Semantic categorization of Spanish se-constructions

Glòria Vázquez*, Ana Fernández Montraveta†, Irene Castellón‡, Laura Alonso‡

GRIAL

*Dept. of English Philology and Linguistics
UdL
Pça Victor Siurana, 1
Lleida 25003
gvazquez@dal.udl.es

†Dept. of English and German Philology
UAB
Emprius 2
Sabadell 08202
ana.fernandez@uab.es

‡Dept. of General Linguistics
UB
Gran Via de les Corts Catalanes, 585
Barcelona 08007
{castel,lalonso@fil.ub.es}

Abstract

In this paper we present a tool to automatically determine the semantic interpretation of some ambiguous sentences in Spanish, the so-called "*se*-constructions". These sentences are syntactically similar, but their argument structure is different. The performance of the disambiguation procedure has been evaluated against a manually annotated corpus, achieving 95% precision.

Keywords: Automatic disambiguation, syntactico-semantic interface, "*se*-constructions".

1. Introduction and Motivation

In this paper we present a system to automatically determine the semantic interpretation of Spanish "*se*-constructions". What these sentences have in is common the presence of the particle *se*. In previous research carried out on Spanish corpora, we have determined that 18% of the sentences are a *se*-construction (Fernández et al., 2003), which means that a total of 4.263 pronominal sentences out of 22.881 sentences. The corpora used for this study are *El periódico* and *La Vanguardia*, two Spanish newspaper corpora totalling about 500.000 words.

"*Se*-constructions" are highly ambiguous. They can express reflexive, reciprocal, passive, anticausative, impersonal or agentive meanings (Fernández et al., 2003), among others. To be able to disambiguate "*se*-constructions" is of interest since differences in meaning concern the thematic analysis of the sentence. Also, it might contribute to determine the sense in the case of polysemous verbs.

As regards thematic structure, a verb can present different ways of argument realization depending on the meaning to be expressed. For example, the verb *lavar* (*wash*) which presents a thematic structure that requires the existence of an agent and a patient, can participate in structures with one or two arguments. As can be seen in the examples below, in (1) a passive meaning is expressed (patient *la ropa sucia*) whereas (2) is an active sentence (agent *the priest* and patient *se*):

(1) *La ropa sucia se lava en casa*
Dirty laundry is washed at home

(2) *Por qué se lava el cura en la misa?*
Why does the priest wash in the mass?

The automatic treatment of these phenomena is even more problematic because Spanish presents a very unconstrained word order, where not only can arguments be found in virtually any position with respect to the verb, but some of them can also be omitted without the sentence being ungrammatical. Moreover, some impersonal structures do not have a subject altogether, not even an omitted one.

As we have said, the disambiguation of "*se*-constructions" can be used to determine which sense is used in a sentence. Not all senses of a verb present the same behaviour regarding "*se*-constructions". For example, only one of the senses of *presentar* can also express a reflexive meaning (3), while in the sense expressed in (4), *se* could never be interpreted reflexively:

(3) *El autor se presenta a sí mismo como un luchador*
The author presents himself as a fighter

(4) *Se presentó en la reunión inesperadamente.*
He arrived at the meeting unexpectedly.

Word Sense Disambiguation is an area of interest to all areas of NLP, while the first aspect, namely, assigning the proper thematic structure to a sentence (Kingsbury and Palmer, 2002; McCarthy, 2000), is specially interesting for some high level NLP applications, like Information Extraction or Question Answering, in which it is crucial to determine the role of the arguments in the sentence. Another area in which this information is determinant is that of Machine Translation, since the translation of each Spanish *se*-construction meaning requires different mechanisms of expression.

We have developed a tool to minimize the impact of the ambiguity of these constructions for NLP. The possible meanings of *se*-constructions are drastically reduced by a heuristic procedure and by access to a lexical database.

The paper is organized as follows. In the following section, we describe the disambiguation tool and, then, the evaluation method is explained in Section 3. Results are discussed in section 4.

2. Disambiguating *se*-constructions

As can be seen in Figure 1, the disambiguation process integrates heterogeneous kinds of linguistic knowledge in a neatly modular architecture.

First, a morpho-syntactic categorization is carried out by means of a pipeline of shallow analyzers (Atserias et al., 1998). Then, the following elements are tagged with semantic information (lexical-based tagging):

- Nominal heads are assigned to one semantic class.

