The Sensem Corpus: a Corpus Annotated at the Syntactic and Semantic Level

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

*Department of Linguistics, Universitat de Barcelona, Spain
**Department of English and German Philology, Universtitat Autònoma de Barcelona, Spain
†Department of English and Linguistics, Universitat de Lleida, Spain
††Department of Computer Science, Universidad Nacional de Córdoba, Argentina

e-mail: icastellon@ub.edu; ana.fernandez@uab.es; gvazquez@dal.udl.es; alemany@famaf.unc.edu.ar;jcapilla@dal.udl.es

Abstract

The primary aim of the project SENSEM (Sentence Semantics, BFF2003-06456) is the construction of a Lexical Data Base illustrating the syntactic and semantic behavior of each of the senses of the 250 most frequent verbs of Spanish. With this objective in mind, we are currently building an annotated corpus consisting of sentences extracted from the electronic version of the newspaper *El Periódico de Catalunya*, totalling approximately 1 million words, with 100 examples of each verb. By the time of the conference, we will be about to complete the annotation of 25,000 sentences, which means roughly a corpus of 800,000 words. Approximately 400,000 of them will have been revised. We expect to make the corpus publicly available by the end of 2006.

1. Introduction

The corpus we present in this paper is being annotated as part of the SENSEM project. Ultimately, we seek to obtain a databank consisting of a lexicon of Spanish verbs where each verb sense will be described as exemplified in the corpus with respect to its syntactic behavior, the participants in the event and the constructions. Information will be inferred from the annotated sentences and each observed pattern will be associated to the set of corpus sentences that exemplify it.

In the project we work with the most frequent 250 verbs of Spanish. Frequency was calculated from the occurrences of verbs in a journalistic corpus of 13 million words. We extracted 100 examples randomly for each verb from the electronic version of the newspaper *El Periodico de Catalunya*. The periphrastic uses of these verbs have no been taken into account.

The first step in the annotation process of a sentence consists in determining the corresponding verb sense. Then, each constituent is assigned a category and syntactical function, and, in the case of arguments, also a semantic role. Finally, the semantics of the whole construction is provided, tagging the corresponding aspect and the meaning of the syntactic frames along the lines of construction grammar (Fillmore 1988, Goldberg 1995). This last aspect is precisely what distinguishes this project from other current corpus annotation projects for Spanish (Subirats and Petruck, 2003; García De Miguel y Comesaña, 2004) and some projects of English (Propbank P. Kingsbury et al. 2002, a and b). Metaphoric uses of both constituents and verb senses are also marked.

Using the annotated corpus, we will create a verb lexicon that captures generalizations from what has been annotated. The description of verbs will focus on the syntactical-semantic interface. It will comprise information about the constructions in which a verb is to be found, as well as the syntactic and semantic characterization of sentence participants. Information about prepositions and selectional preferences will be included as well.

The description of the syntactical-semantic interface provided by this databank can be implemented in natural language processing applications that require an understanding of sentences beyond syntactic analysis. In the areas of semantic representation and machine learning, this kind of resource is also very valuable.

In this paper we will describe the process of annotation together with some examples. The following section depicts the overall process, describing the resources available to human annotators to increase the consistency of the resulting annotation. Then, we present some of the difficulties in the annotation process, and discuss some approaches to solving them and correcting the errors they produce. We finish with some conclusions and future work.

2. Annotation Process

The manual annotation of corpora is a costly process in which a certain percentage of errors is inevitable. In order to increase the consistency of the corpus and to make the task easier for human annotators, we have created a verb lexicon, providing the prototypical

Aktionsart and semantic roles for each sense, and a friendly annotation tool that allows automatic pre-assignation of some annotation labels.

2.1. Verb lexicon

The first step in the process of annotation consists in assigning a verb sense to the verb in the sentence being annotated.

The inventory of verb senses available for each verb is stored in the form of a verb lexicon, in which we have specified the most common senses of the verbs dealt with in this project (see Figure 1). In order to do so, we have used previous work on the description of verb items and their syntax-semantic interface (Fernández et al 2004). Verb senses are associated to their prototypical aspect and role template in order to guide the annotator assigning this information and thereby increase the consistency between annotators.

