-like way.

We also have available the dependency version of the treebank for Catalan and Spanish. Our syntax annotation scheme, which includes syntactic functions, made this task very easy (see section 2.3 in this paper for more details.).

Argument and Adjuncts. We did not make any distinction between arguments and adjuncts, so that the node containing the subject, that containing the verb and those containing verb complements and adjuncts are sister nodes. Such a decision also implied that no verbal phrase node was considered. At this point, the shallow analysis was able to eschew the problems of classical full-blown constituent structures: topicalizations, dislocations and wh-constructions.

Maintaining the surface word order. According to the previous points, no word order alterations were made during the annotation process. The strategy now was quite conservative. However, we were not ruling out this possibility in further developments of the Treebank. Generally speaking, sentence structure can be easily represented. The problem appeared 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 was adjoined to the node containing the adverb of comparison. In the second case, the constituent at the beginning of the sentence was given a special tag when annotating syntactic functions.

Being theory-neutral. Linguistic theories give solutions for some specific problems but they 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 corpora (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 English *PennTreeBank* 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 arguments and adjuncts, and with the PP-attachment, as stated in (Marcus et al., 1993; and Taylor et al., 2003).

Polish (Marciniak et al., 2001) and *BulTreeBank* (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; it provides a uniform way to represent different types of linguistic information; and it is widely used in computational linguistics. The latter claims that HPSG allows to represent simultaneously constituents as well as dependency relations, that this theory permits a consistent description of linguistic facts, that it enables translation to other formalisms, and that it can be used to support annotators' work. It is worth noticing that these treebanks consist of particularly selected sentences instead of by large collections of running texts. 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. In relation with annotation systems which do not follow any specific theory, it should be said (as in Abeillé et al., 2000) that this option allows adopting solutions equally profitable for linguists, computer scientists, psycholinguists, etc. Following this proposal, we did not want an application of one or another linguistic theory, but rather to fix a standard of constituency and functional annotation, neutral enough to be used for any research on Spanish and easy to translate into other formalisms. We think that the more neutral the annotation scheme, the more suitable 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 fine-grained annotation levels over a first shallow one.

4.2. Constituent tagging: sentence, clause and phrase structure

It is commonly accepted that any sentence has two main constituents: a subject and a predicate, the latter including the verb, its arguments and its adjuncts. The relationship between the verb and its arguments is closer than that between the verb and its adjuncts. Since Spanish and Catalan are free-constituent order languages, establishing a predicate node could have let us to alter the surface word order in the sentence. Hence, we decided not to deal with an intermediate predicate constituent. Consequently, the resulting analysis has less information as all sentence constituents are directly attached to the root node *S*, but on the other hand it allowed us to avoid the problem of distinguishing arguments and adjuncts, and to keep the surface word order. The subsequent semantic annotation specified if a constituent was argumental or not. Figure 10 shows a complete representation of the constituent structure of *S*.

```
(S
 (sn
  (espec.fs
   (da0fs0 La))
  (grup.nom.fs
   (ncfs000 declaración)))
 (grup.verb
  (vmis3s0 propugnó))
 (S.NF.C.co
  (S.NF.C
   (infinitiu
    (vmn0000 trabajar))
   (sp
    (prep
     (sps00 por))
    (sn
     (espec.fs
      (da0fs0 la))
     (grup.nom.fs
      (ncfs000 igualdad)
      (s.a.fs
       (grup.a.fs
        (aq0cs0 social))))))
   (Fp . .))
```

Figure 10: Example of a sentence annotated with constituents

Regarding the main structure of the sentence, three root nodes (*S.*, *S.co,* and *S**) were established, all of them appearing between strong punctuation marks: full stop, question and exclamation marks, and three dots. In this classification of root nodes, no distinction was made in relation with modalities, which meant that interrogative, imperative and declarative sentences received all the same tag. Punctuation marks allow us to infer information about modality. *S* (Figure

10) and *S** (Figure 11) differ in the fact that *S** contains only sentences without a verb form:

```
(S*
  (coord
    (cc I))
  (sadv
    (grup.adv
      (rg després)))
  (sn
    (espec.mp
      (da0mp0 els))
    (grup.nom.mp
      (ncmp000 empleats)))
  (Fp . .)))
```

***Figure 11*:** A sentence constituent structure with no verb form

S.co includes coordinated *S* or *S**, as in Figure 12:

```
(S.co
        (S
          (sn-SUJ
            (espec.mp
              (da0mp0 els el))
            (grup.nom.mp
              (ncmp000 béns bé)
              (sp
                (prep
                  (sps00 d' de))
                (sn
                  (grup.nom.ms
                    (ncms000 equipament equipament))))))
          (grup.verb
            (vaip3p0 van anar)
            (vmn0000 pujar pujar))
          (sn-CD.Q
            (espec.ms
              (di0ms0 un un))
            (grup.nom.ms
              (Zp 0,1% 0.1/100))))
        (coord
          (cc i i))
        (S*
          (sn-SUJ
            (espec.mp
              (da0mp0 els el))
            (grup.nom.mp
              (ncmp000 intermedis intermedi)))
          (Fc , ,)
          (sn-CD.Q
            (espec.ms
              (di0ms0 un un))
            (grup.nom.ms
              (Zp 0,9% 0.9/100))))))
```

```
(Fp . .)))
```

Figure 12**:** A sentence coordination structure

S-tag was only used for main clauses, whether root nodes or not. Subordinate clauses were sorted into two groups according to the verb form they had: either non-finite or finite. The former includes infinitive (S.NF.C), gerund (S.NF.A) and participle clauses (S.NF.P) (Figures 13, 14 and 15 respectively). The subject of these clauses rarely appears, as it corresponds to a constituent of the main clause (i.e. control structures). No trace or mark was included for this empty element. The latter includes completive (S.F.C), relative (R.F.R.) and adverbial clauses. Adverbial structures were split into those depending on the verb (S.F.A) and those with a sentential function: comparative (S.F.AComp), conditional (S.F.Acond), concessive (S.F.AConc) and consecutive clauses (S.F.ACons).

```
(S.NF.C
      (infinitiu
         (vmn0000 conservar))
      (sn-CD
         (espec.fs
            (dp3cs0 su))
         (grup.nom.fs
            (s.a.fs
               (grup.a.fs
                  (aq0fs0 vieja)))
            (ncfs000 casa)
            (sp
               (prep
                  (sps00 de))
               (sn
                  (grup.nom.ms
                     (ncms000 alquiler))))))
```

Figure 13**:** Example of a non-finite infinitive clause

```
(S.NF.A-CC-ArgM-ADV
      (gerundi-D2
         ( vmg0000 advertint))
      (sn-CI-Arg2-BEN
         (grup.nom.s
            (pp3csd00 -li)))
      (sadv-CC-ArgM-ADV
         (grup.adv
            (rg de_nou)))
      (S.F.C-CD-Arg1-PAT
         (conj.subord
            (cs que))
         (sn.e-SUJ-Arg0-AGT *0*)
         (neg-NEG
            (rn no))
         (grup.verb-D3
            (vmsi3s0 digués))
         (sn-CD-Arg1-PAT
            (grup.nom.s
                    (pi0cs000 res)))))
```

Figure 14**:** Example of a non-finite gerund clause

```
(S.NF.P-CPRED.SUJ-Arg2-ATR
    (participi
        (aq0mpp varados varado))
    (sp.co-CC-ArgM-LOC
        (sp
            (prep
                (sps00 en en))
            (sn
                (espec.fp
                    (da0fp0 las el))
                (grup.nom.fp
                    (ncfp000 afueras afueras))))
```

Figure 15 Example of a non-finite participle clause

Finite clauses are those that contain a finite verb form. Their subject may be elliptical, in which case a new constituent was added to the tree. If no verb form appears, then the subject was not recovered.

Completive clauses (S.F.C.) include clauses with a nominal function in the sentence. In Spanish they are typically expressed with the subordinating conjunctions *que* or *si* (Figure 16), with interrogative pronouns in reported speech (Figure 17), and with a subset of relative clauses with no explicit referred noun (antecedent) or with the pronoun *quien* ('who') (Figure 18).

```
(S.F.C.co
    (conj.subord
        (cs que))
    (S.F.C
        (sp
            (prep
                (sps00 en))
            (sn
                (espec.mp
                    (da0mp0 los))
                (grup.nom.mp
                    (ncmp000 países)
                    (s.a.mp.co
                        (sadv
                            (grup.adv
                                (rg más)))
                        (s.a.mp.co
                            (s.a.mp
                                (grup.a.mp
                                    (aq0mp0 prósperos)))
                            (coord
                                (cc y))
                            (S.NF.P
                                (participi
                              (aq0mpp desarrollados))))))))
```

Figure 16**:** A completive clause introduced by *que*

```
(S.F.C-CD-Arg1-PAT
        (sn-ATR-Arg2-ATR
            (grup.nom.fp
                (pt0fp000 quines)))
        (grup.verb-C3
            (vsif3p0 seran))
        (sn-SUJ-Arg1-TEM
            (espec.fp
                (da0fp0 les))
            (grup.nom.fp
                (ncfp000 pautes)
                (sp
                    (prep
                        (sps00 d'))
                    (sn
                        (grup.nom.fs
                            (ncfs000 actuació))))
```

Figure 17**:** A completive clause introduced by an interrogative pronoun (reported speech)

```
    (sn
        (espec.fp
            (da0fp0 Les)
            (Z 2.564))
        (grup.nom.fp
            (ncfp000 places)
            (sp
                (prep
                    (sps00 d'))
                (sn
                    (grup.nom.ms
                        (ncms000 aparcament))))
            (S.F.R
                (relatiu
                    (pr0cn000 que))
                (morfema.verbal
                    (p0000000 es))
                (grup.verb
                    (vmif3p0 construiran))
                (sadv
                    (grup.adv
                        (rg ara))))))
Fp.))))))))
```

Figure 18**:** A relative clause with no explicit antecedent

With respect to adjectival clauses (see Figure 18 above), no distinction was made between defining and non-defining clauses. They cover all sentences with an adjectival function. Adjectives and relative clauses nominalized by the neuter definite article *lo* were considered as nominal clauses.

