SenSem: a Databank for Spanish Verbs[1]

Ana Fernández-Montraveta[1], Gloria Vázquez[2], Irene Catellón[3]

[1] Departament of English Philology, Universitat Autònoma de Barcelona
Ana.Fernandez@uab.es
[2] Departament of English Philology and Linguistics, Universitat de Lleida
gvazquez@dal.udl.es
[3] Departament Linguistics, Universitat de Barcelona
icastellon@ub.edu

Abstract: In this paper we present a databank for Spanish verbs created within the SenSem project. This databank is made up of a lexical database, in which the most 250 frequent Spanish verbs are described from a syntactic and semantic point of view, and an annotated corpus with examples of use for these verbs. A total of 1,122 verb senses have thus been described.

Key words: verb databank, lexical database, annotated corpus, syntactic and semantic interface.

1. The SenSem Project

The SenSem Project (2003-06) aims to build a databank that accounts for the description of verbs in Spanish. This databank is composed of a lexical database, in which the syntactic and semantic behavior of verbs is codified, and an annotated corpus to exemplify the verb senses in the lexical database.

The verbs we describe in the databank are the most frequent 250 verbs in Spanish. We established this list of verbs from the analysis of a journalistic corpus, "El Periódico de Catalunya". The corpus we have annotated belongs also to this register, journalism, even though it is a different one that the one we use to establish frequency. We are very much aware that the use of just one linguistic register implies lesser variation in language use than a better balanced corpus. Nevertheless, we have chosen to work with this register mainly for two reasons, it is fairly easy to obtain huge amounts of data using electronic versions of newspapers and, also, we believe it is, in all probability, the closest to the use of language by a standard speaker.

We have described all the senses for each verb, a total of 1,122 with an average of 4.5 senses per lemma. In each entry we include lexical information, mainly a list of senses with a lexicographic definition, event structure and a list of semantic roles for each sense. Also, we include information regarding the syntactic and semantic behavior of verbs, for example the information regarding subcategorization that is

[1] This project has been funded by grant (BFF2003-09456).

established from the annotated sentences. Thus, the information that has been acquired from the annotated examples will only be found for those senses found in the corpus, which amount to a total of 749[2].

For the description of each lexical item, we have taken into account all the sentences ascribed to a sense. The total number of sentences annotated for each lemma is 100, so the corpus will be composed of 25,000 sentences, with a total of approximately 700,000 words. We believe that the data that can be extracted from this corpus is a good starting point to describe the behavior of the verbs dealt with. Nonetheless, we consider that more sentences would be desirable in order to extract statistical data. A more diversified typology of registers in order to achieve a wider diversification of syntactic and semantic patterns would also be useful.

The annotation of sentences has been carried out at three levels: the verb as a lexical unit, sentence constituents in relation to the verb and sentence semantics. More precisely, the annotation process includes the disambiguation of the lexical item, the tagging of the syntactic structure (non-lexical syntactic categories, the constituent's heads and the establishment of the argument or adjunct status of the participants), the characterization of participants in terms of semantic roles and the discrimination of the semantics of the construction. The latter aspect is the brand mark of this project and it is precisely what sets it apart from other similar projects being currently developed for Spanish (Subirats and Petruck 2003; García de Miguel y Albertuz, 2005).

All this information allows us to describe the syntactico-semantic interface with great detail, both at the verb and sentence levels. The information collected is of interest not only for linguistic studies but also for those applications that require the understanding of sentences beyond purely syntactic analysis. In areas such as automatic understanding, semantic representation and also in automating learning systems, a resource of this kind can be highly valuable.

In what follows we describe the composition of the corpus and annotation process (section 2), the design of the lexical entry (section 3) and the design of the search interface (section 4).

2. The Corpus: Composition and Annotation

The SenSem corpus[3] is made up of 25,000 sentences belonging to the journalistic register and annotated at the syntactic and semantic levels [Alonso et al. 2005]. These 25,000 sentences exemplify the behavior of the 250 verbs most commonly used in Spanish according to the statistic data extracted from a corpus with over 13,000,000 words. We have randomly extracted 100 sentences for each verb from the electronic version of "El Periódico de Catalunya". We have not considered for annotation periphrastic uses of verbs, idiomatic expressions or collocations. Currently, the

[2] Of these verbs, 435 have more than 9 sentences associated.