Sentido	Ejemplo	RS	EE
1.- Concernir a alguien o a algo.	'*La medida afecta a los pensionistas / Esta ley afecta a las zonas verdes de las ciudades'.	[t,t-af]	estado
2.- Impactar, producir una impresión en una persona.	'El suceso me afectó considerablemente / Se afectó enormemente con la noticia del terremoto en Asia'.	[caus,t-af]	evento
3.- Simular una actitud.	'Le gusta afectar una tranquilidad que no tiene.'	[ag,t]	proceso
4.- Modificar o alterar una cosa de una forma determinada.	'Los incendios afectan a todo el ecosistema del bosque./ El granizo afectó por completo a unas 10 hectáreas de frutales.'	[caus,t-af,ma]	evento

Figure 1: List of *afectar*—affect—verb senses.

This lexicon does not constitute a fixed inventory of verb senses, since it is sometimes refined and tuned, after reaching a consensus, in order for it to meet the requirements of the annotation process. Modifications are carried out in two ways: by expanding the list with senses that had been discarded beforehand, but that have been reflected in the form of corpus examples; and by refining the sets of definitions and/or semantic roles whenever the ascription of examples to a particular meaning has not found a clear and distinct choice within the list of available senses.

The establishment of senses has been made according to syntactic and semantic criteria, more precisely we have considered different sucategorization frames, different semantic roles and selectional preferences as indicators of different senses. Other properties such as differences in the Aktionsart of a predicate or the meaning of the syntactic frame in terms of construction grammar are also taken into account.

2.2. Pre-assignation of labels

The annotator has to delimit first-order constituents for the verb, as is described in Vázquez et al. (2005). Constituents considered as arguments are assigned a semantic role, selected from the set of roles prototypically associated with the verb sense. If an annotator considers that an argument is best described by a different role from the ones supplied, or that a new role which had not been considered should now be taken into account due to its frequency of appearance in the examples, a request is placed in the area designed for comments in the annotation interface, and the issue is given proper consideration during the review process.

Once semantic roles have been assigned, a category and syntactic function are automatically pre-selected. For example, for the role *Agent* the category *Noun Phrase* and the function *Subject* are pre-selected, for *Finality* the category is *Prepositional Infinitive Clause,* and the function, *Prepositional Object*. If the pre-selection is correct, the annotator only has to validate it; if it is not, it must be manually refined or re-annotated.

Pre-assignation were created taking into consideration the most frequent co-occurrences of semantic roles with functions and categories in the first step of the annotation process.

2.3. Annotation tool

In order to carry out the annotation process, an interface has been developed for annotators to have all the available resources at hand. The resources available via the interface are diverse. In the first place, the annotator selects one of the verb senses available for the verb in the verb lexicon. Then, the prototypical aspect and role template associated to the verb sense are displayed. Once a semantic role is selected, the default values for categories and functions are automatically pre-selected. Finally, it allows the edition of the sentences found in the raw corpus, discarding those not fitted to be annotated.

The interface allows annotators to review the annotation associated to each sentence in two ways: a text format where all the information is exhaustively declared, and a graphical mode that facilitates the visualization for annotators and reviewers (see Figure 2).

