MiniCors and *Cast3LB*: Two Semantically Tagged Spanish Corpora

Taulé, M.; Civit, M.; Artigas, N.; García, M.; Márquez*, L.; Martí, M.A.; Navarro, B.**

CLiC, Centre de Llenguatge i Computació-Universitat de Barcelona
*TALP, Departament LSI-Universitat Politècnica de Catalunya
** Departament LSI-Universitat d'Alacant
Universitat de Barcelona, Gran Via 585, 08007 Barcelona
mtaule@ub.edu, {civit,nuripa,mar}@clic.fil.ub.es, lluism@lsi.upc.es, amarti@ub.edu

Abstract

In this paper we present two Spanish corpora, *MiniCors* and *Cast3LB*, semantically tagged according to different annotation criteria and objectives. In order to guarantee the quality of the results, we have established a methodology for the development of these corpora. The resulting resources consist of a semantically tagged corpus according to the lexical sample task, and a semantically tagged corpus according to the all words task, both of them defined within the Senseval framework.

1. Introduction

In this paper we present two Spanish corpora, *MiniCors* and *Cast3LB*, semantically tagged according to different annotation criteria and objectives. In order to guarantee the quality of the results, we have established a methodology for the development of these corpora. The resulting resources are a semantically tagged corpus according to the lexical sample task, and a semantically tagged corpus according to the all words task, both of them defined within the Senseval[1] framework. In the corpus for the lexical sample task, *MiniCors*[2], only a word per sentence is tagged (a noun, a verb, or an adjective), whereas in the corpus based on the all words task, *Cast3LB* (Civit and Martí, 2004), all the words are tagged, except those that belong to the functional classes (i.e., prepositions, articles, conjunctions, etc.). The development of these basic resources for the Spanish language constitutes a primary objective, since there is an absence of this kind of resources and an increasing interest on the methods for automatic Word Sense Disambiguation (WSD).

MiniCors has been used to test different lexical sources, which have been evaluated in order to establish which one is the most adequate for semantic tagging tasks. The goal of this process has been to define a semantic tagging methodology in order to systematize the tagging process and to evaluate the quality of the results according to the level of agreement among annotators. Due to the complexity of the task, we have only treated 49 words. For every word a minimum of 200 examples have been tagged.

The *Cast3LB* corpus is part of the project Cast3LB, which also includes syntactic and pragmatic tagging. In this case, due to the quantity of words that had to be tagged, the coverage (all words approach) has prevailed over the quality of the results: a subgroup of sentences has been tagged twice in order to determine the main causes of disagreement and we have developed a tagging handbook in order to avoid inconsistencies.

In section 2 the features of *MiniCors* are described. In 2.1 we describe the methodology followed in the corpus annotation: section 2.1.1 presents the basic features of the lexical source, *MiniDir.2.1*, and section 2.1.2. describes the process of analysis of agreement among annotators. Finally, section 2.2. presents the results of the annotation process. Section 4 deals with the corpus *Cast3LB*, its methodology and main characteristics.

2. Corpus MiniCors

MiniCors is a Spanish corpus tagged in a partial way, where only a sample of words has been tagged. These words have been previously selected according to their frequency of use and polysemy degree (See Table 1). With this corpus we have tested different lexical sources, in order to evaluate them and to determine which one is the most adequate for semantic tagging tasks. The goal was to define a semantic tagging methodology in order to systematize the tagging process and to evaluate the quality of the results according to the level of agreement among annotators. The lexical sources evaluated are the *Diccionario de la Real Academia Española* (*DRAE*) and the *MiniDir.2.1* (Artigas et al., 2003b), which is a lexical source created specifically for WSD tasks and where every sense is linked to EuroWordNet. The results of the evaluation have been significantly better for *MiniDir.2.1*, and it has been the dictionary used as reference for the annotation of *MiniCors*.