Finally, adverbial clauses received the tag S.F.A. if they were a complement of time, place, mode, cause or goal (Figure 19). When they have a sentential function (Figure 20), the tag is longer and includes the type of the clause: S.F.ACond (conditionals), S.F.AConc (concesives) and S.F.ACons (consecutives). In function annotation, these adverbial clauses having a sentential function receive the tag AO, for *Adjunto Oracional* (‘Sentential Adjunct’).

```
 (
  (S
     (sn
        (espec.fp
         (da0fp0 Les el))
       (grup.nom.fp
        (s.a.fp
         (grup.a.fp
          (aq0fp0 pròximes pròxim)))
        (ncfp000 entrevistes entrevista)))
          (morfema.verbal-PASS
          (p0000000 s' es))
         (grup.verb
          (vaic3p0 haurien haver)
          (sps00 de de)
          (infinitiu
           (vmn0000 concretar concretar)))
         (S.F.A*
          (conj.subord
           (cs com com))
          (sadv
           (espec
            (rg més més))
           (grup.adv
            (rg aviat aviat)))
          (sadv
           (grup.adv
            (rg millor millor))))
(Fp . .)))
```

Figure 19**:** Time adverbial clause

```
(
 (S
  (sadv
   (grup.adv
    (rg Únicament únicament)))
  (sn.e-SUJ *0*)
  (sn
   (grup.nom.p
    (pp1cp000 ens jo)))
  (grup.verb
   (vmif1p0 desprendrem desprendre))
  (sp
   (prep
    (sps00 de de))
   (sn
    (grup.nom.ms
     (np00000 Saviola Saviola))))
  (S.F.ACond
  (conj.subord
    (cs si si))
  (sn.e *0*)
  (sn
    (grup.nom.p
      (pp1cp000 ens jo)))
  (grup.verb
    (vmip3s0 convé convenir)))
(Fp . .)))
```

***Figure 20*:**Adverbial clause with a sentential function

Nominal, adjectival and adverbial phrases were analyzed following the same syntactic schema: in the first level below phrase we distinguished *specifier* (optional) and *x-group* (*grup.nom*, *grup.adj* or *grup.adv*). Below *x-group* we find the head of the phrase (a noun, an adjective or an adverb) and the complement, which is also a phrase (Figure 21):

```
(sn
   (espec.fs
      (di0fs0 Una))
   (grup.nom.fs
      (ncfs000 información)
         (s.a.fs
            (grup.a.fs
               (aq0fs0 periodística)))))
```

***Figure 21*:** Phrase structure for nouns

In the case of ambiguous attachments, human annotators made their choice considering the context, which often helped them to solve ambiguities. However, some ambiguities remained that called for a decision to be made. Let us consider the sentence: '*La facultad de aprender y reaccionar ante nuevas situaciones'* (*The faculty of learning and reacting in the face of new situations*). Where '*ante nuevas situaciones'* is a PP which may depend either on '*reaccionar'* alone or on both infinitives '*aprender y reaccionar'*, and there is no way to know the appropriate attachment because context does not provide enough information. Two solutions have been widely adopted in the literature. One consists of defining a default attachment: the nearest node, the highest one, etc.; the other solution is marking the two (or more) different possibilities. In AnCora corpora we decided to attach the ambiguous node to the highest right one, since it is the more neutral position. This criterion applied not only to PP attachments but also to coordinated and relative clauses.

```
(sn
      (espec.fs
        (da0fs0 la))
      (grup.nom.fs
        (ncfs000 facultad)
        (sp
          (prep
            (sps00 de))
          (S.NF.C.co
            (S.NF.C.co
              (S.NF.C
  (infinitiu
   (vmn0000 aprender)))
 (coord
  (cc y))
 (S.NF.C
  (infinitiu
   (vmn0000 reaccionar))))
(sp
 (prep
  (sps00 ante))
 (sn
  (grup.nom.fp
   (s.a.fp
    (grup.a.fp
     (aq0fp0 nuevas)))
   (ncfp000 situaciones)))))))
```

Figure 22**:** *La facultad de aprender y reaccionar ante nuevas situaciones*

For more detailed information about constituent annotation see Bufí et al. 2007.

4.3 Function tagging

Function annotation was carried out after constituent annotation. It was agreed to tag only functions corresponding to sentence structure constituents, being it finite or non-finite: only subject and verbal complements were taken into consideration. We defined a total amount of 14 function tags, most of them corresponding to traditional syntactic functions such as subject, direct object, indirect object, etc. (see Table 6). The rest of tags (Table 7) were used to mark discourse elements (AO, ET, VOC) or modality (NEG, PASS, MOD, IMPERS).

Subject	**-SUJ**
Direct object	**-CD**
Indirect object	**-CI**
Attribute	**-ATR**
Predicative	**-CPRED**
Prepositional Complement	**-CREG**
Agent Complement	**-CAG**
Adverbial Complement	**-CC**
Adverbial Complement (Locative)	**-CCL**
Adverbial Complement (Time)	**-CCT**

Table 6: List of syntactic functions corresponding to sentence structure constituents

Textual element	**-ET**
Modal	**-MOD**
Negation	**-NEG**

Passive reflexive ‘se’	**-PASS**
Impersonal ‘se’	**-IMPERS**
Vocative	**-VOC**
Sentence Adjuncts	**-AO**

Table 7: Other syntactic functions corresponding to discourse and modality elements

In the current version of the treebank, noun, adjective and adverbial complements are not tagged with functions (see Figure 18 above). We represent functions as suffixes added to the constituent labels (Figure 23). This annotation schema was also followed in the argument and thematic-role tagging.

```
(S
        (sn-SUJ
            (espec.fs
                (da0fs0 La el))
            (grup.nom.fs
                (ncfs000 declaración declaración)))
        (grup.verb
            (vmis3s0 propugnó propugnar))
        (S.NF.C.co-CD
            (S.NF.C
                (infinitiu
                    (vmn0000 trabajar trabajar))
                (sp-CC
                    (prep
                        (sps00 por por))
                    (sn
                        (espec.fs
                            (da0fs0 la el))
                        (grup.nom.fs
                            (ncfs000 igualdad igualdad)
                            (s.a.fs
                                (grup.a.fs
                                    (aq0cs0 social social))))))
(Fp . .))
```

Figure 23: A complete parse tree annotated with constituents and functions

Table 8 below shows the relationship between constituents and the syntactic functions they convey:

Function Tags	Constituent tags
-SUJ	sn, sn.e, relatiu, S.F.C, S.NF.C
-CD	relatiu, S.F.C, S.F.C.co, sn, sp
-CI	sn, sp
-ATR	sa, sn, S.F.C, S.NF.C, S.NF.P, sp
-CPRED	sa, sn, S.NF.P, sp
-CREG	relatiu, sadv, S.F.C, sn, sp

-CAG	sp
-CC, -CCT, -CCL	sadv, S.F.A, S.NF.A, S.F.C, sn, sp
-ET	sadv, sp
-MOD	sadv, sp
-NEG	neg (negation)
-PASS	morfema.verbal (passive verb morpheme)
-IMPERS	morfema.verbal (impersonal verb morpheme)
-VOC	sn

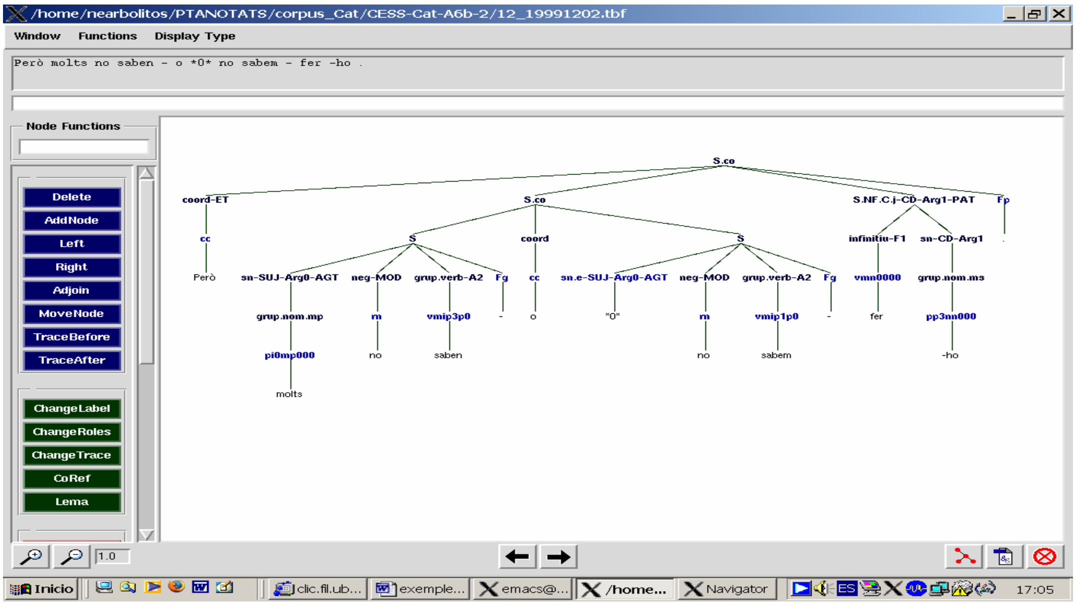
Table 8: Function tags and the constituents they can appear in

where *sn*, *sa*, *sp*, *sadv* stand for 'noun phrase', 'adjective phrase', 'prepositional phrase' and 'adverbial phrase', respectively; *sn.e* indicates an elliptical noun phrase; *relatiu* stands for a relative clause. S.F.C, S.NF.C distinguishes between finite and non-finite complement clauses; S.F.A., S.NF.A, S.F.P and S.NF.P establish the same distinction but between adjective and prepositional phrases. Finally, S.F.C.co stands for coordinated finite completive clauses.