[3] The corpus can be consulted online and downloaded in XML at http://grial.uab.es/search.

annotation phase, which has been carried out manually, has been completed and we are in the revision phase.

As for the range of the annotation, we have annotated those arguments and adjuncts that are directly related to the verb. Those elements that are beyond the scope of the verb, that is, extrasentential elements, such as some adverbs, have not been considered. Let's see an example of this:

Bono agradeció *a éste su gesto porque "*representa *a más de seis millones de catalanes"*.
Bono *thanked* him for his gesture because "he *represents* more than six million Catalan people".

As we can see, we find two verbs in these sentences. In order to annotate the verb "agradecer" (to thank), we have considered the whole sequence of words. The subordinate clause is considered an only constituent and, therefore, its internal structure has not been taken into account. Nevertheless, if we are to annotate the verb "representar" (to represent), only the fragment included in the inverted commas will be taken into account.

Sentences are annotated at three levels: the lexical level where we associate a sense with each verb form; the constituent level, in which we characterize syntactically and semantically the participants in the event; and the sentence level, in which we annotate the sentence meaning as a whole.

At the lexical level, each sentence is associated with the sense it is exemplifying. Senses have been determined aprioristically.

At the constituent level, for each participant in the sentence we annotate the syntactic category (NP, PP, etc.), syntactic function (subject, direct object, indirect object, etc.) and the type of argument relation it holds with the verb, namely if it is an argument or an adjunct. Arguments are defined as participants in the event and, from our point of view, they are part of the verb's lexical semantics. For this reason they are associated with a semantic role (agent, theme, source, etc.). We have also annotated the heads of the arguments, since this information is valuable in order to acquire selection preferences once the corpus has been completely annotated.

At the sentential level, we grouped several aspects of interest in order to characterize the meaning of a sentence. For example, we annotate construction semantics according to the argument being focalized (anticausative, passive, etc.); or the aspectual information it conveys (e.g. habitual sentence; stative sentence, etc). Other specific constructions accounted for are: reflexive constructions, reciprocal constructions and datives of interest. This type of information is useful when we have to specify the argument structure. For example, it might happen that an active construction in the present is interpreted as expressing a habitual construction and, therefore, it is no longer interpreted as designating a specific event:

Entre enero y abril, acceden diariamente a la sala de pinturas 25 personas, distribuidas en grupos de cinco.

Between January and April, 25 people come in daily into the exhibition room, in groups of 5.

Codification of semantic information is of special relevance in order to fully understand statements and can carry consequences in any application that requires understanding. This is the case in automatic translation since one particular kind of meaning may have specific formal marks in some languages and a different one in another. If this type of semantics is not taken into account one might overlook important parts of meaning.

For a complete guide to the annotation criteria see Vázquez et al 2005. Currently, we are revising the annotation. At the time of this writing, we have revised only 36% of all the annotated sentences, so there is a considerable amount of work still pending. The process of revising and correcting is organized in two phases. First, we apply some automatic routines in order to detect errors which are systematic or else can be easily foreseen. For example, some tags can be considered contradictive and cannot, therefore, be used together (e.g. if a constituent is labeled as an adjunct it cannot be given a semantic role tag).

Secondly, we carry out a manual revision in order to make sure the changes realized automatically are correct and also to detect errors and inconsistencies that cannot be found automatically. These do not follow any specific pattern and are mostly due to human error. By the end of the year 2006, when the project will finish, we expect to have automatically revised 100% of the corpus and manually corrected 60%. In the manual checking process a judge goes through the annotated sentences once and a different person then revises his/her corrections. In case of discrepancy, a third person intervenes.

3. The Lexical Entry

The lexical entry devised for verbs presents basically two types of information: 1) the strictly lexical information that has been predefined prior to the annotation of the sentences and 2) the syntactic and semantic information which has been obtained from the data acquired from the corpus.

