The SenSem Project: syntactico-semantic annotation of sentences in Spanish

Alonso, Laura* **Capilla, Joan Antoni†** **Castellón, Irene*** **Fernández-Montraveta, Ana**** **Vázquez, Gloria†**

*Department of Linguistics, Universitat de Barcelona, Spain
{lalonso,icastellon}@ub.edu

**Department of English and German Philology, Universitat Autònoma de Barcelona, Spain
ana.fernandez@uab.es

†Department of English and Linguistics, Universitat de Lleida, Spain
{jcapilla,gvazquez}@dal.udl.es

Abstract

This paper presents SenSem, a project[1] that aims to systematize the behavior of verbs in Spanish at the lexical, syntactic and semantic level. As part of the project, two resources are being built: a corpus where sentences are associated to their syntactico-semantic interpretation and a lexicon where each verb sense is linked to the corresponding annotated examples in the corpus. Some tendencies that can be observed in the current state of development are also discussed.

1 Introduction

The SenSem project aims to build a databank of Spanish verbs based on a lexicon that links each verb sense to a significant number of manually analyzed corpus examples. This databank will reflect the syntactic and semantic behavior of Spanish verbs in naturally occurring text.

We analyze the 250 verbs that occur most frequently in Spanish. Annotation is carried out at three different levels: the verb as a lexical item, the constituents of the sentence and the sentence as a whole. The annotation process includes verb sense disambiguation, syntactic structure analysis (syntagmatic categories, including the annotation of the phrasal heads, and syntactic functions), interpretation of semantic roles and analysis of various kinds of sentential semantics. It is precisely this last area of investigation which sets our project apart from others currently being carried out with Spanish (Subirats and Petruck, 2003 and García De Miguel and Comesaña, 2004).

Abstracting from the analysis of a significant number of examples, the prototypical behavior of verb senses will be systematized and encoded in a lexicon. The description of verb senses will focus on their properties at the syntactico-semantic interface, and will include information like the list of syntactico-semantic frames in which a verb can possibly occur. In addition, selectional restrictions will be automatically inferred from the words marked as heads of the constituents. Finally, the usage of prepositions will be studied.

The conjunction of all this information will provide a very fine-grained description of the syntactico-semantic interface at sentence level, useful for applications that require an understanding of sentences beyond shallow parsing. In the fields of automatic understanding, semantic representation and automatic learning systems, a resource of this type will be especially valuable.

In the rest of the paper we will describe the corpus annotation process in more detail and examples will be provided. Section 2 offers a general overview of other projects similar to SenSem. In section 3, the levels of annotation are discussed, and the process of annotation is described in section 4. We then proceed to present the results obtained to date and the current state of annotation, and we put forward some tentative conclusions obtained from the results of the annotation thus far.

2 Related Work

As shown by Levin (1993) and others (Jones et al., 1994; Jones, 1995; Kipper et al., 2000; Saint-Dizier,

[1] Databank Sentential Semantics: "Creación de una Base de Datos de Semántica Oracional". MCyT (BFF2003-06456).

1999; Vázquez et al., 2000), syntax and semantics are highly interrelated. By describing the way linguistic layers inter-relate, we can provide better verb descriptions since generalizations from the lexicon that previously belonged to the grammar level of linguistic description can be established (lexicalist approach).

Within the area of Computational Linguistics, it is common to deal with both fields independently (Grishman et al., 1994; Corley et al., 2001). In other cases, the relationship established between syntactic and semantic components is not fully exploited and only basic correlations are established (Dorr et al., 1998; McCarthy, 2000). We believe this approach is interesting even though it does not take full advantage of the existing link between syntax and semantics.

Furthermore, we think that in order to coherently characterize the syntactico-semantic interface, it is necessary to start by describing linguistic data from real language. Thus, a corpus annotated at syntactic and semantic levels plays a crucial role in acquiring this information appropriately.

In recent years, a number of projects related to the syntactico-semantic annotation of corpora have been carried out. The length of the present paper does not allow us to consider them all here, but we will mention a few of the most significant ones.

FrameNet (Johnson and Fillmore, 2000) is a lexicographic resource that describes approximately 2.000 items, including verbs, nouns and adjectives that belong to diverse semantic domains (communication, cognition, perception, movement, space, time, transaction, etc.). Each lexical entry has examples extracted from the British National Corpus that have been manually annotated. The annotation reflects argument structure and, in some cases, also adjuncts.