4.4. Syntactic annotation tool

Manually building a Treebank requires a tool which facilitates the annotators' work. After checking several freely available interfaces, we decided to use the AGTK toolkit set up by the Pennsylvania University (Cotton et al., 2002). The main advantage was that it could easily accept our chunker output as well as our large tagset. It was slightly modified in order to allow both the processing of special characters and the processing of XML text format. Figure 24 shows a snapshot of this interface. AGTK allows annotators to split up or merge sentences, to join or split text, to add traces, leaves, nodes, to modify tags, to move nodes, etc. in a user friendly way.

This interface makes it possible to adapt the tagset used in the annotation process. For this reason it was used in different steps of the corpus annotation process: syntactic annotation –which includes constituents and functions-, strong and weak Named Entities, Argument Structure and Thematic Roles.



/home/nearbolitos/PTANOTATS/corpus_Cat/CESS-Cat-A6b-2/12_19991202.tbf

```
Però molts no saben - o *0* no sabem - fer -ho .
```

Figure 24: *TreeTrans* annotation tool for syntax, constituents and functions, annotation

5. Semantic Annotation: sentence, constituents and word level

The AnCora corpora are annotated with different kinds of semantic information: a) the semantic class and argument structure of verbal predicates, where the relationship between predicates and arguments is expressed by means of thematic roles; b) Named Entities, both strong and weak; and c) WordNet synsets for all the nouns in the corpora. As previously mentioned (subsection 2.1.3), a semiautomatic methodology was applied for the semantic annotation of predicates, while WordNet synsets and Named Entities were manually assigned (subsection 2.1.2).

5.1 Sentence level: verbal predicates annotation

The semantic annotation of verbal predicates implies the systematic mapping between syntax and semantics, basically expressed in the argument structure. This mapping ultimately motivates the semantic classes. In this proposal, each verbal predicate was assigned to a specific semantic class and every syntactic function was tagged with both arguments and thematic roles (Taulé et al., 2006b). The semantic properties which were used were defined assuming lexical decomposition (Levin & Rappaport Hovav, 1995; and Rappaport Hovav & Levin, 1998) from which the concept of Lexical Semantic Structure (LSS) was taken. The LSS as well as the kind of diatheses alternations in which the predicate can participate, determines the number of arguments that a verbal predicate requires and the thematic role of these arguments. In this line, we followed the lines laid down by Kipper et al., (2002) and Kingsbury et al., (2002) in the construction of *VerbNet*.

5.1.1 Lexical Semantic Structures (LSS)

We characterized predicates by means of a limited number of Lexical Semantic Structures and Event Structure Patterns, according to four basic event classes (Figure 25): accomplishment (1), achievement (2), state (3), and activity (4) (Vendler 1967; Dowty 1991). These general classes were further split into subclasses, depending on the argument structure, the thematic roles and the diatheses alternations (See section 6.1.4).

(1) [x CAUSE [BECOME [y <STATE/THING/PLACE>]]]

(2) [BECOME [y <STATE>]]]

(3) [x <STATE>]

(4) [x ACT <MANNER/ INSTRUMENT>]

Figure 25**:** The Lexical Semantic Structure of the four main classes of events

In Figure 25 the (1) LSS corresponds to the ontological class *accomplishments*, i.e. dynamic and telic events that refer to states in indirect or external cause processes indirectly or externally caused. They are prototypically transitive verbs. The (2) LSS corresponds to *achievements*, i.e. non-dynamic and telic events that refer to a state in processes without external cause. They are basically unaccusative verbs. The (3) LSS corresponds to *states*, i.e. non-dynamic and atelic events, with just one entity involved in the event, and focused in the state. Finally, the (4) LSS corresponds to *activities* or *processes*, i.e. dynamic and atelic events, they are always inergative predicates. In this way, the number of arguments that a verbal predicate requires as well as their thematic role are associated in the LSS.

The thematic roles are determined by the event class to which the predicate belongs and the type of diathesis alternations that the predicate has. In our proposal, each LSS restricts the set of possible diatheses. For example, predicates defined as *accomplishments* (LSS1) allow causative, inchoative (or anti-causative) and resultative diatheses, which focus on the primitives CAUSE, BECOME, and STATE respectively:

1.a ‘*Juan abre la puerta*’: [Juan CAUSE [BECOME [puerta <*OPEN*>]] (Causative)
Juan opens the door

1.b ‘*La puerta se abre*’: [BECOME [puerta <*OPEN*>]]] (Inchoative)
The door opens

1.c ‘*La puerta está abierta*’: [puerta <*OPEN*>]]] (Resultative)
The door is open

The examples (1.a, 1.b, 1.c) show that the predicate *abrir* ‘open’ surfaces in three different argument structures with different event structures. Consequently, the predicate *abrir* is treated as belonging to at least three different semantic classes: as an accomplishment in (1.a), as an achievement in (1.b), and as a state in (1.c); although the three LSSs are associated to the same sense of WordNet.

Following (Vàzquez et al., 2000 and Aparicio, 2007), diatheses are understood as the syntactic expression of a semantic opposition. Diathesis alternations are thus considered as pairs of structures related to each other by one of these oppositions. We considered the existence of three possible oppositions depending on whether there was a change of focus in the participants (change of focus), a change in the number of subcategorized arguments (change argument number) or a change in the event structure (aspectual opposition). For example, the sentences in (1.a-1.c) are related by a change of focus opposition: (1.a) expresses the cause that originates the event that is expressed, (1.b) focuses the change undertaken by the entity; and (1.c) focuses on the state.

The list of diathesis alternations is the following: *causative/inchoative*, *inchoative/causative*, *resultative*, *passive*, *holistic*, *benefective*, *transitive/intransitive*, *object extension* and *cognate object*.

5.1.2 Argument Structure and Thematic Roles

Argument structure encodes the prominence relations among arguments and also the thematic roles of each argument with respect to the predicate. For each verbal predicate the mapping between functions and thematic roles is declared, as well as the corresponding argument position, taking into account the semantic class related to the verbal predicate.

Following the *PropBank* style (Palmer *et al.*, 2005) and *VerbNet* (Kipper et al., 2002), the arguments selected by the verb are incrementally numbered –ArgA, Arg0, Arg1, Arg2, Arg3, Arg4– according to their degree of proximity in relation to their predicate. Adjuncts are labelled as ArgM. We use a very small set of thematic roles following the *VerbNet* (Kipper et al., 2002) proposal. Nevertheless, unlike *PropBank*, in our proposal each argument includes the thematic role in its label. The list of thematic roles consists of 20 different labels: AGT (Agent), AGI (Induced Agent), CAU (Cause), EXP (Experiencer), SCR (Source), PAT (Patient), TEM (Theme), ATR (Attribute), BEN (Beneficiary), EXT (Extension), INS (Instrument), LOC (Locative), TMP (Time), MNR (Manner), ORI (Origin), DES (Goal), FIN (Purpose), EIN (Initial State), EFI (Final State) and ADV (Adverbial).

Arg0 is assigned to arguments which are understood as agents, causers o experiencers, whereas Arg1 usually corresponds to the patient (arguments being affected by the action) or theme arguments (arguments undergoing a change of state). Arg2 is mainly assigned to beneficiary and attributive arguments, but also to extension and locative arguments. Arg3 is assigned to instrument, source and initial state arguments, whereas Arg4 corresponds to purpose and final state arguments. A special tag (ArgA) is used to capture the agent of an induced action[14]. Finally, ArgM corresponds to adjunct-like arguments: time, locative, manner, cause, goal, source, purpose, etc. Table 9 shows the correspondence between argument position and thematic roles is shown.

ArgA	-ArgA-AGI
Arg0	-Arg0-AGT, -Arg0-CAU, -Arg0-EXP, -Arg0-SRC
Arg1	-Arg1-PAT, -Arg1-TEM, -Arg1-EXT
Arg2	-Arg2-BEN, -Arg2-ATR, -Arg2-LOC, -Arg2-EXT, -Arg2-INS, -Arg2-EFI
Arg3	-Arg3-BEN, -Arg3-INS, -Arg3-ORI, -Arg3-EIN
Arg4	-Arg4-DES, -Arg4-EFI
ArgM	-ArgM-ATR, -ArgM-LOC, -ArgM-TMP, -ArgM-CAU, -ArgM-MNR, -ArgM-EXT, -ArgM-FIN, -ArgM-ADV

***Table 9*:** Correspondence between argument position and thematic roles

Thematic assignments depend not only on the LSS (and therefore on the type of event) but also on the diathesis alternations in which the predicates appear. Thus, one specific thematic role can appear in different argument positions depending on the verbal predicate. As it can be seen in the examples below (6.1, 6.2), the thematic role *extension* [EXT] can be realized as Arg1 or Arg2 depending on the predicate:

[14] For instance, 'p*assejar*' (*to walk*): '[El nen]$_{\text{Arg0}}$ passeja pel parc' vs. '[La mare]$_{\text{ArgA}}$ passeja [el nen]$_{\text{Arg0}}$ pel parc' (*The child goes for a walk in the park vs. The mother walks the child in the park*, where '*nen'* (*child*) is the agentive role and '*mare'* (mother) the agent of an induced action.