Due to the complexity of the task, we have limited our work to the treatment of 49 words of different syntactic categories: 22 nouns, 9 adjectives, and 18 verbs. The *MiniCors* corpus is formed by 13,477 sentences and 565,782 words (with an average of 41.9 words per sentence). The examples have been extracted from the corpus of the EFE Spanish news agency, which includes 289,066 news spanning from January to December of 2000[3]. It is, therefore, a compilation of sentences which belong to a standard language, and, in theory, deal about general subjects and topics[4]. The objective was to obtain 200 sentences for each of the selected words, that is, to obtain a total of 200 examples per word. The context

[1] Senseval is an evaluation exercise that makes possible the comparison of different automatic systems and methods of Word Sense Disambiguation. http://www.senseval.org/

[2] *MiniCors* is a corpus that has been developed in the framework of the project Senseval-3. It will be used as the *gold standard* and the training and evaluation corpora will be created from it.

[3] The available volume of the EFE corpus is 2.814.291 sentences, 95,344,946 words, with an average of 33.8 words per sentence.

[4] The corpus has undergone a previous automatic filtering process in order to remove adjectival and adverbial phrases in which the word to be tagged appears.

considered for each word is larger than a sentence, as it has also been included the previous and following sentences. For each word, we tried to obtain 15 occurrences for each sense. The corpus is marked in XML format and contains about 1,000 examples more for each word, which for the moment have not been tagged. In order to simplify the arbitration process, each example has been tagged by three annotators. The annotation process has been carried out through an interface specifically designed for this task, and a tagging handbook for the annotators (Artigas et al., 2003a).

As regards the polysemy of the selected words, the average of senses per word is 4.5 and, specifically, 4 senses for the nouns subgroup, 6.5 for the verbs and 3.8 for the adjectives. The following table specifies the selected words with their number of senses:

22 nouns		**9 adjectives**		**18 verbs**	
words	senses	words	senses	words	senses
arte	4	brillante	2	actuar	4
autoridad	4	ciego	5	apoyar	4
banda	7	claro	5	apuntar	9
bomba	3	local	2	bajar	5
canal	6	natural	6	canalizar	3
circuito	5	popular	3	conducir	5
columna	8	simple	4	duplicar	2
corazón	6	verde	5	explotar	5
corona	4	vital	3	ganar	8
gracia	5			jugar	5
grano	4			perder	11
hermano	3			saltar	15
letra	5			subir	5
masa	4			tocar	13
mina	4			tratar	12
naturaleza	4			usar	3
operación	4			vencer	7
órgano	3			volar	6
partido	2				
pasaje	4				
programa	3				
tabla	6				

Table 1: List of the selected words and their number of senses.

2.1. Methodology for the development of *MiniCors*

The Senseval competition has highlighted the absence of evaluation of the quality of linguistic resources used for WSD, both of the lexicons and of the tagged corpora. Senseval has focused on the evaluation and comparison of WSD systems and techniques rather than on the linguistic resources. Taking into account that the quality of the linguistic resources determines to a large extent the effectiveness and quality of WSD systems and techniques, our aim has been to define a methodology in order to develop quality linguistic resources. This methodology has implied the simultaneous semantic tagging of the same corpus with different lexical sources (*MiniD.2.1* and *DRAE*) and by three different annotators (so as to facilitate the arbitration task). In short, each word has been tagged by three different lexicographers for every lexical source. The annotator's team was made up of a total of 14 lexicographers with wide experience in the field. We have considered that previous experience was a key feature in order to achieve the maximum possible agreement in the annotation process (Bruce & Wiebe, 1989; Kilgarriff, 1999). In fact, we have considered the degree of annotator's agreement a quality criterion.

The development of *MiniCors* has taken as starting point a previous phase, whose aim was to evaluate two lexical sources of different characteristics, *MiniDir.2.1 and DRAE*, in order to prove which one produced the highest degree of agreement and, therefore, which one was the most adequate for WSD tasks.

MiniDir.2.1 is a dictionary designed for the manual tagging of corpora and, therefore, created specifically for WSD. *DRAE* (*Diccionario de Referencia y Normativo de la Lengua Española*) is a public dictionary of common use.

The objective of this phase was not only to carry out a comparative study of the lexical sources, but also to define a methodology for the evaluation of the agreement degrees, in order to establish a group of categories of agreement among annotators that would reflect the different possible cases that can arise in the annotation process. In short, the aim was to establish a methodology that would enable us to systematize the annotation process and provide at the same time criteria to analyze the degree of agreement among annotators.