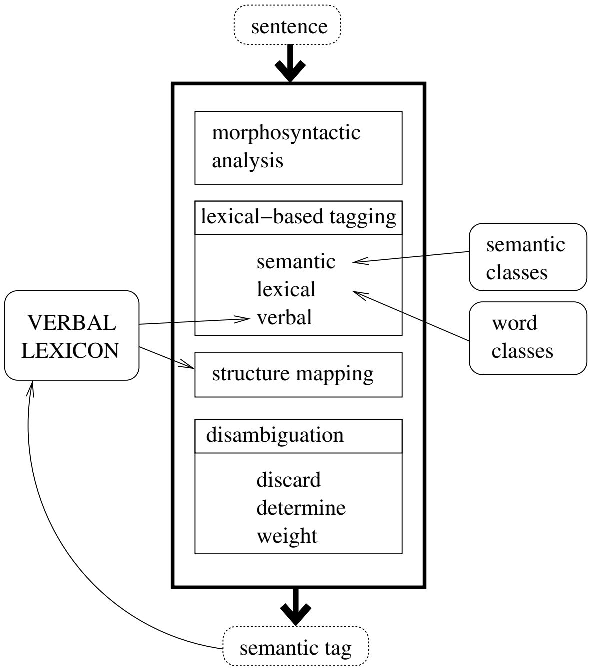

Figure 1: Architecture of the disambiguation system.

- Those specific words, useful to determine semantic patterns, are identified as key words.
- For each verbal head, a list is built with every possible sense and its associated *se*-constructions.

The list of *se*-constructions provides information about the expected argument structure of the sentence. All possible structures are mapped to the sentence under consideration (structure mapping). Once the mapping has been completed, the proper task of sentence disambiguation starts. The output of the process is the semantic tagging of the sentence under consideration.

2.1. Lexical-based tagging: resources

The main resources that are exploited by the disambiguation procedure are *semantic classes*, *words classes* and the *verbal lexicon*. Semantic classes are obtained from the Spanish version of EuroWordNet (Vossen, 1998), using the information on semantic fields and hyponymy relations. Semantic classes are used as selection restrictions for the predicates in the database (e.g. human in preverbal position). They are also used to determine the semantic class of some adjuncts (e.g. time adjuncts).

Word classes are formed up by key words and expressions (*mutuamente -each other-*, *a sí mismo -oneself-*, etc.). By key words we refer to those lexical items that have been identified as prototypical of some semantic configurations, i.e. whenever they appear we can be certain of the semantics of the sentence.

An important source of information for the process is the verbal lexicon SenSem, carried out within the projects VOLEM (Fernández et al., 2002) and SenSem (MCyT BFF2003-06456). The basic unit of description in the database is the sense and each sense is defined by several properties:

Semantic class to which the verb belongs (Vázquez et al., 2000).

Argument structure expressed in terms of thematic roles.

Syntactico-semantic frames that contain the information about the number and type of arguments and the semantics of each syntactic structure: anticausative (anti-pr-2np), impersonal (imp-se-pp), etc.

The use of this resource provides the system with the set of only those semantic tags that are possible for a sense so the range of possibilities is reduced.

2.2. Heuristics for disambiguation

In order to establish the heuristic, a 500 sentences corpus has been collected. The sources selected are diverse (RAE, Lexesp, El Periódico). The pronominal corpus is made up of sentences in which the form *se* is to be found. The rules implemented in the heuristic were settled from the study of these sentences.

There are two types of rules: restrictive and preference rules. The former permit us to eliminate or assign an interpretation. The latter help us to assign weight to possible interpretations. The process ends if one of the following possibilities is met:

- the verb only undertakes one pronominal *se*-construction;
- all possible tags, except one, are eliminated;
- the definitive tag is assigned;
- all the relevant rules are applied and no determination is reached. All the possibilities left are then ordered according to their respective weights.

The disambiguation module is made up of a total of 22 rules that are structured in three phases ordered according to the application process and the sort of action they carry out (see Figure 2).

In the first phase, we aim at eliminating as many interpretations as possible so that the list of candidates is reduced to a minimum. For example, when a sense can express an anticausative or a reflexive meaning and thus, the list created contemplates both possibilities, one of them might be ruled out depending on the semantic categorization of the arguments. For example, in example (5) the reflexive meaning can be eliminated since the preverbal NP is non-human.