PropBank (Kingsbury and Palmer, 2002; Kingsbury et al., 2002) is a project based on the manual semantic annotation of a subset of the Penn Treebank II (a corpus which is syntactically annotated). This project aims to identify predicate-argument relations. In contrast with FrameNet, the sentences to be annotated have not been pre-selected so examples are more varied.

Both FrameNet and Propbank work with the use of corpora, although their objectives are a bit different. In FrameNet, a corpus is used to find evidence about linguistic behavior and to associate examples to lexical entries, whereas in Propbank, the objective is to enrich a corpus that has been already annotated at a syntactic level so that it can be exploited in more ambitious NLP applications.

For Spanish, only a few initiatives address the syntactico-semantic analysis of corpus. The DataBase "Base de Datos Sintácticos del Español Actual" (Muñiz et al., 2003) provides the syntactic analysis of 160.000 sentences extracted from part of the ARTHUS corpus of contemporary texts. Syntactic positions are currently being labeled with semantic roles (García de Miguel and Comesaña, 2004).

FrameNet-Spanish (Subirats and Petruck, 2003) is the application of the FrameNet methodology for Spanish. Its target is to develop semantic frames and lexical entries for this language. Each verb sense is associated to its possible combinations of participants, grammatical functions and phrase types, as attested in the corpus.

The SenSem project provides a different approach to the description of verb behavior. In contrast with FrameNet, its aim is not to provide examples for a pre-existing lexicon, but to shape the lexicon with the corpus examples annotated. Another difference from the FrameNet approach is that the semantic roles we use are far more general, they are related to syntactic functions, and are less class-dependent.

Finally, to the best of our knowledge, no large-scale corpus annotation initiative associates semantics to sentence such as their aspectual interpretations or types of causativity.

3 Levels of annotation

As mentioned previously, we are describing verb behavior so only constituents directly related to the verb will be analyzed. Elements beyond the scope of the verb (i.e. extra-sentential elements such as logical linkers, some adverbs, etc.) are disregarded. The following is an example of the scope of annotation contemplated:

...El presidente, que ayer **inició** una visita oficial a la capital francesa, hizo estas declaraciones…

*...The president, who **began** an official visit to the French capital yesterday, stated…*

Were we annotating the verb *iniciar* –begin– we would ignore the participants of the main sentence and only take into account the elements within the clause.

If we were annotating the verb *hacer* –make– we would annotate the subject to include the entire relative clause, with the word "*presidente*" as the head of the whole structure. The relative clause will not be further analyzed.

Sentences are annotated at three levels: sentence semantics, lexical and constituent level.

3.1 Sentential semantics level

At this level, different aspects of sentential semantics are accounted for. With regard to aspectual information, three types of meanings are distinguished: *eventive*, *procedural* or *stative.* Apart from aspectual information, we also annotate sentential level meanings using labels like *anticausative, antiagentive, impersonal, reflexive, reciprocal* or *habitual.* This serves to further specify the argument structures of each verb sense.

3.2 Lexical level

At the lexical level, each example of a verb is assigned a sense. We have developed a verb lexicon in which the possible senses for a verb are defined, together with its prototypical event structure and thematic grid, and a list of synonyms and antonyms and its related synsets in WordNet (Fellbaum, 1998).

Various lexicographic sources have been taken as references to build the inventory of senses for each verb, mainly the *Diccionario de la Real Academia de la Lengua Española* and the *Diccionario Salamanca de la Lengua Española*. Less frequent meanings are discarded, together with archaic and restricted uses.

This inventory of senses for each verb is only preliminary, and can be modified whenever the examples found in the corpus prove the existence of a distinct sense which has not been considered. Different senses imply either different thematic grids, different event structures, different selectional restrictions or different subcategorizations.

3.3 Constituent level

Finally, at the constituent level, each participant in the clause is tagged with its constituent type (e.g.: *noun phrase, completive, prepositional phrase*) and syntactic function (e.g.: *subject, direct object, prepositional object*).
Arguments and adjuncts are also distinguished. Arguments are defined as those participants that are part of the verb's lexical semantics. Arguments are assigned a semantic role describing their relation with the verb (e.g.: agent, theme, and initiator). In SenSem, each sense is associated with a prototypical thematic grid describing the possible arguments a verb may take, but, as in the case of senses, this thematic grid is only preliminary and is modified when corpus examples provide enough evidence.

The head of the phrase is also signaled in order to acquire selectional restrictions for that verb sense.

Sometimes, information that has been considered relevant in that it may alter some other information declared at a different level has also been included; for example, negative polarity or negative adverbs are also indicated.

