Guidelines for the syntactico-semantic annotation of a corpus in Spanish

Vázquez, Gloria* **Fernández-Montraveta, Ana**** **Alonso, Laura†**

*Department of English and Linguistics, Universitat de Lleida, Spain
gvazquez@dal.udl.es
**Department of English and German Philology, Universitat Autònoma de Barcelona, Spain
ana.fernandez@uab.es
†Department of Linguistics, Universitat de Barcelona, Spain
lalonso@ub.edu

Abstract

The aim of the SenSem project[1] is to build a databank that reflects the syntactic and semantic behavior of Spanish verbs. This databank will eventually consist of a verbal lexicon linked to a significant number of examples from corpus. These examples are being manually analyzed following the guidelines presented here. We will describe the levels of analysis associated with each example, focusing on the decisions taken to bridge the gap between the theoretical view characteristic of a verbal lexicon and the more practical perspective required in corpus annotation.

1 Motivation

Corpus annotation has become a crucial research area in the last decade. Annotation of texts at a morphological level provides satisfactory results for induction of automatic morphological analyzers (taggers). Other levels of language, however, such as syntax, semantics or discourse still remain to be resolved.

At a syntactic level, the simplest task consists of the automatic identification and labeling of phrases or *chunks*, which can be satisfactorily achieved for most languages with basic NLP resources, including Spanish. In contrast, full parsing of clauses, identifying the syntactic function of syntactic constituents, is a much more complex task. In English, significant results have been achieved using both symbolic (Lin 1998) and statistical (Collins 1996, Eisner 1996) approaches.

For languages with a less fixed word order the problem has been tackled with interesting results, especially using induction of probabilistic grammars from a large treebank (Collins et al. 1999). For Spanish, full parsing is a problem still to be solved because of the relatively unconstrained word order and the realization of constituents (Moreno et al. 2000), and because there is a lack of rich linguistic resources wherefrom grammars could be induced.

Finally, semantic analysis is still more difficult to solve than syntax. Many efforts have been devoted to the analysis of lexical semantics through the task of Word Sense Disambiguation for a number of languages, including Spanish, (Senseval 1998 – 2004), but the problem is still far from solved. In recent years, much attention is being paid to the task of semantic role annotation for English (CoNLL 2004, 2005). To our knowledge, no large-scale effort has been carried out to determine meaning at clausal level.

In order to achieve deeper interpretation of clauses in Spanish, a reference corpus annotated with the relevant information is needed. Applying data-driven techniques to such corpora, tools and resources for automated analysis can be obtained that in turn can be used to pre-annotate further corpora.

In this article we describe the annotation criteria, developed within the Sensem project framework, which we have designed in order to create a databank (a lexicon linked to manually analyzed

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

corpus examples) that reflects the syntactic and semantic behavior of Spanish verbs.

The rest of the paper is organized as follows. In the following section we explain the project. In Section 3 we give an overview of the annotation process, and in further sections the various levels of annotation are described, focusing on the decisions that have been taken to systematize the relation between theory and fact. We finish with some conclusions and future work.

2 The SenSem project

The aim of the SenSem project is to build a reference corpus for Spanish annotated at syntactic and semantic levels, along the lines of FrameNet Spanish (Subirats and Sato 2004) and ADESE (García de Miguel and Comesaña 2004). A notable difference that sets the SenSem corpus apart from similar projects for Spanish is that it carries out the semantic annotation of clauses, a task which had previously been addressed at a theoretical level only. This level of analysis aids in completing the interpretation of clauses. For example, it is very useful for interlingua-based multilingual applications (Vázquez et al. 2000).

As a result of this project, a corpus of approximately 1,000,000 words will be created. This corpus will consist of sentences containing a tensed form of one of the 250 most frequent verbs of Spanish, so that at least 100 instances of each verb are available. These sentences have been randomly selected from a corpus of approximately 13,000,000 words of the electronic versions of the Spanish newspapers *El Periodico* and *La Vanguardia.* We chose the journalistic register because it provides a high number of examples and reflects standard language usage, but a future development of this project takes into account the need to diversify the corpus.

Instances of each verb will be related to the entry of the corresponding verb in a computational lexicon, which will be one of the two main results of the project, the other being the annotated corpus itself. The fact that the corpus is explicitly related to a verbal lexicon presents two main advantages. In the first place, the lexicon acquires the authenticity that is proper of corpus-based methods, thus overcoming the bottleneck of hand-made resources. On the other hand, machine learning methods can be systematized in relation to a structured resource, so that the inferred knowledge is more understandable and transportable to all kinds of applications.