2. Recorrimos [tres kilómetros] $_{\text{CD-Arg1-EXT}}$ (*We walked three kilometres*)

3. Los precios aumentaron [un 5,6%] $_{\text{CD-Arg2-EXT}}$ (*Prices increased 5,6%*)

Our methodology allows for the possibility of not specifying the thematic role when a solution is not conclusive. For instance, in example (4).:

4. La Iglesia habla [del problema del Mal en el mundo] $_{\text{CREG-Arg1-}}$

the prepositional complement of the predicate '*hablar de*' (*to talk about*) has no thematic role since in the current version of the corpus prepositional objects have not been assigned a thematic role.

5.1.3. Argument structure and syntactic functions

Arguments might appear in different syntactic positions depending on the event structure and on the diathesis alternations in which they occur. For instance, Arg0 corresponds to external arguments, and it is prototypically assigned to the grammatical subject of predicates expressing an accomplishment or activity; but~~,~~ Arg0 can be assigned to the direct object when the subject is realized by an induced agent; in this case the subject is labelled as ArgA (see examples (5.a) and (5.b)).

5.a [***Pedro***]$_{\text{SUJ-Arg0-AGT}}$ paseó hasta la oficina (*Peter walked into the office*)
5.b [***Juan***]$_{\text{SUJ-ArgA-AGI}}$ paseava [***a su perro***]$_{\text{CD-Arg0-AGT}}$ *(John walked his dog)*

Arg1 corresponds to direct internal arguments, which prototypically are the direct object of accomplishments or the grammatical subject of achievements and states. Arg2 is assigned to the indirect object of accomplishments and also to the attribute of state predicates. Arg3 and Arg4 correspond to prepositional complements selected by the predicate. Finally, ArgM always corresponds to adjuncts. Table 10 shows the correspondence between syntactic functions and semantic arguments.

Argument	**Syntactic Function**
ArgA	SUJ
Arg0	SUJ, CD, CAG
Arg1	SUJ, CD, CREG
Arg2	SUJ, CD, CI, ATR, CREG, CC
Arg3	CI, CC
Arg4	CC
ArgM	CPRED, CC

ArgL	CD, ATR, CC,CREG
ArgX	CD

Table 10: Correspondence between arguments and syntactic functions

Annex 1 includes a complete table with the correspondence between *Arguments*, *Thematic roles* and *Syntactic Functions* illustrated with examples.

5.1.4. Spanish and Catalan Semantic Classes

In this section, we present the basic Lexical Semantic Classes derived from the LSSs mentioned above. These classes result from combining the LSS with the argument structure and the thematic roles. Each verbal class is also characterized by specific diathesis alternations. All this information is captured in the AnCora-Verb lexicon where the syntactic-semantic interface is expressed. For each verbal sense a semantic class is established, and the mapping between its syntactic functions[15] with the corresponding argument structure and thematic roles is declared.

The semantic classes used to characterize verbal predicates are hierarchically arranged in two levels. The first level contains information about the LSS structure, which is closely related to the event structure, corresponding to the main 4 classes (accomplishments, achievements, states and activities). At the second level these main classes are subspecified with information about argument structure and thematic roles, giving rise to a total of 13 classes. Thus, a verb related to a semantic class provides access to both syntactic and semantic information, which makes it possible to infer its event structure.

Next we present the 13 semantic classes that we have compiled so far. These classes result from analysing the 1,809 Spanish verbs and the 2,073 Catalan verbs found in the AnCora corpora.

Accomplishments: LSS1 (A)

LSS1 corresponds to the event structure of accomplishments, i.e. dynamic and telic events. Transitive constructions are the prototypical way to express telicity, where Arg0 corresponds to the subject –causative or agentive– and Arg1 to the direct object –thematic or patient–, depending on the possible diathesis alternations. Within LSS1, we distinguish three main classes: the *transitive-causative* class (A1), the *transitive-agentive* class (A2) and *ditransitive-agentive* class (A3).

These classes are established taking into account the possible diathesis alternations as well as the argument structure of predicates, that is to say, considering the number of subcategorized arguments (two arguments in A1 and A2, three arguments in A3) and the kind of thematic roles that can fulfill each argument (the causative subject in A1, and the agentive subject in A2 and A3). The ditransitive-agentive class (A3) is split into two further subclasses: the *locative ditransitive-agentive* class (A3.1), when the Arg2 is a *locative*, and the *beneficiary ditransitive-agentive* class (A3), when the Arg2 is a *beneficiary*.

LSS1.1 (A1)
[x CAUSE [BECOME [y <STATE >]]]
Arg0##CAU
Arg1##TEM
Diatheses: [+Inchoative] [+Resultative] [+/- Passive]

[15] We extracted the verbal syntactic frames from the corpus as described in Taulé et al., (2005) and Civit et al., (2005).

Spanish verbs: *abrir, aburrir, afear, agobiar, alisar, ascender, aumentar, bajar, cansar, causar, cerrar, congelar, convertir, desilusionar, deteriorar, elevar, emocionar, encandilar, entretener, excitar, freír, hundir, impresionar, inspirar, inundar, irritar, limpiar, mejorar, mezclar, motivar, nivelar, oxidar, pulir, romper, tranquilizar, turbar...*[16]
Catalan verbs: *agobiar, allisar, avorrir, convertir, deteriorar, emocionar, enfonsar, esgotar, espantar, fregir, millorar, obrir, oxidar, purificar, tancar, tranquilitzar, trencar...*

LSS1.2 (A2)
[[*x* DO-SOMETHING] CAUSE [BECOME [*y* < *STATE* >]]]
Arg0##AGT
Arg1##PAT
Diatheses: [-Inchoative] [+/-Resultative] [+Passive] [+/-Benefactive] [+/ Intransitive] [+/- Oblique Subject]

Spanish verbs*: acariciar, admirar, analizar, barrer, bautizar, beber, cantar, cazar, cepillar, comer, decidir, desear, educar, escribir, escuchar, forzar, fregar, leer, mirar, odiar, oler, orientar, peinar, plantar, probar, procesar, reparar, repetir, visitar…*
Catalan verbs: *acariciar, admirar, analitzar, beure, caçar, cantar, escombrar, escriure, escoltar, fregar, llegir, mirar, odiar, orientar, pentinar, repetir, visitar...*

LSS1.3.1 (A3.1)
[[x DO-SOMETHING] CAUSE [BECOME [y < PLACE > z]]]
Arg0##AGT
Arg1##PAT
Arg2##BEN
Diatheses: [-Inchoative] [+/-Resultative] [+Passive]

Spanish verbs: *aconsejar, adjudicar, confiar, contar, dar, decir, entregar, enviar, explicar, notificar, reclamar, recomendar, regular, sugerir, vender…*
Catalan verbs: *aconsellar, adjudicar, contar, dir, donar, entregar, enviar, explicar, recomenar, regular, vendre…*

LSS1.3.2 (A3.2)
[[x DO-SOMETHING] CAUSE [BECOME [y < PLACE >]]] or [[x DO-SOMETHING] CAUSE [BECOME [<THING > IN y]]]

Arg0##AGT
Arg1##PAT
Arg2##LOC
Diatheses: [-Inchoative] [+/-Resultative] [+Passive] [+/- Transitive]

Spanish verbs: *abordar, acercar, agregar, aislar, alejar, almacenar, apartar, arrojar, colocar, dejar, deportar, desplazar, empapelar, encarcelar, encerrar, ensobrar, esconder, esparcir, exponer, incluir, incorporar, incrustar, poner, publicar, recoger, registrar, señalar, tirar...*
Catalan verbs: *abordar, acollir, col·locar, emmagatzemar, emmantegar, empaquetar, empresonar, ensobrar, hospitalitzar, incloure, informar, moure, passar, posar, registrar ...*

[16] We assume that it is one of the possible senses of these verbs. Obviously, we can find that the same verb belongs to different semantic classes because of its polysemy.

(2)Achievements: LSS2 (B)

Verbs belonging to LSS2 correspond to the event structure of *achievements*, i.e. non-dynamic and telic events. They are basically unaccusative verbs. This class includes prototypically intransitive verbs whose subject behaves as an internal argument. In some languages, such as Catalan, this subject is characterized by allowing the clitization with the pronoun *'en'*; for example: '*Han arribat quinze turistes*' vs. '*N'han arribat quinze*' (lit. *Have arrived fifteen tourists* vs. *Of-them have arrived fifteen*).. The subject of these verbs usually appears in post verbal position. This is also the case in Spanish. We distinguished two subclasses: the *unaccusative-motion* class (B1), which includes verbs of inherent directed motion and verbs of appearance and disappearance (Levin 1993), and the *unaccusative-state* class (B2), which contains verbs indicating the final state. Verbs belonging to the LSS2 participate neither in the passive, inchoative nor in the resultative alternation. The subject maps onto Arg1 with the *theme* thematic role. This class also includes the inchoative constructions of predicates prototypically represented as accomplishments, like '*hundir'* (*to sink*). For instance, in the construction '*Los barcos se han hundido*' (*The ships sank*), the predicate '*hundirse'* is characterized as belonging to the semantic class B2.

LSS2.1 (B1)
[BECOME [*y* <*PLACE*>]]
Arg1-TEM/PAT
Diatheses: [- Passive] [+ Causative (hacer)]

Spanish verbs: *aparecer, caer, desaparecer, desembocar, llegar, partir, salir, venir, volver...*
Catalan verbs: *aparèixer, arribar, caure, desaparèixer, entrar, sortir, tornar, venir ...*

LSS2.2 (B2)
[BECOME [*y* <STATE>]]
Arg1-TEM/PAT
Arg2##EFI
Diatheses: [-Passive], [+ Causative (hacer)]

Spanish verbs: *caer enfermo, crecer, entrar en coma, entrar en silencio, florecer, hundirse…*
Catalan verbs: *créixer, enfonsar-se, entrar en coma, florir…*

(3) States: LSS3 (C)

The verbal classes related to LSS3 denote *states*, i.e. non-dynamic and atelic events. These predicates cannot be controlled by an *agent* and do not accept the passive alternation. Verbs belonging to this class have in common that their subject is an Arg1 with the *theme* thematic role. We distinguished four classes depending on the type of thematic role for Arg2. We basically differentiated between: the *existence state* class (C1), when Arg2 fits the thematic role *locative*; the *attributive state* class (C2), when Arg2 corresponds to an *attribute*; the *scalar state* class (C3), when the thematic role of Arg2 is an *extension*; and, finally, the *beneficiary state* class (C4), when the Arg2 fits the *beneficiary* thematic role.

