AnCora: Multilevel Annotated Corpora for Catalan and Spanish

Mariona Taulé, M. Antònia Martí, Marta Recasens

CLiC, Centre de Llenguatge i Computació
Department of Linguistics
University of Barcelona
08007 Barcelona, Spain
{mtaule, amarti, mrecasens}@ub.edu

Abstract

This paper presents AnCora, a multilingual corpus annotated at different linguistic levels consisting of 500,000 words in Catalan (AnCora-Ca) and in Spanish (AnCora-Es). At present AnCora is the largest multilayer annotated corpus of these languages freely available from `http://clic.ub.edu/ancora`. The two corpora consist mainly of newspaper texts annotated at different levels of linguistic description: morphological (PoS and lemmas), syntactic (constituents and functions), and semantic (argument structures, thematic roles, semantic verb classes, named entities, and WordNet nominal senses). All resulting layers are independent of each other, thus making easier the data management. The annotation was performed manually, semiautomatically, or fully automatically, depending on the encoded linguistic information. The development of these basic resources constituted a primary objective, since there was a lack of such resources for these languages. A second goal was the definition of a consistent methodology that can be followed in further annotations. The current versions of AnCora have been used in several international evaluation competitions

1. Introduction

AnCora results from two different corpora: AnCora-Es, the Spanish corpus, and AnCora-Ca, the Catalan one. Both consist of half million words annotated at different levels of linguistic description: morphological (PoS and lemmas), syntactic (constituents and functions), and semantic (argument structures, thematic roles, semantic verb classes, named entities, and WordNet nominal senses). However, the two corpora find themselves at different annotation stages: AnCora-Ca is fully annotated, whereas only 187,000 words of AnCora-Es are at present completely annotated.

The annotation process was carried out sequentially from lower to upper layers of linguistic description: morphology first; different levels of syntactic description next; and semantic annotation third. All resulting layers are independent of each other, thus making easier the data management. The annotation was performed manually, semiautomatically, or fully automatically, depending on the encoded linguistic information. The development of these basic resources constituted a primary objective, since there was a lack of such resources for these languages. A second goal was the definition of a consistent methodology that can be followed in further annotations.

AnCora is the largest multilayer annotated corpus of Spanish and Catalan freely available from `http://clic.ub.edu/ancora`. It is comparable to other corpora which are being developed at present: the English OntoNotes (Hovy et al., 2006), and the Czech Prague Dependency Treebank (Linh and Zabokrstsky, 2007). The current versions of AnCora have been used in several international evaluation competitions such as CoNLL-2006, CoNLL-2007 and SemEval-2007, concerning different syntactic and semantic NLP tasks.

This paper presents AnCora by summarizing the data resulting from each level of annotation. Section 2 describes the sources from which the texts of AnCora were extracted. The methodology and criteria guiding the annotation process are presented in Section 3. Section 4 is split into three subsections, providing an overview of the three levels of annotation –morphological, syntactic, and semantic– respectively. Section 5 provides a fully annotated example. Finally, main conclusions are drawn in Section 6.

2. AnCora: Sources

AnCora was built in an incremental way from the previous 3LB (Civit and Martí, 2004a) and CESS-ECE (Martí and Taulé, 2008) corpora, which come mostly from newspaper and newswire articles.

AnCora-Es contains 75,000 words from *Lexesp* –a Spanish balanced 6-million-word corpus (Sebastián et al., 2000)– 225,000 words from the EFE Spanish news agency[1], and 200.000 from the Spanish version of the *El Periódico* newspaper. AnCora-Ca consists of 75,000 words from the EFE news agency, 225,000 words from the ACN Catalan news agency[2], and 200,000 words from the Catalan version of the *El Periódico* newspaper. The subset of 200,000 words coming from *El Periódico* corresponds to the same news in Catalan and Spanish, spanning from January to December 2000.

3. Methodology

An incremental process guided the annotation of AnCora, since semantics depends on morphosyntax, and syntax relies on morphology. This procedure made it possible to check, correct and complete the previous annotations, thus guaranteeing the final quality of the corpora and minimizing the error rate.