4 Annotation process

The SenSem corpus will describe the 250 most frequently occurring verbs in Spanish. Frequency has been calculated in a journalistic corpus. For each of these verbs, 100 examples are extracted randomly from 13 million words of corpora obtained from the electronic version of the Spanish newspapers, *La Vanguardia* and *El Periódico de Catalunya*. The corpus has been automatically tagged and a shallow parsing analysis has been carried out to detect the personal forms of the verbs under consideration. We do not take into account uses of the verb as an auxiliary. We also disregard any collocations or idioms in which the verb might participate.

The manual annotation of examples is carried out via a graphical interface where the three levels are clearly distinguished.

The interface displays one sentence at a time. First, when a verb sense is selected from the list of possible senses, its prototypical event structure and semantic roles are displayed for the annotator to take into account. Then, the clause is assigned its aspectual semantics, and constituents are identified and analyzed by selecting the words that belong to it. The head of the arguments and its possible metaphorical usage are also signaled in order to facilitate a future automatic extraction of selectional restrictions. Finally, annotators specify any applicable semantics at clause level (e.g.: anticausative, reflexive, stative, etc.), and state any particular fact that they consider might be of use in future revision and correction processes.

The distribution of the corpus among annotators has evolved since the earlier stages of the project. In an initial stage, when the annotation guidelines were not yet consolidated, each of the 4 annotators was given 24 different sentences of the same verb, plus 4 common sentences that were separately annotated by all of them. Later on, these sentences were compared in order to identify those aspects of the annotation that were unclear or prone to subjectivity, as explained in the following section. In the current stage, the annotation guidelines have been well established. Annotators work with sets of 100 sentences corresponding to a single verb. All annotations are revised and any possible errors are corrected.

The final corpus will be available to the linguistic community by means of a soon to be created web-based interface.

5 Preliminary Results of Annotation

At this stage of the project, 77 verbs have already been annotated, which implies that the corpus at this moment is made up of 7,700 sentences (199,190 words). A total of 900 sentences out of these 7,700 have already been validated, which means that a corpus of approximately 25,000 words has already undergone the complete annotation process.

5.1 Data analysis

In this section we describe the information about verb behavior that can be extracted from the corpus in its present state. We have found that, out of the 199,190 words that have already been annotated, 182,303 are part of phrases which are an argument of the verb and 16,887 are adjuncts.

With regard to aspectuality, there is a clear predominance of events (74.26% of the sentences) over processes (20.67%) and states (8.96%).

As concerns syntactic functions, seen in Table 1, the most frequent category is direct object, with a significant difference in subjects. This is not surprising if we take into account that Spanish is a pro-drop language. However, prepositional objects are less frequent than subjects, and indirect objects are also scarce.

function	**Ratio**
direct object	39.83 %
subject	22.57 %
circumstantial	23.16 %
prepositional obj. t	12.65 %
indirect object	1.97 %

Table 1. Distribution of syntactical functions in the annotated examples.

The distribution of semantic roles can be seen in Table 2. Themes are predominant, as would be expected given that the most common semantic role is that of direct object. Within the different types of the SR theme, unaffected themes (moved objects) appear most frequently.

role	**Ratio**
Not- affected theme	53.47 %
affected theme	14.36%
agent and cause	14.02%
initiator	2.97%

Table 2. Distribution of semantic roles in the annotated examples.

At the constituent level, the semantic role chosen for each phrase is often predictive of the other labels of that phrase, following what was expected from linguistic introspection: agents tend to be noun phrases with subject function, themes tend to be noun phrases with subject or object function (if they occur in a passive, antiagentive, anticausative or stative sentence), etc.

Thus, the associations made between labels in different levels have been used as a first step to semi-automate the annotation process: once a role is selected, the category and function most frequently associated with it and its role as a verb argument are pre-selected so that the annotator only has to validate the information.

5.2 Inter-annotator agreement

In order to measure inter-annotator agreement, four sentences of 59 verbs have been annotated by 4 different judges so that divergences in criteria could be found. These common sentences were used in the preliminary phase with the aim of both training the annotators and detecting points of disagreement among them. This comparison has helped us refine and settle the annotation guidelines and facilitate the subsequent revision of the corpus.

In order to detect these problematic issues, we calculated inter-annotator agreement for all levels of annotation. An overview of the most representative values for annotator agreement can be seen in Table 3.