LSS3.1 (C1)
[x <STATE >y]
Arg1-TEM
Arg2-LOC
Diatheses: [-Passive]

Spanish verbs: *estar, existir, haber…*
Catalan verbs: *estar, existir, haver-hi...*

LSS3.2 (C2)
[x <STATE >y]
Arg1-TEM
Arg2-ATR
Diatheses: [-Passive]

Spanish verbs: *acabar, comenzar, comportar, dar, durar, estar, hacer, inspirar, parecer, pasar, poseer, preceder, ser, terminar, tener...*
Catalan verbs: *acabar, començar, estar, semblar, posseir, precedir, ser,...*

LSS3.3 (C3)
[x <STATE >y]
Arg1-TEM
Arg2-EXT
Diatheses: [-Passive]

Spanish verbs: *costar, medir, pesar, valer…*
Catalan verbs: *costar, medir, pesar, valer…*

LSS3.4 (C4)
[x <STATE >y]
Arg1-TEM
Arg2-BEN
Diatheses: [-Passive]

Spanish verbs*: bastar, constar, corresponder, doler, extrañar, faltar, gustar, importar, interesar, parecer, pasar, repugnar…*
Catalan verbs: *agradar , correspondre, encantar, semblar, passar, repugnar…*

(4) Activities: LSS4 (D)

Verbal classes related to LSS4 denote activities, i.e. dynamic and atelic events. These predicates are monadic and inergative with the subject in Arg0 position. Consequently, the predicates cannot participate in the passive alternation, but they might accept the object extension alternation or cognate object alternation.

Three different semantic classes of activities were distinguished: *agentive inergative*, *experiencer inergative*, and *source inergative*. The difference between them basically depends on the thematic role of Arg0, which maps respectively onto the thematic roles of *agent*, *experiencer* and *source*.

LSS4.1 (D1)
[x ACT <*MANNER/INSTRUMENT* >]
Arg0-AGT
Diatheses: [-Passive], [+/-Extension Object]

Spanish verbs: *ayunar, ceder, caminar, cenar, contorsionarse, correr, desembarcar, escapar, establecerse, ir, moverse, nadar, navegar, revolotear, trabajar, vacilar…*

Catalan verbs: *anar, caminar, córrer, desembardar, nadar, navegar, treballar...*

LSS4.2 = D2
[*x* <STATE > *y*]
Arg0-EXP
Diatheses: [-Passive], [+Cognate Object]

Spanish verbs: *dormir, oír, parpadear, respirar, soñar, vivir...*
Catalan verbs*: dormir, roncar, somiar, viure...*

LSS4.3 = D3
[*x* <STATE > *y*]
Arg0-SRC
Diatheses: [-Passive], [+Cognate Object]

Spanish verbs: *chillar, crujir, gritar, jadear, llorar, relucir, romper, rugir, sollozar, sudar, toser…*
Catalan verbs*: brillar, cridar, plorar, roncar, suar, tossir…*

5.2. Named Entities and WordSenses

This subsection describes how the corpora were enriched with semantic information for the linguistic units below sentence level. At PoS level, all the nouns were tagged with WordNet senses, and strong named entities were identified and classified. All noun phrases corresponding to weak named entities were also annotated. These annotation processes were carried out manually and inter-annotator agreement tests were applied.

5.2.1. Named Entities

The AnCora corpora were annotated with both strong and weak Named Entities (Arévalo et al., 2004). We define strong NEs as corresponding to a word, a number, a date, or a string of words that refer to a single individual entity in the real world. From the point of view of the parse tree, strong NEs correspond to a linguistic form with a PoS tag. Examples of strong named entities are personal names and surnames (*John Kennedy Toole*), book~~’s~~ titles (*A Confederacy of Dunces*), some geographical and country names (*Colorado Canyon*, *New Orleans*), dates etc. In these cases, we analysed and annotated the whole string as a single element, thus enriching the PoS tag with information about the semantic class of the entity (see Figure 26).

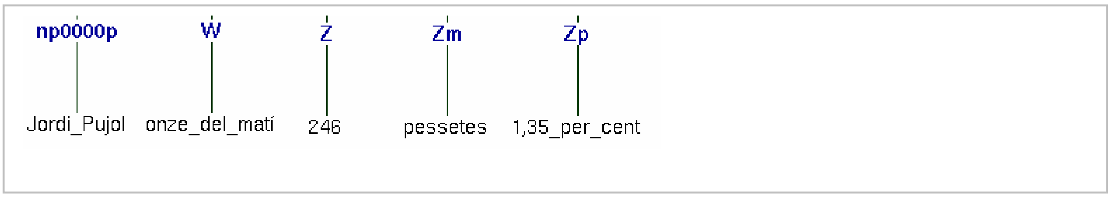


Figure 26: SNEs. (Jordi Pujol // eleven o’clock in the morning // 246 // pesetas // 1,35 percent)

Weak NEs consist of a noun phrase, being it simple or complex. Therefore, they are syntactic elements.

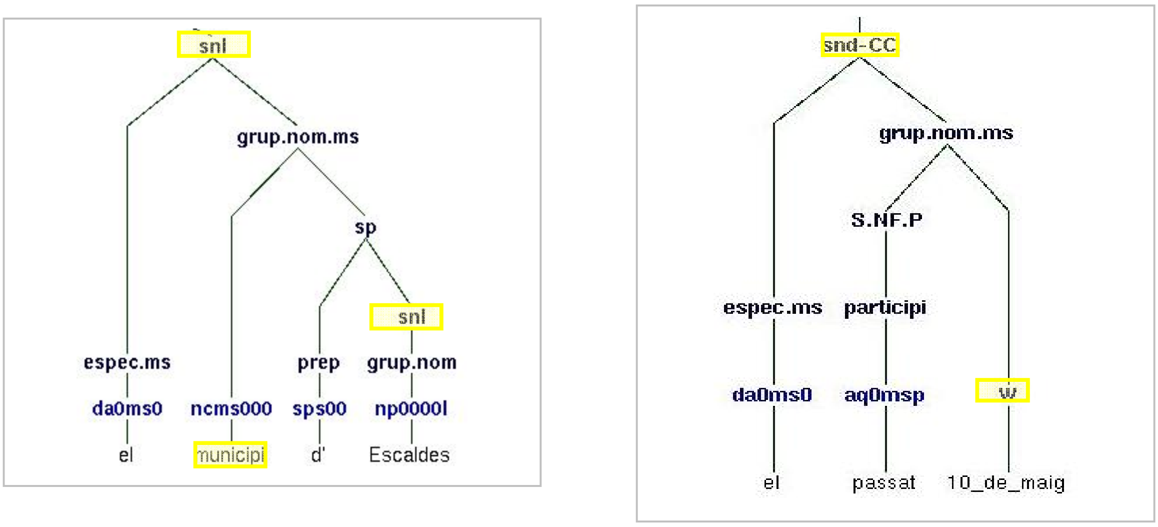


***Figura 27*:** weak Named Entities

Weak NEs (WNE) do not necessarily have a strong NE (SNE) within as a constituent. Some definite noun phrases whose head is a common noun may become a weak NE because of syntactic, semantic or pragmatic reasons. All definite noun phrases whose head is a *trigger word* complemented by either a national adjective (Figure 28) or a relational adjective derived from a proper noun (Figure 29) are considered WNEs.

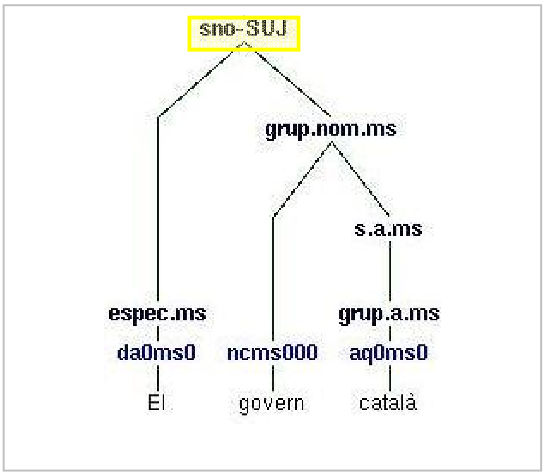


Figure 28: WNE with a *trigger word* as head (not a proper noun) complemented by a national adjective (the Catalan government).

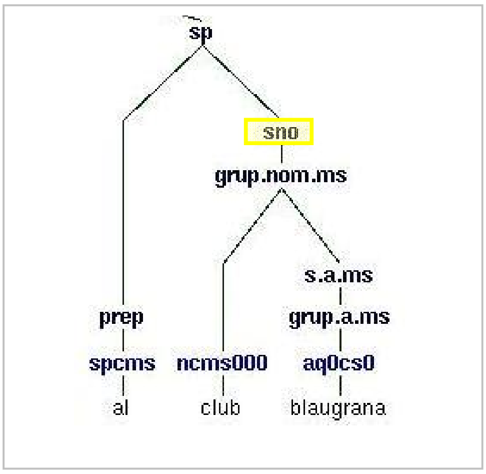


Figure 29: WNE with a *trigger word* as head (not a proper noun) complemented by a relational adjective (the 'blue and scarlet club' – namely, 'Football Club Barcelona').

Language use is essential to decide which relational adjectives are capable of turning their noun phrase into a WNE and which are not. *Republican*, for instance, may be ambiguous, since it can mean "belonging to the Republican Party" as well as "one who is in favour of avoiding monarchy". In such cases, we decided not to annotate the noun phrase as a weak named entity.