Once the lexical sources had been evaluated, we proceeded to the complete annotation[5] of the corpus (in triplicate) with the lexical source that provided the highest results, in this case *MiniDir.2.1*.

2.1.1. The lexical source: *MiniDir.2.1*

MiniDir.2.1 is a dictionary clearly designed for WSD tasks, whose objective is to include a discrete group of senses which are clearly distinguishable, and which do not present the overlapping problems of traditional lexical sources, but at the same time, *MiniDir.2.1.* had to be exhaustive.

In the development of *MiniDir.2.1* we have basically taken into account information extracted from corpora. We have used the corpora from the newspapers *El Periódico* and *La Vanguardia*, with a total of 3.5 millions and 12.5 millions of words respectively, and *Lexesp* (Sebastián et al., 2000), a balanced corpus of 5.5 millions of words, which includes texts on different topics (science, economics, justice, literature, etc.), written in different styles (essay, novel, etc.) and different language registers (standard, technical, etc.). All these corpora are morphologically tagged and disambiguated. The corpora provide quantitative and qualitative information which is essential to differentiate senses and to determine the lexicalization degree.

Apart from the information extracted from corpora, in order to establish and to define the senses we have

[5] In order to systematize the annotation process, we have created a specific interface for the task and a handbook (Artigas et al., 2003a) that specifies the criteria to follow in the annotation.

consulted different traditional lexical sources and two lexical conceptual knowledge bases: WordNet 1.5 (Miller, 1995) and EuroWordNet (Vossen, 1999). The criteria used in the elaboration of *MiniDir.2.1* are listed in (Castelló et al., 2003).
As regards the information of the entries of the dictionary, every sense is organized in the nine following lexical fields:
LEMMA#CATEGORY#SENSE#DEFINITION#EXAMPLE#SYNONYMS#(ANTONYMS)#COLLOCATIONS#SYNSETS
The lexical category is represented by the Eagle tags (Eureka 1989-1995) which have been abridged. In the verbal entries we have also included an additional field with a syntactic category that indicates a classification based on the diathesis alternations that the verb admits.
As regards the field of antonyms, it is only filled in the adjective entries. In the field SYNSET we have established the mapping between each sense and the synset number in the semantic net EuroWordNet (Vossen, 1999). Below, we can see an example of lexical entry:

brillante#AQCS#1#Que brilla: pelo brillante, ojos brillantes
#SIN:reluciente, luminoso, resplandeciente
#ANT:apagado, opaco, mate
#COL:color brillante, ojos brillantes, brillante luz, brillante color, luz brillante
#SYNSET:00219316a/00220071a/00221034a/00221385a/00221761a/00299159a/01697658a/00299159a/00340981a/01383439a/00215174a/00221034a/01697658a#

brillante#AQCS#2#Que destaca por sus cualidades: estudiante brillante, jugador brillante
#SIN:admirable, excelente
#ANT:pésimo, mediocre
#COL:diálogo brillante, futuro brillante, idea brillante, momento brillante, brillante carrera, brillante ejercicio, brillante idea, brillante historial, brillante intervención, brillante labor, brillante porvenir, historial brillante, intervención brillante, carrera brillante, porvenir brillante
#SYNSET:00601428a/01014574a/00852797a#