A major problem in the SenSem project has been to bridge the gap between traditional grammatical concepts and the actual phenomena found in a corpus from real language. In this paper we describe the annotation guidelines used, which aim to bring the theoretical insights to the annotation of the actual examples found in corpus. The final goal of the work presented here is to provide annotators with procedures as objective as possible to deal with phenomena found in corpus.

Examples are annotated at three levels:

1) the **verb** is assigned to one of a pre-defined list of senses, specifying metaphorical usage if necessary,
2) **constituents** are labeled with their category, and syntactico-semantic relation with the verb (syntactic function and semantic role, respectively); their argument status is also specified, and heads of arguments are marked, together with any metaphorical usages, and
3) some **clause-level semantics** are tagged, like aspect or patterns of focalization of participants.

3 Overall annotation process

The text has been morphosyntactically analyzed (Carreras et al. 2004) to segment it in sentences and detect those that contain a personal form of any of the verbs collected in the lexicon.

Spanish journalistic sentences are typically very long, much longer than in English, and subordinate clauses are frequent. Thus, in the annotation process, we annotate only those participants which are dominated by the verb under consideration. For example, if we are annotating the verb *iniciar*– initiate–, in the sentence below only the underlined parts are taken into account:[2]

...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…*

We also disregard any additional internal clauses these constituents may have. In the

[2] All examples between marked between ellipses have been taken from the corpus.

sentence above, the subject of the verb *hacer* – make – will be annotated 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.

The annotation process refers to assignment of an interpretation to the whole sentence or to the clause dominated by the verb (Section 4) and to each of its dependents (Section 5). The verb is also analyzed as a lexical item, by assigning it one of a list of senses pre-determined in the lexicon. The list of possible senses is kept in an independent database in which the event structure, the grid of semantic roles and the link to Spanish WordNet (EuroWordNet v. 1.5) has been declared for each sense. The unit of this lexicon is the sense, so the number of entries is in fact larger –572 senses for 250 verbs.

During the process of annotation, annotators evaluate the adequacy of the information collected by the lexicon and propose any improvement they may consider necessary, such as creating new senses, collapsing them or assigning different roles. Also, the information which has been declared as prototypical is susceptible to being modified according to the specific semantics of the clause. Thus, a lexical item that prototypically expresses an event might, under specific conditions, express a state.

In the next sections we are going to go into further detail of the different levels of the annotation process.

4 Sentence-level tagging

Two kinds of clause-level semantics have been distinguished: one which concerns the aspectual information expressed in the clause (Section 4.1), and another which further specifies particular configurations of participants (Section 4.2).

4.1 Aspectual semantics

Following traditional proposals in aspectual research (Comrie 1976, Vendler 1957, Pustejovsky 1995), we distinguish between three types of classes:

- Events, those actions in which the logical culmination is implied. Verbs such as *put* or *finish* are considered events.

  ...El diálogo **acabará** hoy…

  *...The conversations **will finish** today...*
- Processes, those actions that do not have an implicit limit; they are dynamic actions that go on through a stretch of time with the same properties at any interval. Verbs such as *eat* or *live* express a process.

  …cuando le preguntaron de qué **había vivido** hasta aquel momento ...

  *...when he was asked what he **had been living on** until then...*
- A state denotes relationships between an entity and a quality, or between an entity and a context or between two entities. Verbs such as *consist* or *come close* (where movement is not implied) are considered states.

  ...El gasto de personal **se acerca** a los 2.990 millones de euros...

  *...Personnel expenses **come close** TO 2,990 million euros...*

As we have previously mentioned, lexical aspect is indicated for every lexical item in the lexical database. When a sense is chosen for a verb, the information regarding its Aktionsart is automatically assigned. Annotators can adjust it if they consider that the contextual elements modify the verb's aspectuality. We must take into account that we are annotating clauses and, therefore, some participants in the action might alter the Aktionsart.

For example, some processes are limited, that is to say they express an event when they are modified by a "bounded" object. For example, a verb such as *write*, which is lexically a process, gives an eventive reading when uttered in a clause such as *write a letter*.

Sometimes, it is the semantic type of one of the arguments that changes the lexical aspectual information. This is the case of procedural movement verbs which lexically are processes but that can be limited when the destination of the movement is expressed. When a verb like *walk* is realized together with the goal of the movement, it conveys an event instead of a process (*walk to the fence*).

We are aware that another factor that can also change the aspect of a clause is the verb tense. Nevertheless, we do not consider tense as an element to take into account when analyzing the aspectuality of the clause since we believe it

should be considered at a different level. The only exception to this is the use of present to express a habitual reading (Section 4.2). In such cases, we consider clauses as expressing states, or even processes, when they are expressing an iterative reading. The verb *act* has assigned at a lexical level the label of *process* but in a clause such as the example below it is considered a *state*:

…un árabe no compra este tipo de cosas, ni **actúa** así...