With regard to the semantic types assigned to each named entity, six basic semantic categories were distinguished: Person, Organization, Location, Date, Numerical expression, and Others. Syntactic labels occur at phrase level, whereas morphological labels occur at PoS level (see Table 11).

Strong NE (PoS)		**Weak NE (syntactic nodes)**	
Label	**Meaning**	**Label**	**Meaning**
np0000p	Person	snp	Person
np0000o	Organization	sno	Organization
np0000l	Location	snl	Location
np0000a	Other	sna	Other
Z	Alphanumerical (Numbers)	snn	Numbers
Zm	Alphanumerical (Coins)	snd	Date
Zp	Alphanumerical (Percentages)		
W	Date		

***Table 11*:** Morphological and syntactic labels for strong and weak NEs

In order to guarantee the quality of the results, an annotation guide (Borrega et al. 2007a and 2007b) was provided. In the annotation process we followed the same procedure as that established for manual tagging: annotation in parallel by more than two annotators, inter-annotator agreement tests, discussion, and modification of the guidelines when necessary. Single annotation began as soon as parallel annotation results showed a minimum of 95% agreement.

TreeTrans was the annotation tool used for NE tagging, the same as for syntactic annotation, but with the corresponding tagset.

5.2.2. Semantic lexical tagging

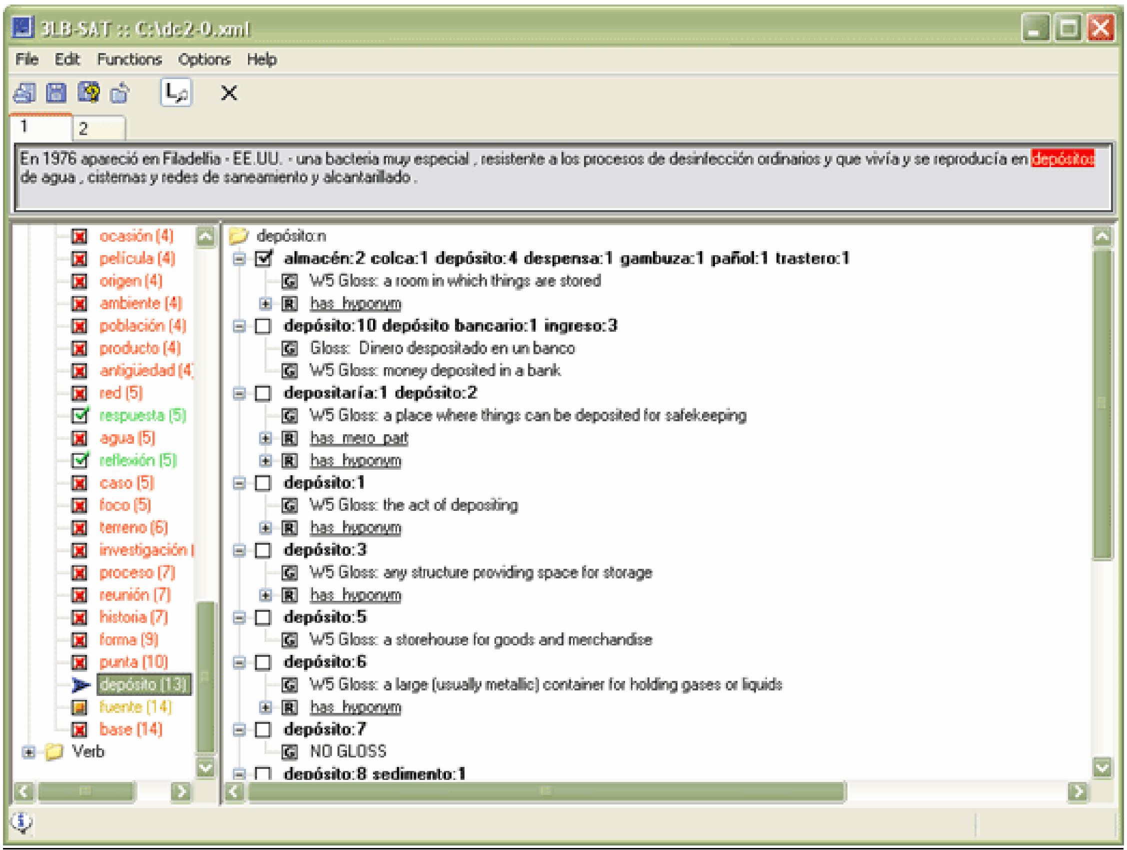
Taking into account that verbal predicates are annotated with the typology of semantic classes described in Section 5, we considered it a priority to tag the NP heads given their relevance to obtain the semantic type of verbal arguments.

The lexical semantic annotation consists in assigning each noun in the corpora its sense. This process was carried out manually and the senses repository is WordNet. We used a steady version of Catalan and Spanish EuroWordNets-1.6 (December 2005). Each noun was assigned either a WordNet sense or a label indicating a special circumstance (see Figure 30):

C2S: “The word does not exist in dictionary”.
C3S: “The word is part of a Multiword Lexical Unit or a lexicalized inflected form”
C4S: “The word is part of a Named Entity”
C5S: “The tagger is strongly uncertain”
C6S: “The word was improperly lemmatized or PoS-Tagged”
C7S: “The word is wrongly used: misspelling or a loanword”.

Figure 30**:** WordNet tagset for non ruled cases

In order to make the annotation task easier, the 3LB-SAT interface –3LB-Semantic Annotation Tool– (Bisbal et al., 2003) was built. This tool is lemma oriented, that is, the annotation process is made lemma by lemma in the whole corpus. Such a strategy facilitates the annotation task because the annotator focuses only on one WordNet entry, thus ensuring the consistency in the tagging process.



En 1976 apareció en Filadelfia - EE.UU. - una bacteria muy especial , resistente a los procesos de desinfección ordinarios y que vivía y se reproducía en depósitos de agua , cistemas y redes de saneamiento y alcantarillado .

☒ ocasión (4)
☒ película (4)
☒ origen (4)
☒ ambiente (4)
☒ población (4)
☒ producto (4)
☒ antigüedad (4
☒ red (5)
☑ respuesta (5)
☒ agua (5)
☑ reflexión (5)
☒ caso (5)
☒ foco (5)
☒ terreno (6)
☒ investigación (
☒ proceso (7)
☒ reunión (7)
☒ historia (7)
☒ forma (9)
☒ punta (10)
➤ depósito (13)
fuente (14)
☒ base (14)

depósito:n
- ☑ **almacén:2 colca:1 depósito:4 despensa:1 gambuza:1 pañol:1 trastero:1**
  - W5 Gloss: a room in which things are stored
  - has_hyponym
- ☐ **depósito:10 depósito bancario:1 ingreso:3**
  - Gloss: Dinero desposítado en un banco
  - W5 Gloss: money deposited in a bank
- ☐ **depositaría:1 depósito:2**
  - W5 Gloss: a place where things can be deposited for safekeeping
  - has_mero_part
  - has_hyponym
- ☐ **depósito:1**
  - W5 Gloss: the act of depositing
  - has_hyponym
- ☐ **depósito:3**
  - W5 Gloss: any structure providing space for storage
  - has_hyponym
- ☐ **depósito:5**
  - W5 Gloss: a storehouse for goods and merchandise
- ☐ **depósito:6**
  - W5 Gloss: a large (usually metallic) container for holding gases or liquids
  - has_hyponym
- ☐ **depósito:7**
  - NO GLOSS
- ☐ **depósito:8 sedimento:1**

***Figure 31*:** 3LB-SAT annotation tool for WordNet senses tagging

Figure 31 shows a snapshot of the annotation interface: in the upper part of the screen the sentence where the noun appears can be seen. On the left, the list of lemmas to be annotated is presented. In the centre of the screen all WordNet senses of the current word are shown, as well as the special cases that can be selected.

6. For concluding

To date, the AnCora corpora have been used as training and test corpora in three international evaluation events on syntactic and semantic NLP problems. On the one hand, the 3LB-ESP dependency Treebank was used in the CoNLL-06 shared task on Multilingual Dependency Parsing[17], and the whole AnCora-CAT dependency Treebank (500,000 words) was also a resource for the same shared task in CoNLL-07[18]. The 3LB-ESP corpus in dependency format was used as language resource to develop a syntactic parser for Spanish (Cowan & Collins 2005). The AnCora-Cat and AnCora-Esp corpora syntactically annotated were implemented in the Natural Language Toolkit (Loper and Bird, 2002).

On the other hand, a subset of 100,000 words from AnCora-Cat and AnCora-Esp fully annotated at the syntactic and semantic levels were used as training and test corpora in the

[17] http://www.cnts.ua.ac.be/conll2006/
[18] http://www.cnts.ua.ac.be/conll2007/

SemEval-07 task number 9, "Multilevel Semantic Annotation of Catalan and Spanish"[19,] involving the prediction of the three semantic levels given the text and its full syntax as input (Morante & Buser, 2007; Màrquez et. al., 2007).

In one year's time, we aim at having one million words for each Catalan and Spanish, publicly available and fully annotated.

Annotation Guidelines

Bufí, N., B. Soriano, M.A. Martí y M. Taulé (2007) Guidelines for the syntactic annotation of Spanish and Catalan CESS corpora: Constituents and Dependencies. *Lang2World* WP 01/2007.

Taulé, M., J. Aparicio, J. Castellví & M.A. Martí (2007) Guidelines for the Thematic Role Annotation of the AnCora corpora. *Lang2World* WP 02/2007.

Borrega, O., M.A. Martí & M. Taulé (2007a) Guidelines for strong and weak NE annotation in the AnCora corpora. Lang2World WP 03/2007.