Figure 1: *Minidir.2.1* Lexical entry

2.1.2. Degrees of agreement

When the corpus has been tagged in triplicate, we have compared the different annotations and we have evaluated the results in order to obtain a disambiguated corpus that we use as *gold standard*[6]. In order to cover all the different agreement degrees, we have established different tags which reflect all the different possible combinations among annotators. Therefore, we have tags that indicate the different types of partial agreement, apart from the disagreement and total agreement tags.
Total agreement takes place when the three annotations match (e.g..: 1, 1, 1 = 1). Partial agreement takes place when not all the annotations match, but an annotation prevails over the others (e.g..: partial agreement 1: 1, 1, 1/2 = 1; partial agreement 2: 1, 1/3, 1/2 = 1; partial agreement 3: 1, 1/2, 1/2 = 1; partial agreement 4: 1/3, 1/2, 1/4 = 1). Minimum agreement takes place when two annotators agree and one does not (e.g..: 1, 1, 2 = 1). Disagreement takes place when any annotator agrees (e.g..: 1, 2, 3 = ?).
All cases of agreement, total, partial and minimum are automatically validated according to the pattern that we have exposed. Only cases of disagreement go to a subsequent arbitration phase.
We have also considered other parameters of analysis:
a) Total minimum agreement that counts all the cases of total agreement among the annotators, and the maximum total agreement, which counts the cases of total agreement and partial agreement among the annotators.
b) Pairwise agreement, which counts the degree of agreement between each pair of annotators. In this case, we have also distinguished among minimum pairwise agreement (cases of total agreement among every pair of annotators) and maximum pairwise agreement (cases of partial agreement among each pair of annotators).

2.2. Results

The table bellow shows the achieved results in the process of annotation of *MiniCors* according to the agreement parameters we have just presented.
The final results consist on 22 nouns, 9 adjectives and 18 verbs in a total of 13,477 examples. In the following graphics (Table 2) we present the global results we have obtained for each category.

	MinTA	TA	MinPA	Max PA	MinA	Dis
N	0.88	0.90	0.92	0.93	0.09	0.01
A	0.80	0.85	0.86	0.90	0.14	0.02
V	0.81	0.83	0.86	0.88	0.15	0.01

MinTA = Minimum Total Agreement
TA = Total Agreement
MinPA= Minimum Pairwise Agreement
MaxPA = Maximum Pairwise Agreement
MinA = Minimum Agreement
Dis = Disagreement

Table 2: Global Agreement degrees in the annotation with *MiniDir* 2.1

3. Cast3LB

The annotated corpus for the all words task, *Cast3LB*, is part of the Cast3LB project (Navarro et al., 2003), which also includes the syntactic and pragmatic (anaphora) annotations. In this case, given the total amount of words that had to be annotated, the coverage of the results has prevailed over quality: a subset of sentences has been annotated twice so as to detect the main causes of disagreement and a handbook of annotation has been created so as to avoid inconsistencies.
Cast3LB is a corpus of 100,000 words (approximately 3,700 sentences) created from two corpora: the CLiC-TALP corpus, a balanced and morphologically annotated corpus containing literary, journalistic, scientific, etc. language, and the corpus of the EFE Spanish news agency corresponding to year 2000. The former contributed with about 75,000 words, while the latter with 25,000.
As for the semantic annotation, the senses used were those defined in the lexicosemantic network of the Spanish EuroWordNet version. In order to make the annotation task easier, a specific interface has been designed, 3LB-SAT (Bisbal et al., 2003).

[6] The training and evaluation corpora for Senseval-3 are obtained from this *gold standard* corpus.

The annotation process has been carried out in two steps. In the first step a subset of the corpus has been selected and annotated twice by two different annotators. The results of this double annotation process have been compared and a disagreement typology in sense assignation has been established. After a process of analysis and discussion, a handbook of annotation has been produced, where the main criteria to follow in case of ambiguity have been described. In the second step, the rest of the corpus has been annotated following the all words strategy. The lexical items annotated are those words with lexical meaning, i.e., nouns, verbs, and adjectives.

From a methodological point of view, and given the fact that EuroWordNet has a high number of senses per word, the strategy followed to annotate has been the assignation of a sense to each occurrence of the word in the corpus, instead of annotating each sentence. Thus, the annotator was able to concentrate on the analysis of a specific word in all its different occurrences in the corpus and, consequently, a better quality and coherence of the results are guaranteed. By default, monosemic words have been automatically assigned the only sense they have in EuroWordNet. Afterwards, its correctness has been checked.