With respect to the degree of automation, we can distinguish three kinds of annotation processes: full automatic,

[1] `http://www.efe.es`
[2] `http://www.acn.cat`

semiautomatic and manual. First, both corpora were morphologically tagged and disambiguated using automatic linguistic tools (Civit and Martí, 2004b) and were later manually revised throughout the syntactic annotation stage. Base constituents were recognized by means of an automatic shallow parser. Shallow parsing served as starting point for handling the annotation at the full syntax level.
The annotation of deep syntactic information (constituents and functions), strong and weak named entities (NEs), and WordNet nominal synsets was carried out manually. Inter-annotator agreement rates were computed for manual annotation in order to assess the quality of the annotation and, by extension, the appropriateness of the annotation guidelines.
The semantic annotation of verbal predicates was done semiautomatically (Martí et al., 2007). Firstly thematic roles were automatically associated with the syntactic functions on the basis of the verbal lexicons AnCora-Verb-Es and Ancora-Verb-Ca. These lexicons make explicit the mapping between syntax and semantics (Aparicio et al., 2008). A set of manually written rules automatically mapped part of the information declared in these verbal lexicons onto the syntactic structure, which made it possible to tag the treebanks with thematic roles and semantic classes. The output from the automatic stage was either full –both arguments and thematic roles– or partial –with either arguments or thematic roles. This level of annotation was finally revised and completed by hand.

4. AnCora level-by-level

Before proceeding to the description of AnCora on a level-by-level basis, we present the general figures of both corpora. On the one hand, AnCora-Ca contains 483,859 tokens (including punctuation marks); 39,258 types (different word forms); and 27,977 lemmas. On the other hand, Ancora-Es consists of 187,278 tokens; 25,349 types; and 16,867 lemmas.
Tables 1 and 2 show the ten most frequent words in each corpus. Notice that despite the difference in corpus size, the top ten words overlap to a great extent. All them are functional. The first noun does not appear until the 36th and 28th positions, being 'year' both for Spanish (*año*) and Catalan (*any*); and the first verb for Spanish is *ser* 'to be' found in the 11th position, unlike Catalan *haver* 'to have' in the 10th position.

Frequency	%	Lemma
18,284	9.74	el ('the')
11,353	6.05	,
10,220	5.45	de ('of')
5,844	3.11	.
5,539	2.95	que ('that')
4,619	2.46	en ('in')
4,318	2.30	y ('and')
3,599	1.92	uno ('a, an')
2,992	1.59	a ('to')
2,972	1.58	"

Table 1: The top ten words in AnCora-Es

The following subsections present in detail each level of annotation.

Frequency	%	Lemma
48,483	10.02	el ('the')
30,178	6.24	de ('of')
26,295	5.43	,
16,574	3.43	.
12,364	2.56	que ('that')
11,739	2.43	i ('and')
9,839	2.03	un ('a, an')
9,649	1.99	a ('to')
8,074	1.67	del ('of-the')
8,054	1.66	haver ('to have')

Table 2: The top ten words in AnCora-Ca

4.1. Morphological level

This level distinguishes the part of speech (PoS) and minor morphological categories such as gender, number, case, person, time, and mode. Each tag consists of a series of digits: the first corresponds to the main category (e.g. noun), and the second to the subcategory (e.g. common noun). Since all the possible tags are automatically assigned to each word, the morphological disambiguation tool RELAX (Padró, 1998) was used to obtain a single tag-lemma pair for each word. Figure 1 shows the output of the morphological disambiguation process for the Spanish sentence *Si trabajo bajo presión bajo el interés* 'If (I) work under pressure (I) decrease the interest'.

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 1: Output of the morphological annotation

If we take into account the complete tag, AnCora has 280 different labels; whereas 47 tags if only the first two digits are considered. Table 3 shows the relative frequencies of all categories and the most relevant subcategories: proper and common nouns, definite articles and other determiners, auxiliary and main verbs, and the verb *to be*. This verb is set apart because in Spanish and Catalan it can function as either auxiliary (passive voice) or main verb.

4.2. Syntactic level

The two AnCora treebanks result from this annotation level on the basis of constituents and functions. Five main principles underlie the development of these treebanks:

- The only implicit information added was elliptical subjects.
- Constituents were preferred to dependencies (these were obtained in a subsequent stage by a conversion process (Civit et al., 2006)).