References

Abeillé, A., L. Clément, A.Kinyon (2000) "Building a Treebank for French", in Proceedings of the Second Conference on Language Resources and Evaluation (LREC2000). Lisbon, Portugal.

Abeillé A., F.Toussenel, M.Chéradame (2002) Corpus le Monde. Annotations En Constituants.Guide Pour Les Correcteurs. Technical Report, LLF, UFRL.

Abeillé (2003) Treebanks. Building and Using Parsed Corpora. Series: Text, Speech and Language Technology, vol. 20. Springer Verlag.

Abney, S. (1991) "Parsing by Chunks", in Berwick R., S.Abney, and C.Tenny (eds.) Principle-Based Parsing, Kluwer Academic. Dordrech.

Abney, S. (1996) "Part-of-Speech Tagging and Partial Parsing", in Proceedings of the ESSLLI'96 Robust Parsing Workshop.

Afonso S., E. Bick, R. Haber, D.Santos (2002) "Floresta Sintá(c)tica': a Treebank for Portuguese", in Proceedings of the Third Conference on Language Resources and Evaluation (LREC2002), Granada, Spain???.

Aparicio, J. (2007) *Clasificación semántica de los predicados en español*. Master thesis, (Cognitive Science and Language program), Universitat de Barcelona, 2007.

Arévalo, M., M.A. Martí, M. Civit (2004) "MICE: a module for Named Entities Recognition and Classification", in International Journal in Corpus Linguistics, pp. 53 -69. John Benjamins, Amsterdam. ISSN: 1384-6655 Depósito legal: 36224

Atserias, J., Carmona, J., Castellón, I., Cervell, S., Civit, M., Màrquez, L., Martí, M.A., Padró, L., Placer, R., Rodríguez, H., Taulé, M. & Turmo, J. (1998) "Morphosyntactic Analysis and Parsing of Unrestricted Spanish Text". Proceedings of the First Conference on Language Resources and Avaluation 2. Granada: LREC: 1.267-1.272. http://garraf.epsevg.upc.es/freeling/

Beil, F., D. Prescher, H. Schmid, S. Shulte um Walde (2002) „Evaluation of the Gramotron parser for German", in Beyond Parseval, LREC02 Workshop. Lisbon, Portugal.

Bick, Eckhard (2003-7), "Arboretum, a Hybrid Treebank for Danish", in: Joakim Nivre & Erhard Hinrich (eds.), *Proceedings of TLT 2003 (2nd Workshop on Treebanks and Linguistic Theory, Växjö, November 14-15, 2003)*, pp.9-20. Växjö University Press

[19] http://nlp.cs.swarthmore.edu/semeval/

Black, E., Abney, S., Flickinger, C., Gdaniec, R., Grishman, R., Harrison, P., Hindle, D., Ingria, R., Jelinek, F., Klavans, J., Liberman, M., Marcus, M., Roukos, S., Santorini B. & Strzalkowski, T. (1991) "A Procedure for Qunatitatively Comparing the Syntactic Coverage of English Grammars". Proceedings of the Speech and Natural Language Workshop, pp: 306-311, Pacific Grove, CS. DARPA.

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

Böhmova A., Hajicova E. (1999) "How Much of the Underlying Syntactic Structure can be Tagged Automatically". Journées Atala, Corpus annotés pour la syntaxe, Paris, June 1999.

Bosco C., Lombardo V., Vassallo D., Lesmo L. (2000) "Building a Treebank for Italian: a Data-driven Annotation Schema", in Proceedings of the Second Conference on Language Resources and Evaluation (LREC2000).

Borrega, O., Taulé, M. & Martí, MA. (2007b) "What do we mean when we speak about Named Entities?", in *Corpus Linguistics Conference*. Birmingham. http://www.corpus.bham.ac.uk/conference2007/

Brants, Th., W. Skut & H. Uszkoreit (2003) "Syntactic Annotation of a German Newspaper Corpus", in: A. Abeillé (Ed.) Treebanks Building and Using Parsed Corpora, Book Series: Text, Speech and Language Technology: Volume 20, Kluwer Academic Publisher, Dordrecht.

Brants, S., S. Dipper, S.Hansen, W. Lezius, and G.Smith, (2002) "The TIGER Treebank" in Proceedings of the Workshop on Treebanks and Linguistic Theories. Sozopol, Bulgaria.

Carmona, J., Cervell, S., Màrquez, L., Martí, M.A., Padró, L., Placer, R., Rodríguez, H., Taulé, M. & Turmo,J. (1998) "An Environment for Morphosyntactic Processing of Unrestricted Spanish Text". Proceedings of the First Conference on Language Resources and Avaluation 2. Granada: LREC: 915-922.

Civit , M., A. Ageno, B. Navarro, N. Bufí & M.A. Martí (2003) "Quantitative and qualitative Analysis of Annotators' Agreement in the development of 3LB", 2nd. Workshop on Treebanks and Linguistics Theories. Växö, Sweden: November 2003.

Civit, M. (2003) Criterios de etiquetación y desambiguación morfosintáctica de corpus del español. Monografías de la Sociedad Española para el Procesamiento del Lenguaje Natural, 3. Alicante.

Civit, M. & M.A. Martí (2004a) "Estándares de anotación morfosintáctica para el español". Proceedings of IX Ibero-American Workshops on Artificial Intelligence. Iberamia Conference. México. pp.217-224.

Civit, M. & M.A. Martí (2004b) "Building Cast3LB: a Spanish Treebank", in Research on Language & Computation (2004) 2. pp.: 549-574. Springer, Science & Business Media. Germany.

Civit, M. & Martí, M.A. (2005) 'GramCat and GramEsp : two Grammars for Chunking', en Intelligent Information processing and Word Mining, Gdansk, Poland. Springer Verlag. ISSN 1615-3871.

Civit, M., Martí, M.A. & Bufí, N. (2006) "Cat3LB and Cast3LB: From Constituents to Dependencies". Advances in Natural Language Processing. Germany: Springer: 141-152.

Cotton S. and Bird S. (2002) "An Integrated Framework for Treebanks and Multilayer Annotations", in *Proceedings of the Third Conference on Language Resources and Evaluation* (LREC2002).

Cowan, B. & M. Collins (2005) "Morphology and Reranking for the Statistical Parsing of Spanish", in *Proceedings of the Empirical Models for Natural Language Processing-EMNLP-2005*.

Demonte, V. (2003) "Preliminares de una clasificación léxico-sintáctica de los predicados verbales del español". In Sybille Grosse & Axel Schönberger (eds.) *Ex oriente lux: Festchrift für Eberhard Gärtner zu seinem 60*. Geburtstag. Frankfurt am Main: Valentia.

Dowty, D. (1991) 'Thematic proto-roles and argument selection'. *Language*, 67.

Hajik, J. (1999) "Building a syntactically annoated corpus: the Prague Dependency Treebanks", in Issues in Valency and Meaning. Studies in honour of Jarmila Panevova.

Kingsbury, P., Palmer, M. & Marcus, M. (2002) 'Adding semantic annotation to Penn TreeBank". 2002, in *Proceedings of the 2002 Conference on Human Language Technology*, San Diego, CA.

Kipper, K., Palmer, M. & Rambow, O. (2002) "Extending PropBank with VerbNet Semantic Predicates". *Workshop on Applied Interlinguas*, held in conjunction with AMTA-2002. Tiburon, CA.

Kromann, Mikkelsen and Lynge (2003) "Danish Dependency Treebank and the Underlying Linguistic Theory", *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*, Växjö, Sweden.

Levin, B. & Rappaport-Hovav, M. (1995) *Unaccusativity. At the Syntax-Lexical Semantics Interface*. Cambridge, MA: MIT Press.

Lin, D. (1998) "A dependency-based method for evaluating broad-coverage parsers", in *Natural Language Engineering*, num. 4 (2).

Loper, E., and Steven Bird (2002) "NLTK: The Natural Language Toolkit", in *ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language*, 2002.

Marcus, M., B. Santorini & M.A. Marcinkiewicz (1993) "Building a large annotated corpus of English: the Penn Treebank", *Computational Linguistics*, num 19 (2). MIT Press.

Martí, M.A., M. Taulé, M. Bertran & L. Màrquez (2007) "Anotación semiautomática con papeles temáticos de los corpus CESS-ECE", in *Procesamiento del Lenguaje Natural*, num. 38, pp. 67-77. Alicante, Spain. ISSN-1135-59-48.

Màrquez, L., Taulé, M., Martí, M.A., Artigas, N., García, M., Real F. & Ferrés, D. (2004) "Senseval-3: The Spanish Lexical Sample Task", in *Proceedings of Senseval-3, ACL Conference-2004*. Barcelona. pp. 21-25.

Màrquez, Ll., L. Padró, M. Surdeanu, L. Villarejo (2007) "UPC: Experiments with joint learning within SemEval task 9", *ACL, SemEval Workshop*, pp. 426-429. Prague.

Marciniak M., A. Mykowiecka, A. Przepiórkowski, A. Kupsc (2001) "Construction of an HPSG Treebank for Polish", in A. Abeillé, (ed.) *Building and Using Syntactically Annotated Corpora*. Language and Speech: Kluwer, Dordrecht.

Monnachini, M., & N. Calzolari (1996) *Sinopsis and Comparision of Morphosyntactic Phenomena Encoded in Lexicons and Corpora*. A Common Proposal and Applications to European Languages. EAGLES 1996.

Montemagni S., 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 Abeillé A. (ed.), Building and Using Syntactically Annotated Corpora, Series: Language and Speech, Kluwer, Dordrecht.

Morante, R. and B. Busser (2007) "ILK2: Semantic Role labelling for Catalan and Spanish using TiMBL", *ACL, SemEval Workshop*, pp. 183-186. Prague.