We determined pair wise proportions of overall agreement, that is, the ratio of cases in which two annotators agreed with respect to all cases.

category	**agreement**	**kappa**
eventual semantics		
event	*66%*	.11
state	90%	.33
process	76%	.06
argumentality		
argument	82%	.54
adjunct	64%	.46
semantic role		
initiator	70%	.37
agent	84%	.81
cause	91%	.89

experiencer	97%	.92
theme	68%	.43
affected theme	74%	.55
non-affected theme	70%	.34
goal	79%	.70
syntactic function		
agentive complement	100%	1.00
subject	87%	.83
direct object	80%	.63
indirect object	77%	.79
prepositional object – 1	67%	.65
prepositional object – 2	66%	.28
prepositional object – 3	78%	.24
circumstantial	62%	.42
predicative	76%	.16
syntactic category		
noun phrase	78%	.67
prepositional phrase	72%	.53
adjectival phrase	88%	.69
negative adverbial	100%	1.00
adverbial phrase	77%	.54
adverbial clause	68%	.66
gerund clause	72%	.65
relative clause	82%	.16
completive clause	95%	.93
direct speech	96%	.95
infinitive clause	94%	.98
prep. completive clause	96%	.44
prep. infinitive clause	81%	.57
personal pronoun	97%	.81
relative pronoun	98%	.96
other pronouns	94%	.82

Table 3. Inter-annotator agreement for a selection of annotated categories

In addition, we also obtained the kappa coefficient (Cohen, 1960), which gives an indication of stability and reproducibility of human judgments in corpus annotation. The main advantage of this measure is that it factors out the possibility that judges agree by chance. Kappa measures range from k=-1 to k=1, with k=0 when there is no agreement other than what would be expected by chance, k=1 when agreement is perfect, and k=-1 when there is systematic disagreement. Following the interpretation proposed by Krippendorf (1980) and Carletta et al. (1996), for corpus annotation, kappa>0.8 indicate good stability and reproducibility of the results, while k<0.68 indicates unreliable annotation.

As a general remark, agreement is comparable to what is reported in similar projects. For example, Kingsbury et al. (2002) report agreement between 60% and 100% for predicate-argument tagging within Propbank, noting that agreement tends to increase as annotators are more trained. In SenSem, the level of annotation that is comparable to predicate-argument relations, semantic role annotation, is clearly within this 60%-100% range.

It is noteworthy that the values obtained for the kappa coefficient are rather low. After a close inspection, we found that these low values of kappa are mainly due to the fact that the annotation guidelines were still not well-established at this point of annotation, and that annotators were still under training. This led us to further describe and exemplify cases detected as having a low agreement value once the preliminary exploration of the corpus had concluded. As a result, we expect values for kappa to increase in future evaluations.

Agreement within aspectual interpretations of sentences is very close to chance agreement. The stative interpretation seems to be more clearly perceived than the rest. Events and processes at times seem to be confused.

It can be seen that there is not a consensus about what an adjunct is. Therefore, the prototypical subcategorization frame for each verb sense has been provided, making it easier for annotators to identify arguments associated with a verb and to label the rest of constituents as adjuncts. A clear distinction has also been made between constituents dominated by a verb (arguments or adjuncts) and those beyond clausal scope.

With semantic roles, linguistic intuition seems to play an important role. There seems to be perfect agreement for very infrequent roles (indirect cause, instrument, location, not shown in the table as space was lacking). More frequent roles show a higher level of disagreement: initiators are significantly less clearly perceived than agents or causes (note differences in k agreement). It is also clear that fine-grained distinctions are more difficult to perceive than coarse-grained ones, as exemplified by low agreement within the superclass of theme.

Among syntactic functions, the agentive complement of passives presents perfect agreement. Agreement is also high for subjects and indirect objects, but the distinction between different kinds of prepositional objects and circumstantial complements

is not clearly perceived. Therefore, a clearer decision-making procedure was established in the annotation guidelines to distinguish among these.

Finally, agreement is rather high for some syntactic categories: pronouns, adverbs of negation, adjectival complements, completive clauses, infinitive clauses and direct speech present k > .7 and ratios of agreement over 90%. However, major categories present a rather high ratio of disagreement, as well as those categories that are mostly considered adjuncts.

6 Conclusions and Future Work

The linguistic resource we have presented constitutes an important source of linguistic information useful in several natural language processing areas as well as in linguistic research. The fact that the corpus has been annotated at several levels increases its value and its versatility.

The project is in its second year of development, with still a year and a half to go. During this time we intend to continue with the annotation process and to develop a lexical database that will reflect the information found in the corpus. We are aware that the guidelines established in the annotation process are going to bias, to a certain extent, the resulting resources, but nevertheless we believe that both tools are of interest for the NLP community.