PoS tag	AnCora-Es (%)	AnCora-Ca (%)
Adjective	7,25	6.10
Conjunction	5,50	4.80
Definite article	9,74	10.02
Determiner	5,37	4.44
Punctuation mark	12,80	11.71
Common noun	17,87	18.34
Proper noun	5,29	6.10
Pronoun	4,51	4.63
Adverb	3,86	3.19
Preposition	15,19	15.73
Auxiliary verb	0,92	2.91
Main verb	9,54	9.16
ser ('to be')	1,27	0.95

Table 3: PoS tags in AnCora

```
(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 2: Output of the syntactic annotation

- Arguments and adjuncts were not distinguised in the tree structure. They are all sister nodes.
- The surface order was maintained.
- Theory neutral.

A complete parse tree annotated with constituents and functions is illustrated in Figure 2 for the Spanish sentence *La declaración propugnó trabajar por la igualdad social* 'The declaration advocated working for the social equality'. The syntactic annotation falls into three groups according to its relation with the verb: verbal functions, external verbal complements (discourse elements), and verbal modifiers (modality). Table 4 presents relative frequencies of each tag.

Interestingly enough, given that the two Romance languages are pro-drop, a quarter of all subjects are elliptical in AnCora-Ca, and more than half in AnCora-Es. The syntactic annotation makes it possible to obtain easily occurrence frequencies of certain linguistic constructions, such as the frequency of verbs with a postponed subject, which amounts to 7.4% in AnCora-Es, and even more in AnCora-Ca, 11%. From a linguistic point of view, finding explanations for these differences can cast light on constructions which are more specific to one of the two languages. For instance, inaccusativity.

Word order is more flexible in Spanish and Catalan than it is in English. AnCora makes it evident that a large number of different syntactic orders are possible: SV, SVO, SOV, OSV, etc. The figures for AnCora-Ca are 27%, 28%, 2%, 2%, respectively. These figures are helpful both for theoretical and computational linguistics.

4.3. Semantic level

The AnCora corpora are annotated with different kinds of semantic information: the argument structure of verbal predicates and their semantic class (4.3.1.), the thematic role associated with each argument (4.3.2.), NEs both strong and weak (4.3.3.), and WordNet synsets for all nouns (4.3.4.).

4.3.1. Arguments

A semiautomatic process was followed to enrich syntactic functions with their semantic argument, distinguishing Arg0, Arg1, Arg2, Arg3, Arg4, ArgM, ArgA, and ArgL. The first five tags are numbered from less to more obliqueness with respect to the verb; ArgM corresponds to adjuncts; ArgA are external agents (e.g. *Juan pasea al perro.* 'John takes the dog for a walk.'); and ArgL codes complements of light verbs (very often lexicalised, e.g. *dar un beso* 'to give a kiss'). Discourse elements and modality tags do not receive any semantic label. Hence, approximately 90% of syntactic tags have a semantic argument.

Table 5 shows the possible functions that can realise each argument with the relative frequencies in each corpus. 47 combinations of arguments and functions are possible, and 86 combinations of functions, arguments and thematic roles.

4.3.2. Thematic roles

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). Each argument position can map onto specific thematic roles. By way of example, Arg0 can be AGT, CAU, EXP or SRC; and Arg1 can be PAT, TEM or EXT. Figure 4 shows the treebank enriched with arguments and thematic roles.

4.3.3. Named entities

The AnCora corpora were annotated with both strong and weak NEs (Borrega et al., 2007). 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 unit with a PoS tag. Examples of strong NEs are personal names and surnames, book titles, some geographical and country names, dates,