*..Arabs do not buy this sort of thing, nor do they **act** this way...*

Another example is the verb *open*. Lexically it has been described as an event but it is labeled as a process when expressing an iterative reading:

…para que no **abran** más locales nocturnos en el distrito durante un año...

*...so that no more night clubs are **opened** in the district for a year...*

4.2 Construction Semantics

In addition to the information about aspect, we are also concerned about other aspects of sentential semantics that more specific to a particular configuration of elements.

We believe syntactic configurations always convey a meaning which is different to the meaning expressed by the same elements arranged differently. A speaker of a language always chooses a particular arrangement of elements for communicative purposes (Goldberg 1995).

In order to describe this level of sentential meaning, various labels related to focalization of arguments, reference binding and aspectuality are provided, as we will see next.

On the one hand, we have distinguished constructions according to which element constitutes the focus of communication. First, we have considered *anticausative* constructions. In Spanish an anticausative construction is typically a pronominal structure in which the participant upon which communicative intention falls is the entity undertaking the action and not the cause that has triggered it.

… las perspectivas que se le **abren** a Catalunya tras la llegada del PSOE al Gobierno...

*... the political horizon **opened up** in Catalunya by the instalment of the PSOE political party in government...*

Secondly, we also include passive constructions, some of which we have grouped together under the antiagentive tag. It is the equivalent to an anticausative construction but instead of a cause we have an agent as the element that starts the actions. Under this tag we have included both pronominal and syntactic passive constructions:

...En el peor de los casos **se construirán** o rehabilitarán en Barcelona un total de 65.000 pisos…

*...At the very least, 65.000 **will be built** or rehabilitated in Barcelona…*

If the action is neither an agentive nor a causative structure, then we use the tag passive to indicate that the logical subject of the clause is no longer the grammatical focus and that the logical object is acting as the functional subject of the clause.

...Hasta el 40 % hay familias que se lo pueden permitir, pero cuando **se supera** este porcentaje,…

*...Some families can afford up to 40%, but **past** this level...*

The last tag used to refer to the communicative focus is the *impersonal* tag. Whenever a clause does not present a functional subject, the clause is tagged as impersonal.[3]

...En este restaurante **se come** barato...

*...In this restaurant one can **eat** cheap...*

On the other hand, some properties affecting reference binding are explicitly tagged, namely *reflexivity* and *reciprocity*.

In relation to aspectuality, two specific states are distinguished: *habitual* and *middle*. The first term refers to those actions that are not truly a state, in that they do not describe a relation. However, they do not refer to a particular real-world action.

…Wimbledon siempre **cierra** sus puertas en el primer domingo del torneo...

…Wimbledon always closes its doors the first Sunday of the tournament...

Middle constructions are states that give information about how an entity's characteristic can be modified, such as "Este material se dobla con facilidad" –This material bends easily-. Examples of this type of construction appear often

[3] Here we are not making reference to typical cases of subject elision in Spanish. It is important to remember that subject elision in Spanish does not imply defocalization or its disappearance as a function.

in the literature. However, in our corpus few have appeared.

Finally, we use two more categories to account for those structures expressing an *indirect cause* and what is known in Spanish as *dative of interest*. We have an instance of indirect cause in those cases in which the syntactic agent is not the real, direct agent of the action.

..., que también **construyó** el puente sobre el Duero en Pino (Zamora) ...

*..., who also **built** the bridge over the Duero river in Pino (Zamora) and ....*

The *dative of interest* includes clauses such as the following in which the indirect pronoun is used to express a possessive relation of the speaker with the object of the clause.[4]

…se me ha detenido el motor al final…

...in the end the motor died on me…

5 Constituent-level tagging

Those constituents of the clause that are directly dependent of the verb under inspection are assigned an interpretation at various syntactic and semantic levels. We distinguish between arguments or adjuncts (Section 4.1), and the syntactic category is labeled (Section 4.2) along with the syntactic function (Section 4.3) and the expression of the semantic role (Section 4.4).