(5) *..., la abarrotada Sala de la prestigiosa*
..., the crowded hall of the prestigious
Musikverein se llenó con los acordes...
Musikverein filled with the chords ...

Another type of semantic information that can be of help in this phase is the semantic class to which the verbs has been ascribed to. For example, for verbs belonging to the *change of possession* class the string of pronouns *se lo* implies an agentive interpretation (6) whereas the string *se le* favors a passive interpretation (7):

a sentence is **not reflexive or reciprocal** *if* — *elimination rule*
    `there is a preverbal noun phrase` *and*
    `this noun phrase is not human` *and*
    `this noun phrase and the verb agree`
a sentence is **impersonal** *if* — *determining rule*
    `there is no noun phrase` *or*
    `( there is only one noun phrase` *and*
    `this noun phrase is temporal )`
a sentence is **more probably anticausative** *if* — *weighting rule*
    `there is an abstract noun phrase` *and*
    `this noun phrase and the verb agree`

Figure 2: Some example rules for assigning the proper semantic interpretation to a sentence.

(6) *Pedro no se lo envió a tiempo*
Pedro didn't send it to her in time.

(7) *se le ha dado un plazo extra para terminar.*
he has been given some extra time to finish.

In the second phase we determine, whenever possible, the right semantic tag, and, if achieved, it implies the end of the process. To that aim, the rules seek for the presence of some very well defined lexical marks. For example, the presence in the context of words or expressions, such as *mutuamente* (mutually) or *entre ellos* (each other), will determine a reciprocal interpretation:

(8) *... , se decían el uno al otro.*
... , they told each other.

In the third phase, rules that award different weights to the constructions are applied. They are based on discriminative values attached to the selection restrictions. For example, properties such as the semantic types of the verb arguments are considered (see example (9)):

(9) *El problema se resolvió*
The problem was solved

The fact that "the problem" is an abstract object undergoing a process, instead of a human, confers the anticausative meaning a more pre-eminent position.

3. Evaluation methodology

The evaluation has been carried out by comparing the interpretation provided by the system with a gold standard created by three human experts. Additionally, a dummy baseline was created in order to assess the improvements on interpretation that the system contributes. This baseline assigns every sentence the *anticausative* interpretation, the most common in the manual tagging of the corpus.

The gold standard corpus for this first evaluation was collected from a press corpus. All the extracted sentences contained the *se* pronoun. From it, we selected 10 verbs that fulfil several criteria: they are described in the database SenSem, they have a high frequency of use and they present different behaviour with respect to the se-constructions. For each verb, 10 example sentences were randomly extracted, so that a total of 100 sentences have been evaluated.

	correct interpretation	incorrect interpretation	
	precision	*error*	*recall*
baseline	65%	35%	65%
system	**95%**	**5%**	**73%**
ambiguous interpretations	25%	0%	
inambiguous interpretations	70%	5%	

Table 1: Results of the evaluation of the system and a dummy baseline against a manual gold standard.

The process of manual interpretation was as follows. First, each verb occurrence was assigned a sense [1]. Then, three judges annotated the sentence with the possible interpretations. Each sentence was assigned a unique interpretation, obtained by consensus from the proposals of judges. Whenever disagreements arose, systematic decisions were taken in order to establish the consensus. For example, the most frequent disagreements between judges were between anticausative and passive interpretations, since sometimes both interpretations are so close that only contextual information, that might be beyond sentence limits, could really help disambiguate them.

4. Results and Discussion

The performance of the baseline and the system is assessed by precision and recall against the gold standard.

$$precision = \frac{correct\ interpretations}{sentences}$$
$$recall = \frac{correct\ interpretations}{interpretations}$$

In the case of the baseline, precision and recall are coincident because there is no ambiguity in the output of the baseline. In contrast, the output of the system may be ambiguous, in cases where no evidence could be found to assign a unique interpretation to the sentence, or where different interpretations could not be weighted differently. In those cases, all the interpretations provided by the system

[1] In order to make the output of the system comparable with the gold standard, sentences were manually disambiguated for verbal senses before automatic interpretation.

have been considered, which makes a difference in precision and recall.

Even if the system does not resolve the ambiguity provided by the verbal database totally, it significantly reduces it. Indeed, 58% of the sentences in the corpus are assigned more than one interpretation by the database, while the system provides ambiguous interpretations for only 26%, thus reducing 32% of the initial ambiguity.

