

Description of the guidelines for the syntactico-semantic annotation of a corpus in Spanish

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

*Dept. of English and Linguistics, Universitat de Lleida, Spain

gvazquez@dal.udl.es

**Dept. of English and German Philology, U. Autònoma de Barcelona, Spain

ana.fernandez@uab.es

† Department of Linguistics, Universitat de Barcelona, Spain

lalonso@ub.edu

Abstract

The aim of the SenSem project¹ 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.

1 The SenSem project

The final aim of the SenSem project is to create a databank (a lexicon linked to manually analyzed corpus examples) that reflects the syntactic and semantic behavior of the verbs selected. As the initial phase of the project, a reference corpus for Spanish annotated with syntactico-semantic information is being constructed.

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. The final goal of the guidelines used is to bring the theoretical insights to the annotation of the actual examples found in corpus and to provide annotators with procedures as objective as possible to deal with phenomena found in corpus.

Each sentence is linked to the verb sense it exemplifies. Each verb sense is in turn associated with its Aktionsart and its argument structure in the form of a semantic role list. So sentences inherit this information from the sense.

Participants in the sentence are annotated with respect to their syntactic function, the semantic role they hold with respect to the verb and their argument or adjunct status, along the lines of the Spanish FrameNet (Subirats and Sato 2004) and ADESE (García de Miguel and Comesaña 2004). The argument head is marked, together with any metaphorical usages.

SenSem differs from other projects that treat syntactico-semantic annotation in that sentences are also tagged with respect to their semantics, both aspectual and construction.

Annotators have been trained and provided with an annotation manual. Once the sentences have been annotated, the annotation methodology requires they be validated by other linguists in order to detect possible errors and provide a more uniform treatment of problematic cases (Alonso et al. 2005).

As a result of this project, a corpus of approximately 1,000,000 words will be created, containing 100 sentences for each of the 250 most frequent verbs of Spanish. These sentences have been randomly selected from a corpus of approximately 13,000,000 words of the electronic versions of different newspapers. The journalistic register provides a high number of examples and reflects standard language usage, but a future development of this project will take into account the need to diversify the corpus. Also, we want to apply mechanisms that automatically increase the number of sentences per verb.

2 Sentence-level tagging

Two kinds of sentential semantics have been distinguished: one which concerns the aspectual information expressed in the sentence (Section 2.1), and another which specifies the semantics of its construction (Section 2.2).

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

2.1 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 take place over 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...*

- State denote 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 sentences 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 verbal phrase 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*).

Verb tense can also change the aspect of a sentence. Nevertheless, we do not consider tense as an element to take into account when analyzing the aspectuality of the sentence 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 2.2).

2.2 Construction semantics

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 installment of the PSOE political party in government...*

Secondly, we also include passive constructions and we account for both pronominal and syntactic passive constructions. They have been 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:

...En el peor de los casos **se construirán** o rehabilitarán en Barcelona un total de 65.000 pisos...
...*At the very least, 65.000 apartments will be built or rehabilitated in Barcelona...*

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

...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 sentence does not present a functional subject, the sentence is tagged as impersonal.²

...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*.

Piloto y copiloto **se cambiaron** el sitio...

*The pilot and the copilot **exchanged** places....*

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.³

Finally, we use two more categories to account for those structures expressing an *indirect cause* and *dative of interest*. We have an instance of indirect cause

² 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.

³ We have not found any constructions of this type in the corpus so we use an invented example here.

in those cases in which the syntactic agent is not the real, direct agent of the action.

..., el Gobierno también **construyó** el puente sobre el Duero...

..., *the Government also **built** the bridge over the Duero river....*

The dative of interest includes sentences such as the following in which the indirect pronoun is used to express a possessive relation of the speaker with the object of the sentence.

...se me ha detenido el motor...

... *the motor died on me...*

3 Constituent-level tagging

Those constituents of the sentence that are directly dependent on the verb are assigned an interpretation at various syntactic and semantic levels. First, we determine whether a constituent is an argument or an adjunct (Section 3.1). Arguments are further labeled with respect to syntactic category (Section 3.2), syntactic function (Section 3.3) and semantic role (Section 3.4).

3.1 Arguments and adjuncts

Constituents are either arguments or adjuncts depending on whether they are required or not by the verb semantics. 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 contextual references. Typically, the aspects that can be conveyed by such 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 8:00 AM* – time
He uses it *for writing* – purpose

Adjuncts

Today, I worked *pretty well* – manner
I bought it *in Barcelona* – place
He had dinner *at 8:00 PM* – time
He came here *to sell it* – purpose

In the annotation, arguments and adjuncts are treated differently. Adjuncts are simply tagged as such without any further analysis.