5.1 Arguments and adjuncts

Constituents are either arguments or adjuncts depending on whether they are required or not by the verb semantics. Mention should be made that for a constituent to be considered an argument it does not always have to be expressed, some arguments are optional:

Maria has eaten bread - Maria has eaten

He has arrived from Paris - He has arrived

Adjuncts usually express aspects related to circumstantial or contextual references. Typically, the aspects that can be conveyed by circumstantial constituents are the expression of place, purpose, manner, and so on. However, this is not always true the other way around. Some verbs require the expression of these types of aspects that are compulsory because of their semantics. Consider these examples:

Arguments:

He is feeling well – manner

He lives in Barcelona – place

It started at 8 – time

He uses it for writing – purpose

Adjuncts

Today, I worked pretty well – manner

I bought it in Barcelona – place

He had dinner at 8 – time

He came here to sell it – purpose

In the annotation, arguments and adjuncts are treated differently. Arguments are only annotated with reference to their semantics while adjuncts are simply tagged as such without any further analysis.

5.2 Syntactic categories

Each constituent is assigned a syntagmatic category: *prepositional phrase, relative clause,* etc. A list of the proposed categories used in the corpus annotation is presented in Table 1.

Nominal phrase
Prepositional phrase
Adverbial phrase
Negative adverbial phrase
Adjectival phrase
That-clause
Infinitive clase
Gerund clause
Indirect interrogative clause
Infinitive clause introduced by a preposition
That-clause introduced by a preposition
Adverbial clause
Relative clause
Relative pronoun
Personal pronoun (subject–yo, tú...- and object–le, me, nos...-)
Pronominal phrase (nothing, nobody, what,...)
Reported speech
Comparative phrase
Reduced clause

Table 1. List of syntactic categories.

We have created categories such as *reported speech*, *comparative phrase* and *reduced clause*. Even though these categories are not traditional syntactic categories, we have considered it necessary to create them in order to adequately solve the tagging of some segments. For example, unifying as an only constituent under the tag

[4] In English, possession is expressed directly by means of a possessive adjective.

reduced clause two separate constituents allows us to account for cases in which they are equivalent to a subordinate clause.

...Carod consideró **normal echar de menos el cargo**...

*...Carod considered **it normal to miss the post**...*

5.3 Syntactic functions

Each constituent is also assigned a syntactic function out of the list seen in Table 2.

subject	*direct object*
indirect object	*prepositional object -1*
prepositional object -2	*prepositional object -3*
attribute	*predicative*
agentive complement	*circumstantial*

Table 2. List of syntactic functions

We would like to point out that besides traditional functions such as *subject*, *predicative*, *attributive*, etc., we have distinguished three different kinds of *prepositional object* (all of which are used when annotating arguments):

- The argument is required by the verb to form a grammatical clause; even though it is not a prepositional verb, the verb does require a prepositional phrase to be syntactically realized. Sometimes more than one preposition is allowed; e.g. *ir a, hasta* – go to, go until you get to.
- The preposition dominating the argument is determined by the verb; e.g. *reírse de* –laugh at–, *acostumbrase a* –get used to–.
- The complement (a PP) is included in the subcategorization frame of the verb, but it is not required by the verb and the form of the preposition is not determined by it; e.g. the verb *correr* –run– can be used with or without complements.

5.4 Semantics

The semantic head of each argument constituent is also signaled. These head will constitute the set of words required to acquire the selection restrictions of a given verb.

To avoid interference with the information provided at this level, whenever a metaphorical or metonymical complement is observed, it is marked as not to be taken into account in this process.

...Documentos TV celebra hoy sus 800 programas...

...the show "Documentos TV" today celebrates its 800th program...

Moreover, each argument is assigned a semantic role. The inventory of semantic roles can be seen in Table 3.

initiator	*agent*
cause	*agent – cause*
experiencer	*agent – experiencer*
indirect cause	
theme	*perceived theme*
moved theme	*agent – moved theme*
affected theme	*creation affected theme*
destruction affected theme	
initial state theme	*resulting state theme*
goal	*source*
localization	*route*
purpose	*instrument*
medium	*manner*
quality	*quantity*
substitutive	*company*
time localization	*goal time*
source time	*circumstance*

Table 3. List of semantic roles

As can be seen, our inventory maintains the majority of the well-established semantic roles, such as *cause*, *agent*, *theme* and *destination*. Other tags are newer and have been created *ad hoc* in order to solve the problems that have appeared. Some of these tags are: *initiator, indirect cause, resulting state theme, initial state theme, affected theme, substitution, comparative,* and *quality.*

The role *initiator* is used to label those cases in which the promoter of the action is neither a cause nor an agent nor an experiencer, as in the case of the verb *lose*. *Indirect cause* is represented by verbs such as *formar* –form–, in which the subject may not the direct agent but rather the instigator of the action.

...el sargento **formó** a los reclutas para pasar revista...

...the sergeant mustered the recruits in order to pass review ...