3.1 The Predefined Lexical Information

The codification of what we consider as strictly lexical information includes:

a) a lexicographic definition
b) the semantic roles
c) the event structure
d) a link to WordNet
e) a list of synonyms

The first step in the development of this project was the establishment of a list of senses for each lemma and the corresponding definitions. It was absolutely necessary to start from an established list of senses to which we could associate each sentence. The lexicographic sources consulted in order to create the above-mentioned list of senses have been the paper-based edition of the “Diccionario Salamanca de la Lengua Española” and the on-line dictionaries “Diccionario de la Real Academia de la

Lengua Española" and "Diccionarios de El Mundo"[4]. We have not taken into account the uses considered archaic or very restricted. On the other hand, even though we have included in our dictionary the uses of verbs as auxiliaries, the collocations and phrases in which the verb participates and sentences exemplifying them are not to be found in the corpus since we believe further study should be carried out in order to properly annotate this specific type of use.

Semantic roles will allow us to describe the semantic relation between the participants in the event and the verb. Moreover, it has proven methodologically convenient to reach a consensus a priori with regard to what semantic roles are present in each verb sense in order to unify the annotation of sentences. This became necessary because the semantic characterization of participants was a major source of disagreement (20.8% of the annotations) [Alonso et al 2005]. The general list of semantic roles comes from the proposal carried out within the project VOLEM [Fernández et al. 2002], developed jointly with other research groups; it has been modified according to the specific needs that have arisen as the annotation process was being performed. Nevertheless, the majority of the semantic labels used can be found in the most common bibliography about the topic. [Fillmore 1968, Anderson 1971, Cook 1979, Larson 1984, Van Valin 1993, Barker et al. 1997].

Event structure has been defined to complete the semantic description at the lexical level. Furthermore, each sense has been linked to Wordnet versions 2.1 and EuroWordNet 1.6 [Fellbaum 1998, Vossen 1999]. Lastly, we intend to include all senses described in the database that hold a synonymy relation.

All the information that has been predefined, especially that related with sense distribution and semantic roles, has been modified occasionally during the annotation process whenever the sentences being dealt with required it. Every change has been submitted to the opinion of two judges. As for the redefinition of senses, modifications have consisted mostly in the aggregation of a new sense whenever we came across a sentence that couldn't be assigned one of the senses available and the merging of two senses whenever the distinction between them proved inconsistent when assigning sentences to them. As far as semantic roles are concerned, some new tags have been incorporated into the original list associated with a sense when it was observed that a semantic type of constituent appeared in the event often enough to justify its inclusion. We have considered it a clear indicator of the argument nature of a participant.

3.2 Information Extracted from the Corpus

With regard to the information acquired from the annotated corpus, the lexical entries include the following information:

a) Subcategorization patterns: under this label we include those structures in which arguments are expressed in terms of syntactic categories and functions, regardless of order.

[4] http://www.rae.es; http://www.elmundo.es/diccionarios.

b) Realization frames: with this term we refer to the actual realization of subcategorization patterns taking into account word order, the information about semantic roles and the different possibilities for syntactic realization; for example a nominal phrase can be realized as a noun, a pronoun, a relative pronoun, etc. and it can also be elided in Spanish.
c) The sentential meanings (constructions).
d) Examples: each of the above is linked to the sentences in the corpus that exemplify it.
e) Selection restrictions.
f) Prepositions.
g) Information about frequency regarding the number of sentences linked to a sense, to a subcategorization frame and to a construction.

Currently, we have finalized the first phase in acquiring the information listed above, which includes all of the sections except for e) and f). Next, we will explain in greater detail the interface to visualize the data above.

4. The Online Databank

The interface we present in this paper (http://grial.uab.es/adquisicio) takes information from three different databases[5]:

a) The database in which we have included what we call proper lexical information[6] (see section 3.1).
b) The database in which we have codified, for each verb sense, the syntactic and semantic information extracted from the corpus (see section 3.2).
c) The annotated corpus (see section 2).

The consultation of the data is carried out first by selecting a lemma and then subspecifying the sense for which we want the information shown (figure 1).

Once we have selected the sense we want to see, we are shown a new page divided into two parts. In the upper part, we can see the content of the lexical data base and the information regarding frequencies for that sense (figure 2).

In the lower part, we see two lists that present the syntactic and semantic information acquired from the corpus: the list of subcategorization frames and all the constructions in which that verb has been found. The number before each element in the list refers to the frequency of appearance in the corpus of a particular frame or construction (figure 3).