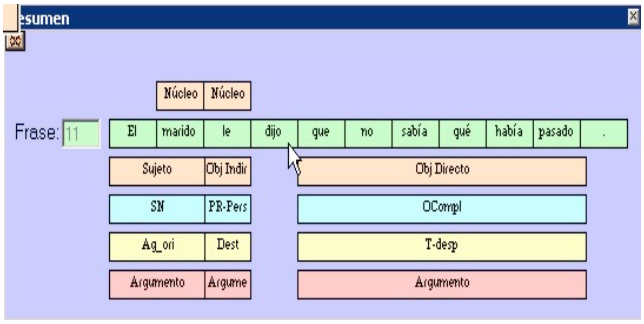

Figure 2: Annotation summary: graphical mode

Additionally, an interface to facilitate the search of the annotated data has also been implemented (Vázquez et al. 2006), and the corpus has been exported to XML format (see figure 3). Both these utilities are of much help to obtain summaries of tendencies in the data and detect inconsistencies and errors in the revision phase.

```
<s ID='5264' semor1='Evento' semor2='Antiagentiva'
anotado='1-23' verbo='9' lema_verbo='Anunciar'
sentido='Anunciar_1'>
<phr id='1' cat='SN' fs='Circunstancial'>
<w Id='1' forma='El'>
<w Id='2' forma= 'año'>
<w Id='3' forma='pasado'>
</phr>
<w Id='4' forma=','>
<phr id='2' rs='Med' cat='SP' fs='Obj Prep-3'
Argumento='1'>
<w Id='5' forma='en'>
<w Id='6' forma='el'>
<w Id='7' forma='Dogc' nucleo='1'>
</phr>
<phr id='3' cat='part'>
<w Id='8' forma='se'>
</phr>
<phr id='4' cat='verbo'>
<w Id='9' forma=' anunció'>
</phr>
<phr id='5' rs='T-desp' cat='SN' fs='Sujeto'
Argumento='1'>
<w Id='10' forma='una'>
<w Id='11' forma='tercera'>
<w Id='12' forma=' línea' nucleo='1'>
<w Id='13' forma='de'>
<w Id='14' forma='P3'>
</phr>
</s>
```

Figure 3: XML annotated corpus instance. Translation: "*Last year, a third line of P3 was announced in the DOGC*".

3. Difficulties in Annotation

The most difficult questions that have arisen in the annotation process concern the labeling of information regarding aspect and construction semantics. Semantic role labeling has also proved to be conflictive. Finally, some discrepancies and errors are found in the delimitation of the annotation range and assigning category and function. Many of these errors are detected by manual inspection of the sentences; others are automatically detected by scripts working on the XML version of the corpus (Alonso et al. 2006). The percentage of errors detected in the SENSEM corpus is about a 20 %.

3.1. Aspectual and Construction semantics

There are some differences found in the application of the annotation criteria concerning aspect and construction semantics. The fact that verb senses are associated to their prototypical aspect makes annotation more consistent, but some exceptions can still be found. In the revision phase, we detect those cases that do not present the expected set of features, for example, the following sentence has been tagged as *event* when it should be a *process* (a):

(a)
Toni **busca** hacerse respetar por sus compañeros […]
*Tony **seeks** the respect of his workmates […]*

It is also frequent to find incompatible combinations of aspectual and sentence semantics, for example, *state* and *antiagentive*. These combinations are clear errors and will be automatically detected and corrected in the XML version of the corpus in the revision phase.

Construction semantics is sometimes interpreted differently by different annotators (about 5.8% of the detected errors). Habitual and anticausative sentences constitute two of the most problematic constructions. The tag *habitual* is assigned to stative sentences expressing iterative events, as in example (b):

(b)
En Juribga , ciudad del interior marroquí, cada verano **se celebra** el mercado de los italianos, donde los emigrantes revenden lo que han traído .
*At the Italian market **celebrated** each summer in Juribga, a Moroccan interior city, emigrants resell what they have brought back with them.*

3.2. Delimitation of the annotation scope

Differences have been observed in the way annotators consider whether certain constituents belong to the sentence being annotated or not (about 3.8% of the detected errors). This affects above all subordinate clauses (example (c)) and adverbs modifying the whole sentence (example (d)):

(c)
La ministra de Medio Ambiente **anunció** ayer el inmediato relevo […], al tiempo que se pone en marcha una auditoría interna "para clarificar la situación".
*The minister for Environment **announced** yesterday the immediate dismissal [….], while at the same time an internal audit has been started in order to clarify the situation*.

(d)
Afortunadamente, en eso las cosas **han cambiado**.
*Fortunately, things **have changed** with respect to that.*

Some errors have also been found in the delimitation of constituents. For example, some annotators have analyzed all the constituents within subordinate clauses, despite that fact that only first-level constituents need to be analyzed.