Cast3LB has more that 100,000 words, from which 42,291 have been semantically tagged: 20,467 are nouns, 13,471 are verbs and 8,353 are adjectives.

```
<Annotation id="ejemplo1:EJ1:Annotation4"
type="wrd" start="ejemplo1:EJ1:Anchor2"
end="ejemplo1:EJ1:Anchor3">
<Feature name="label">gato</Feature>
<Feature name="synset">01457160n </Feature>
<Feature name="synset">01458079n </Feature>
<Feature name="synset">06051878n </Feature>
<Feature
name="parent">ejemplo1:EJ1:Annotation5</Featu
re>
</Annotation>

<Annotation id="ejemplo1:EJ1:Annotation5"
type="pos" start="ejemplo1:EJ1:Anchor2"
end="ejemplo1:EJ1:Anchor3">
<Feature name="lema">gato</Feature>
<Feature name="label">ncms000</Feature>
<Feature
name="parent">ejemplo1:EJ1:Annotation6</Featu
re>
</Annotation>
```

Figure 2: *Cast3LB* semantic annotation

4. Conclusions

We have developed two semantically tagged Spanish corpora *MiniCors* and *Cast3LB*. *MiniCors* is formed by 13,477 sentences and 565,782 words. It has been tagged with a dictionary developed specifically for WSD tasks and with criteria of maximum granularity. The tagging process has been carried out in parallel by three annotators, with an agreement degree of 0.90 for nouns, 0.85 for adjectives and 0.83 for verbs. These results guarantee the quality of the corpus. *MiniCors* will be used as the training and evaluation corpus for the lexical sample task of Senseval-3.

Cast3LB is formed by more than 100,000 words from which 42,291 have been semantically tagged. It has been tagged equally by different annotators taking from a starting point a previous phase in which the methodology of the tagging process has been defined. The agreement degree is lower, due to fact that EuroWordNet has been the source used, and it has a higher granularity than *MiniDir 2.1*.

5. Acknowledgments

The development of the *MiniCors* corpus has been possible thanks to the support of the project XTRACT-2 (BFF2002-04226-C03-03).

Cast3lb has been developed thanks to the support of the project FIT-150-500-2002-244.

We also want to acknowledge the linguists of CLiC and UNED who had collaborated in the annotation task.

References

Artigas, N. García, M., Taulé, M. & Martí, M. A. (2003a). Manual de anotación semántica. XTRACT-WP-03/03. CLiC, Universitat de Barcelona.

Artigas, N. García, M., Taulé, M. & Martí, M. A. (2003b). Diccionario MiniDir 2.1. X-TRACT2 WP-03/08. CLiC, Universitat de Barcelona.

Bisbal, E., Molina, A., Moreno, L., Pla F., Saez-Noeda, M. & Sanchis, E. (2003). 3LB-SAT: una herramienta de anotación semántica, Procesamiento del Lenguaje Natural, num. 31, Alicante.

Bruce, R. & Wiebe, J. (1998). Word sense distinguishability and inter-coder agreement. In Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing (EMNLP-98). Association for Computational Linguistics SIGDAT (pp. 53-60). Granada, Spain.

Castelló, G., Artigas, N., García, M., Martí, M.A, Taulé, M. Guía del Diccionario CLIC-Thera. XTRACT2-WP-03/01, CliC, Universitat de Barcelona.

Kilgarriff, A. (1999). 95% Replicability for manual word sense tagging. In Proceedings of EACL'99. Morgan Kaufman Publishers, San Francisco.

Miller, G.A. (1995). WordNet: A Lexical Database. Communications of the ACM, 38: n.11.

Navarro, B., Civit, M., Martí, M.A, Marcos, R. & Fernández, B. (2003). Syntactic, semantic and pragmatic annotation in Cast3LB. In Proceedings of the Shallow Processing of Large Corpora (SProLaC), a Workshop of Corpus Linguistics. Lancaster, UK.

Real Academia Española, Diccionario de la lengua española (2001). 22ª ed., Espasa Calpe, Madrid.

Sebastián, N., Martí, M.A., Carreiras, M.F., & Cuetos Gómez, F. (2000). Lexesp, léxico informatizado del español. Edicions de la Universitat de Barcelona.

Véronis, J. (2001). Sense tagging: does it make sense?. In Proceedings of the Corpus Linguistics Conference. Lancaster, UK.

Vossen, P. (ed.) (1999). EuroWordNet General Document, http://www.hum.uva.nl/~ewn.