3.2 Syntactic categories

Each constituent is assigned a syntagmatic category: *prepositional phrase, relative clause, etc.*

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.

As a category, reported speech is very useful for labelling such complements, which are common in journalistic discourse.

... aunque sólo se han alcanzado récords en Lleida, destaca Antoni Gázquez.
... although records have only been reached in Lleida, highlights Antoni Gázquez.

In the ‘comparative phrase’ label, we join together the two elements of a comparison into a single tag.

Esta cuestión afecta más a mi padre que a mi madre.
This matter affects my father more than my mother.

As for the tag ‘reduced clause’, we unify as an only constituent two separate constituents. Consider the example:

...Carod consideró normal echar de menos el cargo...
...*Carod considered it normal to miss his position*...

Considerar has two complements, an adjective and an infinitive clause that can be converted into a single completive clause: *Carod considered that it was normal to miss his position*. With the aim of standardizing the treatment of both types of construction, we label the two complements of the former as a reduced clause.

3.3 Syntactic functions

Each constituent is also assigned a syntactic function. In addition to traditional functions such as *subject, predicative, attributive, etc.*, we have distinguished

three different kinds of *prepositional object* –PO– (all of which are used when annotating arguments):

- PO-1: The argument is required by the verb to form a grammatical sentence; 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).
- PO-2: The preposition dominating the argument is determined by the verb; e.g. *acostumbrarse a* – get used to–, *reírse de* –laugh at.
- PO-3: The complement is included in the sub-categorization frame of the verb, but it is not necessarily compulsory as it is in the case of PO-1; e.g. the verb *correr* –run– can be used with or without complements and it accepts prepositions such as *a* –to– or *hasta* –until (you get to).

3.4 Semantic roles

Each argument is assigned a semantic role. 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 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*⁴ is represented by verbs such as *formar* –muster–, in which the syntactic subject may not be the direct agent but rather the instigator of the action; in fact, the true agent is the object.

... 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 *convertir* –convert:

El mago **ha convertido** el pañuelo en una paloma.

⁴ We have distinguished between an indirect cause at the constituent level and another at the sentence level.

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

The role *affected them* 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 sentence such as “He hablado por ti” –I spoke on your behalf–. *Company* 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 sentences is tagged *quality*.

*...en el que Capella **ha actuado** como detective.
... *in which Capella **has acted** as a detective.**

Besides, we have further used two mechanisms to account for specific semantic relations between verbs and arguments. We have foreseen the possibility of double-tagging an argument using tags as *ag_exp* and *ag_t-des* when we want to express that an argument is both an agent and an experiencer or an agent and a moved-theme.

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 which are diverse in nature, such as time (“The fire started at 10”) and place (“The fire started in the forest”).

The semantic head of each argument constituent is also signaled. These heads 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,

We have presented a description of annotation guidelines designed to bring a theoretical perspective to the annotation of actual corpus examples. To our knowledge, no comparable guidelines have ever been made public for Spanish.

The guidelines described are flexible and they are being progressively enriched as new phenomena arise.

References

- (Alonso et al. 2005) L. Alonso, J.A. Capilla, I. Castellón, A. Fernández, G. Vázquez. The SenSem Project: syntactico-semantic annotation of sentences in Spanish. RANLP 2005.
- (Carreras et al. 2004) X. Carreras, I. Chao, L. Padró and M. Padró, FreeLing: An Open-Source Suite of Language Analyzers. Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal, 2004.
- (Comrie 1976) B. Comrie, Aspect. Cambridge University Press, 1976.
- CoNLL. 2005. <http://cns.uia.ac.be/conll/>
- (García de Miguel & Comesaña 2004) J. M. García 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 Cognição: Estudos de Linguística Cognitiva*. Almedina, 2004, pp. 399-420.
- (Goldberg 1995) A. Goldberg, *Constructions: a construction grammar approach to argument structure*. University of Chicago Press, 1995.
- (Pustejovsky 1995), J. Pustejovsky, *Generative Lexicon*. Cambridge University Press, 1995.
- Senseval. <http://www.senseval.org/>
- (Subirats & Sato 2004) C. Subirats and H. Sato, *Spanish FrameNet and FrameSQL. Proceedings of 4th International Conference on Language Resources and Evaluation, Workshop on Building Lexical Resources from Semantically Annotated Corpora*, Lisbon, Portugal, 24-30 May, 2004, 2004.
- (Vendler 1957) Z. Vendler, 1957, *Verbs and Times*. *Philosophical Review* 56, 1957, pp. 143-160.

4 Conclusions