Oflazer K., B. Say, D.Z. Hakkani-Tür, G. Tür (2001) "Building a Turkish Treebank". in Abeillé A. (ed.), Building and Using Syntactically Annotated Corpora, Series: Language and Speech, Kluwer, Dordrecht.

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

Rappaport Hovav, M. & Levin, B. (1998) "Building Verb Meanings", in M. Butt and W. Geuder, eds., The Projection of Arguments: Lexical and Compositional Factors, CSLI Publications, Stanford, CA, 97-134.

Padró, L. (1998) *A Hybrid Environment for Syntax--Semantic Tagging* PhD. Thesis, Dep. Llenguatges i Sistemes Informàtics. Universitat Politècnica de Catalunya. February, 1998.

Sampson, G.(1987) "Probabilistic Models of Analysis", in Garside R., Leech, G., Sampson, G. (ed.), The Computational Analysis of English, Chapter 1. Longman, New York.

Sampson, G. (1995) English for the Computer. The Susanne corpus and Analitic *Scheme*. Clarendon Press, Oxford.

Sebastián, N., M.A.Martí, , M.F.Carreiras, & F.Cuetos, (2000) LEXESP: Léxico Informatizado del Español. Barcelona: Ediciones de la Universitat de Barcelona.

Simov, K., P. Osenova, M. Slavcheva, S. Kolkovska, E. Balabanova, D.Doikoff, K. Ivanova, A. Simov, M. Kouylekov (2002) Building a Linguistically Interpreted Corpus of Bulgarian: the BulTreeBank. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC02).

Tadic M. (2002) "Building the Croatian National Corpus", in Proceedings of the Third *International Conference on Language Resources and Evaluation* (LREC02).

Taulé, M., Aparicio, J., Castellví, J. & Martí, M.A. (2005) "Mapping syntactic functions into semantic roles", Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT05)". Barcelona: Universitat de Barcelona, 2005: 185-196.

Taulé, M., J. Castellví, M. A. Martí & J. Aparicio (2006a) "Fundamentos teóricos y metodológicos para el etiquetado semántico de CESS-CAT y CESS-CAST", in Procesamiento del Lenguaje Natural, num. 37 pp. 75-82

Taulé, M., J. Castellví, & M.A. Martí, (2006b) "Semantic Classes in CESS-LEX: Semantic Annotation of CESS-ECE", in J. Hajic & J. Nivre (eds.) Fifth Workshop on Treebanks and Linguistic Theories, pp. 139-150. Prague: Czech Republic. ISBN: 80-239-8009-2.

Taylor, A., M.Marcus & B.Santorini (2003) "The PennTreeBank: an overview", in A. Abeillé (Ed.) Treebanks, Springer Verlag Series Text, Speech and Language Technology, vol.20. ISBN: 978-1-4020-1334-8.

Telljohann, E., E. Hinrichs and S. Klüber, H. Zinsmeister (2006) Stylebook for the Tünbingen Treebanks of Written German (Tüba-D/Z). Universität Tübingen, Seminar für Sprachwissenschaft.

Váradi T. (2002) "The Hungarian National Corpus", in *Proceedings of the Third International Conference on Language Resources and Evaluation* (LREC02).

Vàzquez, G., Fernández, A. & Martí, M.A. (2000) Clasificación verbal. Alternancias de diátesis. Edicions de la Universitat de Lleida: Lleida.

Vendler, Z. (1967) Linguistics in Philosophy. Cornell University Press: New York.

Annex-1

Verbal Semantic Classes, Syntactic Functions, Thematic Roles

Argument-Thematic role	**Description**	***Syntactic Function***	***Example***
ArgA-AGI	Induced Agent	Subject (SUJ)	**Su abolición**$_{\text{SUJ-ArgA-AGI}}$ pondría a tributar a todos los empleados
Arg0-AGT	Agent	Subject (SUJ) Direct object (CD) Agent compl. (CAG)	**Juan**$_{\text{SUJ-Arg0-AGT}}$ corre **Juan**$_{\text{SUJ-Arg0-AGT}}$ lee una novela Su abolición pondría a tributar **a todos los empleados**$_{\text{CD-Arg0-AGT}}$ Clara es amada **por todos**$_{\text{CAG-Arg0-AGT}}$
Arg0-CAU	Cause	Subject (SUJ) Agent compl. (CAG)	**Juan**$_{\text{SUJ-Arg0-CAU}}$ rompió la ventana **El trabajo**$_{\text{SUJ-Arg0-CAU}}$ agota a María
Arg0-EXP	Experiencer	Subject (SUJ)	**Juan**$_{\text{SUJ-Arg0-EXP}}$ sueña
Arg0-SCR	Source	Subject (SUJ)	**El enfermo**$_{\text{SUJ-Arg0-SCR}}$ sudaba
Arg1-TEM	Theme	Subject (SUJ) Direct object (CD)	**Los niños**$_{\text{SUJ-Arg1-TEM}}$ llegaron tarde **El pan**$_{\text{SUJ-Arg1-TEM}}$ cuesta dos euros El fuerte viento hundió **el barco**$_{\text{CD-Arg1-TEM}}$ Llueve **barro**$_{\text{CD-Arg1-TEM}}$
Arg1-PAT	Patient	Subject (SUJ) Direct object (CD)	**Clara**$_{\text{SUJ-Arg1-PAT}}$ es amada por todos Juan lee **una novela**$_{\text{CD-Arg1-PAT}}$
Arg1-EXT	Extension	Direct object (CD)	Juan caminó **3 km**$_{\text{CD-Arg1-EXT}}$
Arg1-	Unspecified[20]	Prepositional compl. (CREG)	Trata **de evitar una canasta**$_{\text{CREG-Arg1-}}$
Arg2-BEN	Beneficiary	Indirect object (CI)	Juan da un pastel **al niño**$_{\text{CD-Arg2-BEN}}$

[20] The methodology allows for the possibility of not specifying the thematic role when a solution is not conclusive, that is the case of most argumental prepositional complements (CREG).

Arg2-ATR	Attribute	Attribute (ATR)	Juan es **listo**$_{\text{ATR-Arg2-ATR}}$
Arg2-LOC	Locative	Subject (SUJ) Direct Object (CD) Prepositional compl. (CREG) Adverbial complement (CC)	**La novela**$_{\text{SUJ-Arg2-LOC}}$ aborda esa temàtica El acusado abandonó **la sala**$_{\text{CD-Arg2-LOC}}$ Laura entró **en la habitación** $_{\text{CREG-Arg2-LOC}}$ El escritor aborda la violencia **en su última novela** $_{\text{CC-Arg2-LOC}}$
Arg2-EXT	Extension	Direct object (CD)	El pan cuesta **75 céntimos**$_{\text{CD-Arg2-EXT}}$
Arg2-INS	Instrument	Prepositional compl. (CREG)	Juan abre la puerta **con la llave**$_{\text{CC-Arg2-INS}}$
Arg2-EFI	Final State	Adverbial complement (CC) Prepositional compl. (CREG)	Juan entró **en coma** $_{\text{CC-Arg2-EFI}}$
Arg3-BEN	Beneficiary	Indirect object (CI)	**Nos**$_{\text{CI-Arg3-BEN}}$ cuesta trabajo evolucionar
Arg3-INS	Instrument	Adverbial complement (CC)	Juan dió una mano de pintura **con la brocha**$_{\text{CC-Arg3-INS}}$
Arg3-ORI	Origin	Adverbial complement (CC)	Juan arrastró la silla tres metros **desde mi despacho**$_{\text{CC-Arg3-ORI}}$
Arg3-EIN	Initial State	Adverbial complement (CC)	Las ventas aumentaron un 10% **de un millón de euros**$_{\text{CC-Arg3-EIN}}$ a 1,1 millones de euros
Arg4-DES	Purpose	Adverbial complement (CC)	Juan arrastró la silla de un sitio **a otro**$_{\text{CC-Arg4-DES}}$
Arg4-EFI	Final state	Adverbial complement (CC)	Las ventas aumentaron un 10% de un millón de euros **a 1,1 millones de euros**$_{\text{CC-Arg4-EFI}}$
ArgM-ATR	Attribute	Predicative (CPRED)	Hablaba **tranquilo**$_{\text{CPRED-ArgM-ATR}}$ El gobierno mantuvo los precios **estables**$_{\text{CPRED-ArgM-ATR}}$
ArgM-LOC	Locative	Adverbial complement (CC)	Juan vive **en Barcelona**$_{\text{CC-ArgM-LOC}}$
ArgM-TMP	Time	Adverbial complement (CC)	Llueve **cada día**$_{\text{CC-ArgM-TMP}}$
ArgM-CAU	Cause	Adverbial complement (CC)	Toma antibiótico **porque està resfriado**$_{\text{CC-ArgM-CAU}}$
ArgM-MNR	Manner	Adverbial complement (CC)	Juan duerme **profundamente**$_{\text{CC-ArgM-MNR}}$
ArgM-EXT	Extension	Adverbial complement (CC)	Compró un coche **por 2.000 euros**$_{\text{CC-ArgM-EXT}}$
ArgM-FIN	Goal	Adverbial complement (CC)	Juan amortiza capital **para reducir cuota**$_{\text{CC-ArgM-FIN}}$
ArgM-ADV	General	Adverbial complement (CC)	Las leía **de nuevo** $_{\text{CC-ArgM-ADV}}$

ArgL	This argument indicates that the constituent is part of the verb	Direct object (CD) Attribute(ATR) Adverbial complement (CC) Prepositional compl. (CREG)	Lo atacaron cuando bajó **la guardia**$_{\text{CD-ArgL}}$ Dio **las gracias**$_{\text{CD-ArgL}}$ a su amigo Dio **a luz**$_{\text{CC-ArgL}}$ a las 7 p.m.
ArgX	The argument of aspectual verbs	Direct object (CD)	Suele **cantar**$_{\text{CD-ArgX}}$ en la ducha