Syntactic type	Syntactic tag	Ancora-Es (%)	Ancora-Ca (%)
Functions	Attribute	5.20	4.39
	Agent complement	0.98	1.08
	Adverbial complement	27.49	23.80
	Direct object	20.30	21.06
	Indirect object	2.33	1.82
	Predicative	1.34	1.70
	Prepositional complement	3.80	5.23
	Subject	30.46	30.43
Discourse elements	Sentence adjunct	1.62	2.42
	Textual element	1.20	1.34
	Vocative	0.02	0.01
Modality	Impersonal tag	0.32	0.39
	Modality	3.54	3.40
	Negation	0.01	0.03
	Passive tag	1.34	2.87

Table 4: Syntactic tags in AnCora

Argument	Syntactic function	AnCora-Es (%)	AnCora-Ca (%)
Arg0	Agent complement, Direct object, Indirect object, Subject	16.72	17.18
Arg1	Adjunct, Direct object, Prep. comp., Subject	33.36	34.38
Arg2	Attribute, Adjunct, Direct object, Indirect object, Predicative, Prep. comp., Subject	13.70	13.45
Arg3	Adjunct, Indirect object, Predicative	0.54	0.41
Arg4	Adjunct	0.68	0.58
ArgM	Adjunct, Predicative	25.80	23.05
ArgA	Prep. comp., Subject	0.01	0.02
ArgL	Adjunct, Direct object, Predicative, Prep. comp., Subject	0.49	0.42

Table 5: Mapping from syntax to semantics

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. Examples from the AnCora-Ca corpus (Figure 3) include person (np0000p), date (W), number (Z), number-currency (Zm), and number-percentage (Zp).

Weak NEs consist of a noun phrase, being it simple or complex. Therefore, they are syntactic elements. Weak NEs do not necessarily have a strong NE 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 or a relational adjective derived from a proper noun are considered weak NEs.

With regard to the semantic types assigned to each NE, six basic semantic categories were distinguished (Table 6).

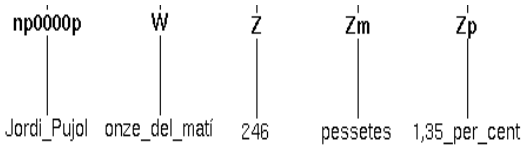

Figure 3: Output of the NE annotation

	NE	AnCora-Es (%)	AnCora-Ca (%)
Strong	Organization	29.84	29.81
	Location	18.10	24.79
	Person	34.10	24.14
	Number	5.07	4.79
	Date	4.08	4.69
	Other	8.81	11.78
Weak	Organization	32.04	34.32
	Location	21.03	26.68
	Person	35.85	25.69
	Other	9.00	12.31

Table 6: Strong and weak NEs

4.3.4. WordNet

The lexical semantic annotation consists in assigning a WordNet sense to each noun. This process was carried out manually. 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:

C2S: The word does not exist in WordNet.
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: a misspelling or a

loanword.

A total amount of 76,307 nouns in AnCora-Ca received a WordNet synset. The number of C2S tokens is 30,078, which points out a need for enlarging the Catalan EuroWordNet. For Spanish no data are available as yet.

5. A fully annotated example

Figure 4 illustrates the multilevel annotated Catalan sentence:
Rigol considerà que en la definició d'aquest nou projecte no es pot oblidar ni passar per alt el nostre passat.
'Rigol considered that in the definition of this new project we can neither forget nor ignore our past.'
At the level below the sentence (S), we find the main constituents containing the type (e.g. sn [NP]), the syntactic function (e.g.-SUJ), the argument (e.g.-Arg0-), the thematic role (e.g. -AGT). Phrases corresponding to a weak NE specify the class by adding its code at the end of the constituent label (e.g. snp for 'persons'). At word level, all tokens are preceded by their PoS and followed by their lemma. Nouns and verbs are additionally annotated:

- Nouns: proper nouns have their PoS enriched with the NE class (e.g. np0000p). Common nouns receive their WordNet synset following the lemma (e.g. 05054071n).

- Verbs: the semantic verb class is indicated after the lemma (e.g. -A2).