By selecting any element from either of the two lists, we will be able to see all the corresponding realization frames which, in turn, are linked to all the possible sentences in the corpus exemplifying a specific realization frame (figure 4). Each one of the sentences contains a link to the annotation that allows us to see the annotation graphically represented (figure 5).

---

[5] All the databases are available in MySQL.
[6] This data base can also be consulted online at http://grial.uab.es/sentits.

Fig. 1. Access interface to the databank

acabar	
ID:	1
Definición:	Finalizar algo.
RS:	[ag/caus,t-af,circ]
EE:	evento
Wordnet:	00211850v
Sinónimos:	
N° ocurrencias en el corpus:	68/100

Fig. 2. Lexicographic information for the lexical entry: “acabar 1”.

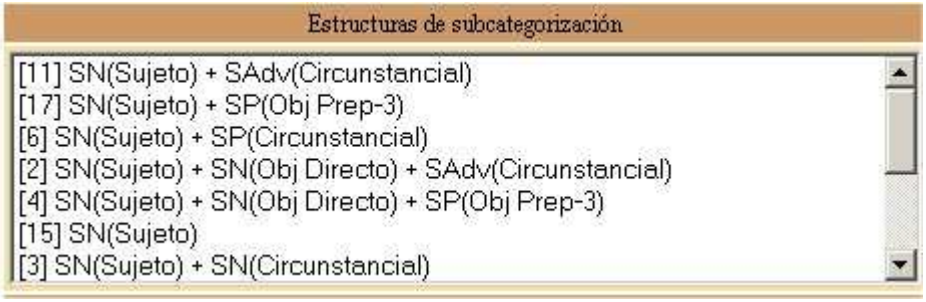

Fig. 3. Subcategorizaton frames for “acabar 1”.

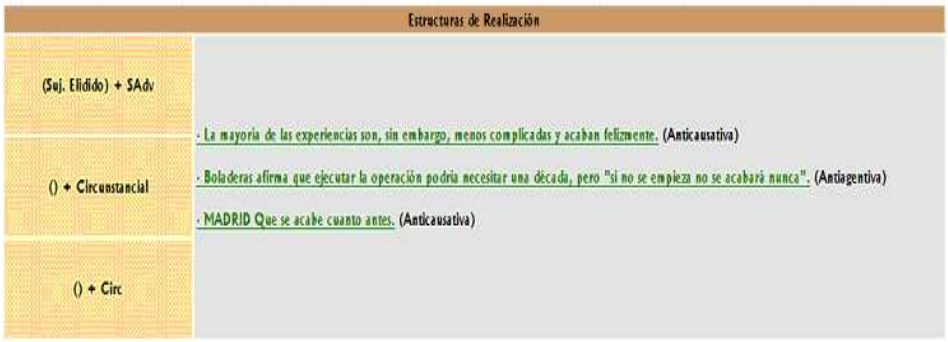

Fig. 4. Realization frames corresponding to the first subcategorization frame for the verb "acabar 1".

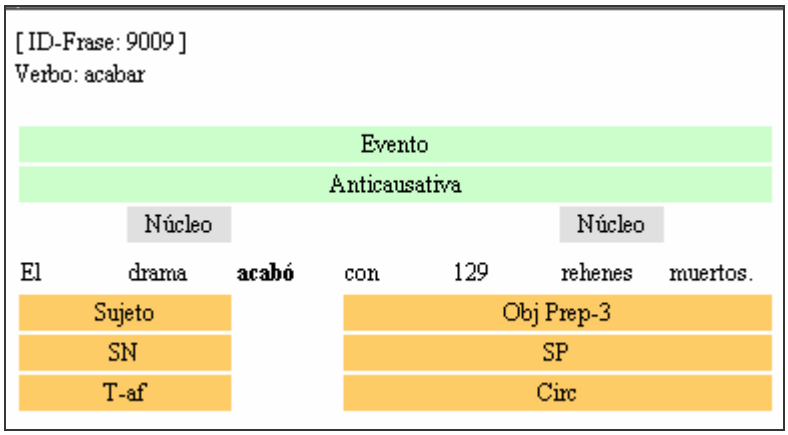

Fig. 5. Annotation of the sentence "El drama acabó con 129 rehenes muertos" ("The drama ended with 129 dead hostages").