The themes *resulting-state* and *initial-state* are required to annotate the complements of verbs such as *converir* –convert*:*

El mago ha convertido el pañuelo en una paloma

fThe magician has ***turned*** *the handkerchief into a dove*

When discussing the role of affected them, this label is very useful as it serves to differentiate objects whose properties are modified in order to achieve the action.

...las entidades y los feriantes han **acabado** contentos...

... the organizers and the fair show stand sponsors ***are pleased*** *with the result ...*

The term *substitution* is used to tag arguments such as *por ti* –for you– in a clause such as "He hablado por ti" –I spoke on your behalf–. C*ompany is a* role used in cases such as "Está con Luisa" –He's with Luisa–. Lastly, the role that identifies an object as part of attributive clauses is tagged *quality*.

Besides, we have further used two mechanisms to account for specific semantic relations between verbs and arguments. On the one hand, we have foreseen the possibility of double-tagging an argument using tags as *ag_exp* and *ag_t-des*. With these tags we express that an argument is both an agent and an experiencer or an agent and a moved-theme. They are used with verbs such as *read* and *run*, respectively.

We also use more generalizing tags such as *ag_caus* or *circ*. The former is used for those verbs that can be either agentive or causative (*romper* –break–). The latter expresses circumstances of the action that are diverse in nature, such as time ("The fire started at 10") and place ("The fire started in the forest").

6 Conclusions and future work

We have presented an annotation schema that aims to bring the insight of theoretical perspective to the annotation of actual corpus examples.

One of the main interests of the schema lies in the fact that the annotation process considers various, interrelated aspects of verbal behavior at the same time, and not as unconnected items. Also, the fact that various annotators deal with the same items allows for different views of the same phenomena to be put together in order to build an objective interpretation.

The guidelines presented here are flexible and they are being progressively enriched as new phenomena arise. To our knowledge, no comparable guidelines have ever been made public for Spanish.

Future work includes dealing with problematic cases. So far we foresee several areas that require further study. With regard to syntactic functions, we should review the list of functions since we have observed that there is some duplicity with respect to some categories. For example, the function called *prepositional object* actually indicates that the phrase is a prepositional phrase and not really a function.

Also, there are some phrases that perform the same syntactic function and are differentiated in traditional grammar. This is the case of the PP of verbs such as *acabar* –finish, end– in a clause such as "acabó en la piscina" –he ended up in the swimming pool– (with the syntactic function prepositional object) and the AP of the same verb in another clause such as "acabó mojado" –he ended up all wet– (with the syntactic function predicative).

Finally, another function we need to review is that of *circumstantial*. In the majority of cases it implies duplicates the information provided by the *adjunct* tag even though this is not always the case.

With regard to semantic roles, some of the tags need to be reviewed since we have detected that we are using some of them as sets, as sometimes happens with the tags *theme* and *initiator*. Moreover, the semantic role known as goal includes different types such as beneficiary, etc.).

References

X. Carreras, I. Chao, L. Padró and M. Padró. 2004. FreeLing: An Open-Source Suite of Language Analyzers. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal.

M. Collins. 1996. A New Statistical Parser Based on Bigram Lexical Dependencies. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 184-191.

B. Comrie. 1976. *Aspect*. Cambridge University Press, Cambridge, UK.

CoNLL. 2004. http://cnts.uia.ac.be/conll2004/

CoNLL. 2005. http://cnts.uia.ac.be/conll/

J. Eisner .1996. Three New Probabilistic Models for Dependency Parsing: An Exploration. *Proceedings of COLING-96*, pages 340-345.

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

A. Goldberg. 1995. *Constructions: a construction grammar approach to argument structure*. Chicago: University of Chicago Press.

D. Lin. 1998. Dependency-based evaluation of MINIPAR, *Proceedings of the Workshop on The Evaluation of Parsing Systems at LREC'98*, Granada, Spain.

A. Moreno, R. Grishman, S. López, F. Sánchez and S. Sekine. 2000. A Treebank of Spanish and its Application to Parsing. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*. pp. 107 – 112. Athens, Greece.

J. Pustejovsky. 1995. *Generative Lexicon.* Cambridge University Press, Cambridge, UK.

Senseval. http://www.senseval.org/

C. Subirats and H. Sato. 2004. Spanish FrameNet and FrameSQL. *Proceedings of 4th International Conference on Language Resources and Evaluation, Workshop on Building Lexical Resources from Semantically Annotated Corpora*, Lisbon.

G. Vázquez, A. Fernández and M. A. Martí. 2000. *Clasificación verbal. Alternancias de diátesis*, Universitat de Lleida.

Z. Vendler, 1957. Verbs and Times. *Philosophical Review* 56, 143-160.