```
(
  (S
    (snp-SUJ-Arg0-AGT
      (grup.nom.ms
        (np0000p Rigol Rigol C2S)))
    (grup.verb
      (vmis3s0 considerà considerar-A2))
    (S.F.C.CO-CD-Arg1-PAT
      (conj.subord
        (cs que que))
      (S.F.C.CO
        (S.F.C
          (sp-CC-ArgM-LOC
            (prep
              (sps00 en en))
            (sn
              (espec.fs
                (da0fs0 la el))
              (grup.nom.fs
                (ncfs000 definició definició 05054071n)
                (sp
                  (prep
                    (sps00 d' de))
                  (sn
                    (espec.ms
                      (dd0ms0 aquest aquest))
                    (grup.nom.ms
                      (s.a.ms
                        (grup.a.ms
                          (aq0ms0 nou nou)))
                      (ncms000 projecte projecte 00508925n)))))))))
          (neg-MOD
            (rn no no))
          (morfema.verbal-PASS
            (p0000000 es es))
          (grup.verb
            (vmip3s0 pot poder)
            (infinitiu
              (vmn0000 oblidar oblidar-B2))))
        (coord
          (cc ni ni))
        (S.F.C*
          (infinitiu
            (vmn0000 passar_per_alt passar_per_alt))))
      (sn.j-SUJ-Arg1-PAT
        (espec.ms
          (dp1msp el_nostre el_nostre))
        (grup.nom.ms
          (ncms000 passat passat 10849142n))))
    (Fp . .)))
```

Figure 4: Sentence with all the annotation levels.

6. Conclusions

We have presented two new language resources for Catalan and Spanish: the AnCora corpora, a multilevel annotated corpus freely available from the Web. The systematic procedure followed in the annotation procedure and the linguistically well-founded coding schemes ensure the consistency and reliability of the different annotation levels.
The AnCora corpora are useful resources both to train and to test several NLP systems. The treebanks have been used to build two parsers for Spanish: (Cowan and Collins, 2005) from the first 3LB treebank, and (Carreras et al., 2006).
In future work, we plan to enlarge the annotation of AnCora-Es up to 500,000 words, and to enrich both corpora with coreference information as well as the argument structure of nominal predicates.

Acknowledgements

The development of AnCora has been funded by the following projects: 3LB (FIT-150-500-2002-244), CESS-ECE (HUM2004-21127), PRAXEM (HUM2006-27378-E), and Lang2World (TIN2006-15265-C06-06, `http://gplsi.dlsi.ua.es/text-mess`) from the Spanish Ministry of Education and Science, and the funding given by the Catalan Secretary of Linguistic Policy.
We would also like to thank all the annotators: Joan Aparicio Mera, Oriol Borrega Cepa, Isabel Briz Hernández, Núria Bufí Cabrol, Montserrat Civit Torruella, Maria Jesús Díaz Cabrera, Silvia Garcia Casaseca, Raquel Hernández Bitinas, Marina Lloberes Salvatella, Raquel Marcos, Difda Monterde Puig, Borja Navarro Colorado, Aina Peris Morant, Lourdes Puiggros Casals, Marta Recasens Potau, Alba Rodríguez Vidal, Rita Zaragoza Jove, and Bàrbara Soriano Bautista.

7. References

J. Aparicio, M. Taulé, and M.A. Martí. 2008. AnCora-Verb: A lexical resource for the semantic annotation of corpora. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

O. Borrega, M. Taulé, and M.A. Martí. 2007. What do we mean when we talk about named entities? In *Proceedings of the 4th Corpus Linguistics Conference*, Birmingham, UK.

X. Carreras, M. Surdeanu, and L. Màrquez. 2006. Projec-

tive dependency parsing with perceptron. In *Proceedings of CoNLL-X Shared Task*, New York.

M. Civit and M.A. Martí. 2004a. Building Cast3LB: a Spanish treebank. *Research on Language and Computation*, 2(4):549–574.

M. Civit and M.A. Martí. 2004b. Estándares de anotación morfosintáctica para el español. In *Proceedings of the 9th Ibero-American Workshop on Artificial Intelligence. Iberamia Conference*, pages 217–224, Mexico.

M. Civit, M.A. Martí, and N. Bufí. 2006. Cat3LB and Cast3LB: From constituents to dependencies. In *Advances in Natural Language Processing. Lecture Notes in Computer Science*, volume 4139, pages 141–152. Springer Verlag, Berlin.

B. Cowan and M. Collins. 2005. Morphology and reranking for the statistical parsing of Spanish. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP2005)*, pages 795–802, Vancouver, Canada.

E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2006. OntoNotes: the 90% solution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT-NAACL2006)*, pages 57–60.

N.G. Linh and Z. Zabokrstsky. 2007. Rule-based approach to pronominal anaphora resolution applied on the Prague Dependency Treebank 2.0 Data. In *Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC2007)*, pages 77–81, Lagos, Portugal.

M.A. Martí and M. Taulé, editors. 2008. *CESS-ECE TreeBanks*, Barcelona. Publicacions de la Universitat de Barcelona.

M.A. Martí, M. Taulé, M. Bertran, and L. Màrquez. 2007. Anotación semiautomática con papeles temáticos de los corpus CESS-ECE. *Procesamiento del Lenguaje Natural*, 38:67–76.

L. Padró. 1998. *A Hybrid Environment for Syntax–Semantic Tagging*. Ph.D. thesis, Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya.

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