5. Conclusions and Future Work

In this paper we have presented a databank for Spanish, which has been developed in the project SenSem. Its main characteristic is that it includes information defined using lexicographic criteria as well as information acquired in a corpus. In order to consult this database we have developed an interface that allows users access to both the lexical verb data base and the corpus, which have been linked.

Since it is a verb-centered resource, we have paid special attention to the description of the interface between syntax and semantics. Nevertheless, we have also provided other kinds of information such as the definition of the lexical items and the synonyms. Moreover, each verb entry includes a link to WordNet, providing for

information in the databank to be linked to other corpora also connected to this standard and vice versa.

In the short term, several tasks remain pending, some relating to the corpus and its exploitation as a source for linguistic information and others that have to do with the lexical description. For example, we still have to complete the task of creating the synonyms lists. We believe this task to be of importance since, apart from its undoubted value as a complement to verb description, it will allow us to establish a control over the information described by comparing the data provided in each of the entries that are considered synonymous.

As for the corpus, we have to complete the revision of the annotation process, an arduous task that requires considerable time and effort. Some aspects of acquisition also remain unfinished. More precisely, we need to carry out the task of preposition extraction and the acquisition of information regarding selection restrictions.

We are also working in the exploitation of the databank for the construction of other resources. For example, the list of subcategorization frames obtained from this corpus is currently being used as the basic information for the construction of a grammar for Spanish.

In the long run, we mean to improve the resource by adding or refining the types information considered so far. For example, we would like to complete the semantic information by specifying the aspect of the sentence as a whole, and thus taking into account negation adverbs, verb tense and other adjuncts that contribute to meaning. We would also like to complete the corpus by enlarging the registers included in the texts to account for more literary uses of language.

References

1. Alonso, L., J.A. Capilla, I. Castellón, A. Fernández and G. Vázquez (2005). "The Sensem Project: Syntactico-Semantic Annotation of Sentences in Spanish", *Proceedings of the International Conference RANLP*, p. 39-46. Borovets, Bulgaria.
2. Anderson, J. M. (1971). *The Grammar of Case: Towards a Localistic Theory*. Cambridge: Cambridge University Press.
3. Barker, K., T. Copek and S. Szpakowicz (1997). "Systematic construction of a versatile case system". B. Boguraev, R. Garigliano, J. I. Tait (ed.), *Natural Language Engineering*. Cambridge: Cambridge University Press, p. 279-315.
4. Cook, W. A. (1979). *Case Grammar: Development of the Matrix Model*. Washington: Georgetown University Press.
5. Fellbaum, Ch. Eds. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.
6. Fernández, A., P. Saint-Dizier, G. Vázquez, F. Benamara and M. Kamel (2002). "The VOLEM Project: a Framework for the Construction of Advanced Multilingual Lexicons", *Proceedings of the Language Engineering Conference*, p. 89-98.
7. Fillmore, C. J. (1968). "The case for case". E. Bach, R. T. Harms (ed.), *Universals in Linguistics*. New York: Holt, Rinehart, Winston.
8. García de Miguel, J. M. and F. J. Albertuz (2005) "Verbs, Semantic Classes and Semantic Roles in the ADESSE project". *Interdisciplinary Workshop on Verb Features and Verb Classes*. Saarbrücken.
9. Larson, M. L. (1984). *Meaning-Based Translation: A Guide to Cross-Language Equivalence*. Lanham: University Press of America.

10. Subirats-Rüggeberg, C. and M. R. L. Petruck (2003). “Surprise: Spanish FrameNet!” Presentation at Workshop on Frame Semantics, *Proceedings of the International Congress of Linguists*, Praga.
11. Van Valin, R. D. (1993). *Advances in Role and Reference Grammar*. Amsterdan: John Benjamins Publishers.
12. Vázquez, G., A. Fernández and L. Alonso (2005). "Description of the Guidelines for the Syntactico-semantic Annotations of a Corpus in Spanish". Angelova, G., K. Bontcheva, R. Mitkov, N. Nicolov (ed.), *International Conference Recent Advances in Natural Language*. Shoumen (Bulgaria):, p. 603-607.
13. Vossen P. (1999) “EuroWordNet as a multilingual database”. Wolfgang Teubert (ed) TWC.