3.3. Semantic roles, categories and functions

Thanks to the delimitation provided by the verb lexicon, the inventory of roles for each verb sense is quite homogeneous, and the cases where annotation differs (about 3.8% of the detected errors) are commented and dealt with in the revision phase.

The most frequent errors are those concerning the assignation of categories (50% of the errors) or functions (21,2% of the total), mostly due to the lack of revision of the features that are automatically pre-selected. For example, it often happens that some constituents that receive the semantic role of *theme* are left with their automatically selected *Noun Phrase* tag, although they are Prepositional Phrases. *Themes* are also found with their automatic tag *Direct Object* in anticausative constructions, where they function as subjects.

The rest of the errors are related to verb sense disambiguation (7,7%) and head detection (7,7%).

4. Conclusions and Future Work

The linguistic resource presented in this paper is an important source of linguistic information, useful for various NLP applications, as well as for research in linguistics. The fact that the corpus is annotated simultaneously at various levels of linguistic analysis increases its value and versatility. This can be especially useful for applications in the fields of Natural Language Understanding and semantic representation, as well as for systems applying machine learning techniques.

We are currently in the beginning of the third year of the project, with still one year ahead. We are about to finish the annotation phase, the revision phase has already started and we are also extracting and acquiring information from the corpus to build the verb lexicon SENSEM. We are also developing a grammar exploiting the linguistic knowledge encoded in the annotation, like subcategorization frames and argumental structure.

Besides the corpus and lexicon, other resources resulting from this project are annotation and search interfaces (Fernández et al. 2006). A set of heuristics for automating the process of construction annotation has been carried out (Vázquez et al. 2004). A typology of annotation errors, together with the approach to correct them semi-automatically has been also developed (Alonso et al. 2006).

All the tools and resources developed in the project will soon be available for the research community at http://grial.uab.es/projectes/sensem.

5. References

Alonso, L., Capilla, J.A., Castellón, I.; Fernández, A.; Vázquez, G.(2005) "La semántica oracional del español: perspectiva desde el léxico", *International Conference Recent Advances in Natural Language Processing*. G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov eds. Shoumen (Bulgaria)

Alonso, L., Castellón, I., Tincheva, N. (in press), "Detección automática de errores en el corpus SenSem" *Asociación Española de Lingüística Aplicada, AESLA-06.*

Fernández, A., G. Vázquez, I. Castellón (2004) "Sensem: base de datos verbal del español". G. de Ita, O. Fuentes, M. Osorio (ed.), /IX Ibero-American Workshop on Artificial Intelligence, IBERAMIA/. Puebla de los Ángeles, Mexico:, p. 155-163.

Fernández, A., G. Vázquez, D. Teruel (in press) "Interfaz de explotación del corpus Sensem" *Asociación Española de Lingüística Aplicada, AESLA-06.*

Fillmore, C. J. (1988) "The Mechanisms of "Construction Grammar", /BLS/, 14, 35-55.

García de Miguel, J. M. and S. Comesaña (2004), "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, pp. 399-420.

Goldberg, A. E. (1995) /Constructions: A Construction Grammar Approach to Argument Structure/, Chicago, Illinois: Universitand Chicago Press.

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

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

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*, Prague.

Vázquez, G., L. Alonso, I. Castellón, A. Fernández Montraveta (2004). "A Set of Heuristics for Semantic Sentence Disambiguation for Spanish", *4th International Conference on Language Resources and Evaluation (LREC 2004)*. Lisboa, Portugal.

Vázquez, G., Fernández, A. and Alonso, L. (2005) "La semántica oracional del español: perspectiva desde el léxico" *International Conference Recent Advances in Natural Language Processing*. G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov eds. Shoumen (Bulgaria)

Acknowledgements

Acknowledgements: This research was supported by project SENSEM, funded by the Spanish Ministry of Sciency and Technology (grant n.BFF2003-06456).