All tools developed in the project and the corpus and lexicon themselves will be available to all researchers who might have interest in exploiting them.

References

(Carletta et al. 1996) J. Carletta, A. Isard, S. Isard, J. C. Kowtko, G. Doherty-Sneddon and A. H. Anderson, HCRC Dialogue Structure Coding Manual, HCRC Technical report HCRC/TR-82, 1996.

(Cohen 1960) J. Cohen, A coefficient of agreement for nominal scales. in Educational & Psychological Measure, 20, 1960, pp. 37-46.

(Corley et al. 2001) S. Corley, M. Corley, F. Keller, M. W. Crocker, S. Trewin, Finding Syntactic Structure in Unparsed Corpora in Computer and the Humanities, 35, 2001, pp. 81-94.

(Dorr et al. 1998) B. Dorr, M. A. Martí, I. Castellón, Spanish EuroWordNet and LCS- Based Interlingual MT. Proceeding of the ELRA Congress. Granada, 1998.

(Fellbaum 1998) C. Fellbaum, A Semantic Network of English Verbs, in Christiane Fellbaum (ed.). WordNet: An Electronic Lexical Database, MIT Press, 1998.

(Garcia de Miguel & Comesaña 2004) J.M. Garcia de Miguel and S. Comesaña, Verbs of Cognition in Spanish: Constructional Schemas and Reference Points, in A. Silva, A. Torres, M. Gonçalves (eds) Linguagem, Cultura e Cogniçao: Estudos de Linguística Cognitiva, Almedina, 2004, pp. 399-420.

(Grishman et al. 1994) R. Grishman, C. Macleod and A. Meyers. 1994. Comlex Syntax: Building a computational lexicon. Proceedings of COLING.

(Johnson &Fillmore 2000) C. Johnson and C. J. Fillmore, The FrameNet tagset for frame-semantic and syntactic coding of predicate-argument structure. Proceedings NAACL 2000, Seattle WA, USA, 2000, pp. 56-62.

(Jones 1994) D. Jones (ed), Verb classes and alternations in Bangla, German, English and Korean. Memo nº 1517. MIT, Artificial Intelligence Laboratory, 1994.

(Jones 1995) D. Jones, Predicting Semantics from Syntactic Cues --- Evaluating Levin's English Verb Classes and Alternations. UMIACS TR-95-121, University of Maryland, 1995.

(Kingsbury & Palmer. 2002) P. Kingsbury and M. Palmer, From Treebank to Propbank. Third International Conference on Language Resources and Evaluation, LREC-02, Las Palmas, Spain, 2002.

(Kingsbury et al. 2002) P. Kingsbury, M. Palmer and M. Marcus., Adding Semantic Annotation to the Penn TreeBank. Proceedings of the Human Language Technology Conference. San Diego, California, 2002.

(Kipper ET AL. 2000) K. Kipper, H. T. Dang and M. Palmer, Class-Based Construction of a Verb Lexicon. AAAI-2000 Seventeenth National Conference on Artificial Intelligence, Austin, TX, USA, 2000.

(Krippendorf 1980) K. Krippendorf, Content analysis: an introduction, Sage, 1980.

(Levin 1993) B. Levin, English Verb Classes and Alternations: A Preliminary Investigation. The University of Chicago Press, 1993.

(McCarthy 2000) D. McCarthy, Using semantic preferences to identify verbal participation in role switching alternations. Proceedings of NAACL 2000, Seattle, WA, USA, 2000.

(Muñiz ET AL. 2003) E. Muñiz, M. Rebolledo, G. Rojo, M.P. Santalla and S. Sotelo, Description and Exploitation of BDS: a Syntactic Database about Verb Government in Spanish, in Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov, Nikolai Nikolov (eds.). Proceedings of RANLP 2003. Borovets, Bulgaria, 2003, pp. 297-303.

(Saint-Dizier 1999) P. Saint-Dizier, Alternations and verb semantic classes for French: analysis and class formation . Saint-Dizier, P. (ed.). Predicative Forms in Natural Languages and Lexical Knowledge Bases. Holanda. Kluwer: 139-170.

(Subirats-Rüggeberg & Petruck 2003) Subirats-Rüggeberg, C. y M. R. L. Petruck, Surprise: Spanish FrameNet! Presentation at Workshop on Frame Semantics, Proceedings of the International Congress of Linguists, Praga., 2003.

(Vázquez et al. 2000) G. Vázquez, A. Fernández and M. A. Martí, Clasificación verbal. Alternancias de diátesis, Universitat de Lleida, 2000.