As can be seen in Table 1 the system presents a 30% improvement in precision over the baseline, achieving almost full reliability for judgements, at the cost of only a slight improvement in recall, of 8%. Still, the recall of the system covers most of the test data (73%).

An accurate analysis of rule application was carried out, both for correct interpretations and specially for the 5% cases where the system provided erroneous interpretations. For example, it was found that the rule that applies the *th-roles* provided by the database tended to fail in discriminating passive and anticausative interpretations. Another source of error are word classes, mainly due to the fact that they are now detected by simple pattern-matching. We will improve the accuracy in the detection of these words, by additional constraints on their PoS tags.

The productivity of rules was also studied, and it was found that some rules apply seldom. For example, not once was applied the rule that eliminates the *reflexive* and *reciprocal* readings when a verb belonging to the class of *psychological change* is present, and there is a noun phrase labelled as *human* that agrees with the verb. However, we suspect the reason why some of the rules are not applied may be the small size of the evaluation corpus, so productivity of rules will be further studied in more representative evaluations.

5. Conclusions and Future Work

We have presented a first application of a system for the disambiguation of the semantic interpretation of a kind of highly ambiguous sentences in Spanish, the so-called *se*-constructions. This system is based on the verbal database SenSem, and it exploits heterogeneous lexical and syntactico-semantic knowledge.

The system has been evaluated against a manually created gold standard, showing an improvement on a dummy baseline. Even though results are not optimal, this first approach has provided us very valuable information in order to continue in the improvement and enlargement of the system and of the interactions between different kinds of knowledge involved in it.

In the near future we plan to improve the performance of the system by modifying the heuristic procedure according to the error analysis carried out here. Given the modular architecture of the tool, we will be able to investigate to what extent each linguistic level contributes to the disambiguation process. We plan to evaluate the results obtained using only some of the sources at a time, so that it becomes clear what each kind of linguistic knowledge contributes to the overall disambiguation process. Additionally, we would also like to evaluate how the process of segmenting sentences into smaller units can improve the disambiguation process, for example, by applying automated discourse segmentation (Alonso and Castellón, 2001).

The natural development of the work presented here will be to assign thematic roles to the arguments identified in the sentence, as in (Kingsbury and Palmer, 2002), according to the semantic interpretation determined by the tool. This enhancement will arguably contribute to improve the disambiguation rules based on thematic constraints.

We also intend to further enlarge the list of *se*-constructions dealt with by taking into consideration stative constructions, such as *middle* or *habitual* sentences, which will be differentiated from purely eventive ones.

6. Acknowledgements

This research has been funded by MCyT program - BFF2001-5440, and the MCyT grant PB98-1226.

7. References

Alonso, Laura and Irene Castellón, 2001. Towards a delimitation of discursive segment for natural language processing applications. In *First International Workshop on Semantics, Pragmatics and Rhetoric*. Donostia - San Sebastiàn.

Atserias, Jordi, Josep Carmona, Sergi Cervell, Lluís Màrquez, M. Antònia Martí, Lluís Padró, Roberto Placer, Horacio Rodríguez, Mariona Taulé, and Jordi Turmo, 1998. An environment for morphosyntactic processing of unrestricted spanish text. In *First International Conference on Language Resources and Evaluation (LREC'98)*. Granada, Spain.

Fernández, A., P. Saint-Dizier, G. Vázquez, F. Benamara, and M. Kamel, 2002. The VOLEM project: a framework for the construction of advanced multilingual lexicons. In *Proceedings of the Language Engineering Conference*. Hyderabad, India.

Fernández, Ana, Gloria Vázquez, and Irene Castellón, 2003. La disambiguation automática de oraciones pronominales. In *AESLA*. Santiago de Compostela. In press.

Kingsbury, Paul and Martha Palmer, 2002. From treebank to propbank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*. Las Palmas, Spain.

McCarthy, Diana, 2000. Using semantic preferences to identify verbal participation in role switching alternations. In *Proceedings of the first Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Vázquez, Gloria, Ana Fernández, and M. Antònia Martí, 2000. *Clasificación verbal. Alternancias de diátesis*. Universitat de Lleida.

Vossen, Piek (ed.), 1998. *Euro WordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers.