# Automatic Semantic Role Annotation for Spanish

**Eckhard Bick**
Institute of Language and Communication
University of Southern Denmark
eckhard.bick@mail.dk

**M. Pilar Valverde Ibáñez**
Departamento de Lengua Española
Universidade de Santiago de Compostela
pilar.valverde@usc.es

## Abstract

This paper describes and evaluates the automatic annotation of clause-level complements with semantic roles in a Spanish Web corpus, using a rule- and dependency-based approach. In all, 52 different role tags, like agent (§AG), experiencer (§EXP), location (§LOC) etc. are distinguished. The annotator uses a role grammar of 568 hand-written Constraint Grammar rules that take as input the syntactic analysis of the HISPAL parser. A rough evaluation of 5000 running words was performed, where the role annotation achieved an $F_1$ of 81,6% on raw text and 90,0% on syntactically revised input. A Spanish Internet corpus of 11.2 million words has been compiled and automatically annotated with our semantic role grammar, allowing us to provide some linguistic and statistical interpretations about the relationship between semantic roles on the one hand and syntactic functions, part of speech and semantic prototypes on the other.

## 1 Semantic roles

A semantic role is the underlying relationship that a syntactic constituent has with a predicate. Therefore, assigning semantic roles to the arguments of a verb is a way of adding deep semantic information to the analysis of a sentence. With this type of information, we can answer questions like who, when, where or what happened, which is useful in systems that require the comprehension of sentences, like dialogue systems, information retrieval, information extraction or automatic translation.

The idea of semantic roles has a long linguistic tradition, originated in the concept of case roles (Fillmore 1968), later termed thematic or theta roles in Government & Binding theory (Jackendoff 1982).

A higher level of abstraction often implies less consensus on category definitions in the linguistic community, and in semantic role annotation the level of agreement among different projects, as well as inter-annotator agreement and annotation consistency is affected by this tendency. For Spanish, the ADESSE database (García-Miguel and Albertuz 2005) uses a set of 143 roles, the AnCora corpus (Taulé et al. 2008) 20 roles and the Sensem corpus (Alonso et al. 2007) 24 roles. Only the AnCora corpus assigns a semantic role to all the complements of the clause, while the rest only treat valency-bound complements.

In our corpus, we use a set of 52 semantic roles, adopting the set of roles used by Bick (2007) for the annotation of Portuguese texts. These cover the major categories of the tectogrammatical annotation layer of the Prague Dependency Treebank (Hajicova et al. 2000), as well as those of the Spanish AnCora project.

The rules of the grammar use syntactic-semantic information available in the input (lemma, semantic prototype of the head, type of preposition, etc.) as well as information extracted from corpus-based resources, such as the ADESSE database (García-Miguel and Albertuz 2005) and the Spanish CorpusEye corpora (http://corp.hum.sdu.dk).

## 2 The grammar

We have developed a role grammar of 568 hand-written Constraint Grammar rules that exploit syntactic and semantic information to assign role tags to the clause-level complements. Input to the semantic role grammar is provided by the HISPAL parser (Bick, 2006a). Linguistically, the three main difficulties to overcome in the assignment of syntactic roles were (a) the relative lack of lexical-semantic information, (b) the fact that there is no clear correspondence between syntactic functions and semantic roles, and (c) the behaviour of the multi-ambiguous particle *se*.

### 2.1. Semantic information

We used the ADESSE database, that contains syntactic-semantic information about the clauses

**Kristiina Jokinen and Eckhard Bick (Eds.)**
**NODALIDA 2009 Conference Proceedings, pp. 215–218**

and verbs of a Spanish corpus of 1.5 million words, to study the relationship between the syntactic functions and the role of valency-governed clause-level complements. All in all, 96 sets of verb lexemes that typically allow a given role with a given syntactic function have been defined[1], moving part of the lexical information into the grammar. For example, the following LIST of verbs (V-SP-SUBJ) contains verbs whose subject is usually a speaker.

(a) LIST V-SP-SUBJ = "contar" "decir" "hablar" ...; (to tell, to say, to speak)

With the list in (a) and the following rule, the grammar assigns the role "speaker" (§ SP) to any subject (or agent complement in the passive voice) (§ARG1&) whose dependency-parent (p) is one of the verbs of the list.

(b) MAP (§SP) TARGET §ARG1& (p V-SP-SUBJ);

In addition, the semantic features of the head were also used, exploiting the semantic prototype information from the HISPAL lexicon. For example, the following rule (c) assigns the role "destination" (§DES) to a dependent of preposition (@P<) if its semantic prototype is in the set N-LOC (that contains the semantic prototypes related with a locative meaning) and its parent is in the set of prepositions PRP-DES (that contains prepositions that typically introduce this role, like *hasta (till), en dirección a (towards)*, etc.).

(c) MAP (§DES) TARGET @P< (0 N-LOC LINK p PRP-DES);

### 2.2. Diathesis alternation

The tags §ARG0& and §ARG1& are used to systematize diathesis alternation, and assigned to two types of arguments, respectively: the argument semantically closest to the predicate (0) (that corresponds to the subject in active voice) and the second closest one (1) (that corresponds to the accusative object of transitive verbs in active voice). Specifically, §ARG1& is assigned to the subject of passive clauses or unaccusative verbs and to the accusative object of the rest of verbs. §ARG0& is assigned to the rest of subjects and to the passive agent of the passive voice. The grammar takes the active voice as a reference, and the roles that would be assigned to the subject in the active voice are instead assigned to §ARG0&, and the roles that would be assigned to the accusative object in the active voice, are assigned to §ARG1&.

Three annotation principles were followed:

a) All clause-level complements (valency governed or not), are systematically assigned a semantic role, including relative pronouns and adverbial subclauses.

b) The role tags are assigned to semantic dependency heads at the token level, CG-style, i.e. alongside syntactic and other tags. However, the semantic head is not necessarily equivalent to the syntactic head. Thus, pp's were role-tagged not on the preposition, but on its dependent. In (sub)clauses, the syntactic head is the first verb of a verb chain, while the semantic head is the last one.

c) Only one role is allowed for each token, with the exception of clause-heading verbs which besides §PRED (predicator) also carries the "external" function for its clause as a whole.

### 2.3. The particle *se*

So-called "se-constructions", covering not only true reflexive use, but also others (pronominal, unaccusative, passive and impersonal), constitute one of the main sources of ambiguity in the automatic syntactic analysis of Spanish, and thus in further levels of analysis like the semantic one. These sentences are syntactically similar, but their argument structure is different.

## 3 Constraint Grammar (CG)

Our Constraint Grammar uses the new CG3 compiler, that was developed by the Danish company GrammarSoft in an open source framework, in cooperation with the VISL project at the University of Southern Denmark (for documentation, see http://beta.visl.sdu.dk/constraint_grammar.html). In fact, the semantic role annotation project served as a kind of test bed for a number of CG3 features, allowing the authors to influence compiler development according to their needs.

Like previous incarnations of the Constraint Grammar paradigm (Karlsson 1995), CG3 is basically a disambiguation and information mapping methodology operating on token-based grammatical tags that can be added, removed or changed in an incremental and context-sensitive fashion. Unlike previous editions of the formalism, however, CG3 explicitly moves beyond shallow syntax, by allowing the direct creation and use of dependency and other binary rela-

[1] The problem of semantic verb ambiguity was limited, since listing the same verb in 2 different lists was only necessary where a semantic difference corresponded to a difference in syntactic subcategorization frames.

tions. CG3 also provides for hybridization with other major parsing paradigms, integrating corpus-derived statistical information and feature-attribute unification. Finally, CG3 allows the use of regular expressions, increasing rule and tag set economy and permitting the on-the-fly reference to lexical information by reference to e.g. grammatical morphemes and affixes.

In CG3's direct use of dependency links, as we have seen in rule (c), topological methods (here searching leftward from a noun, for the nearest preposition with nothing in between but determiners) are replaced with p (parent), c (child) or s (sibling) relations. Thus, the context "p PRP LINK p V LINK c @SUBJ LINK 0 §AG" could be used to establish a preposition-link independent of the actual distance between the preposition and its argument, and to check for the agent-hood of the clause's subject (through its verb).

## 4 Evaluation

A soft evaluation has been carried out by manually revising the role labels in a fragment of 5000 running words[2]. Overall, the automatic role annotation achieved 89.0% recall, 75.4% precision and 81.6% $F_1$ (tp=1062, fp=347, fn=131)[3].

As expected, the results of an automatic role labelling system depend to a large extent on the precision of the previous syntactic analysis (Gildea and Palmer, 2002). If we only take into account the errors that can be attributed to the role grammar itself, ignoring the errors due to wrong input[4], a promising 91.4% recall, 88.6% precision and 90.0% $F_1$ are achieved (tp=1249, fp=160, fn=117).

One purpose of the evaluation was to identify error sources that could be fixed in a second development phase. For example, 20 of the false negatives (fn) are due to the fact that, by mistake, one of the rules of the role grammar included the passive clitic *se* as a target only when it was placed to the right of the verb. False positives role tags (fp) were often due to the lack of a clear-cut division between related tags. Thus, 32 errors concerned the role §BEN (benificiary) and 12 §DES (destination), both of which conflict with §REC (recipient). Those three functions constitute an important source of error not only in the automatic annotation but also in the manual annotation of Spanish corpus with semantic roles (cf. Vaamonde 2008)[5].

A relatively high pay-off can thus be expected from tackling the most problematic categories. Using the grammar for corpus creation, we intend to create a bootstrapping cycle facilitating such work, followed by more precise evaluation allowing a comparison with other role labelling systems for Spanish that are based on machine learning (e.g. Márquez et al. 2007 and Morante et al. 2007), achieving an $F_1$ of around 86%.

## 5 Corpus results

We used our semantic role grammar to create a new, annotated internet corpus of Spanish (11.2 million words), which allowed us – with a certain margin of error - to infer some tendencies about the relationship between syntactic information and semantic roles. In the table below, the most frequent correspondences are listed for some major roles (by order of role frequency), covering (a) syntactic function, (b) part of speech and (c) semantic prototype (nouns).

| Role | Syntactic function[6] | Part of speech[7] | Semantic prototype[8] |
|---|---|---|---|
| **§TH** | ACC (61%) | N (57%) | sem-c (10%) |
| **§AG** | SUBJ> (91%) | N (45%) | Hprof (7%) |
| **§ATR** | SC (75%) | N, ADJ, PCP | act (7%) |
| **§BEN** | ACC (55%) | INDP (35%) | HH (13%) |
| **§LOC-TMP** | ADVL (64%) | ADV (34%) | per (31%) |
| **§EV** | ACC (54%) | N (85%) | act (33%) |
| **§LOC** | ADVL (57%) | PRP-N (55%) | L (10%) |
| **§REC** | DAT (73%) | PERS (41%) | H (9%) |
| **§TP** | FS-ACC (34%) | VFIN (33%) | sem-c (14%) |
| **§PAT** | SUBJ> (73%) | N (55%) | sem-c (7%) |

Table 3: roles, syntax and lexical categories

[2] Due to our diferent and much larger role set, it was not possible to use pre-existing evaluation material from SemEval 2007 (the AnCora corpus)

[3] tp = number of correctly detected cases; fp = number of incorrectly detected cases; fn = number of non-detected cases. Recall = tp / (tp+fn) ; Precision = tp / (tp+fp) ; F1 = (2 * precision * recall) / (precision+recall)

[4] In the manual evaluation, errors were classified into 2 types: those attributed to a previous incorrect syntactic analysis and those attributed to the role grammar alone.

[5] Obviously, role-specific differences in performance will be of interest also beyond the evaluation phase, but figures from the current development-level grammar were not deemed to be of interest as such.

[6] SUBJ=subject, ACC=direct object, DAT=dative object, SC=subject complement, ADVL=adverbial, FS-ACC=que-subclause

[7] N=noun, PERS=personal pronoun, INDP=non-inflecting nominal pronoun, VFIN=finite verb, PRP-N=prepositional phrase (pp) with noun,

[8] H=human, Hprof=professional, HH=human group/organisation, sem-c=cognitive semantic product, f=feature, act=action, L=location,, per=period

As percentages in the first column indicate, every role can be fulfilled by multiple syntactic functions, §AG (agent) and its subcategories §SP (speaker) and §COG (cognizer) having the smallest spread (subject and passive agent). "Easiest" are roles like §SP and §COG, since they can be determinde from the head verb alone, while roles like §AG and §TH (theme) cover a wide range of verbs and semantic features. Inversely, the dominant functions, subject and object, both can correspond to around 20 different roles, depending on the target's semantic features, the governing verb, etc. However, certain tendencies can be observed, which might be of interest to descriptive linguistics, and could also be exploited in parser design. Thus, §AG has the highest and §BEN and §TP (topic) the lowest subject/object ratio.

| **Role** | Frequency | Subject/object ratio | Left/Right ratio |
|---|---|---|---|
| **§TH** | 14.6 % | 25.4 % | 31.0 % |
| **§AG** | 6.6 % | 97.2 % | 78.4 % |
| **§ATR** | 6.0 % | - | 21.7 % |
| **§BEN** | 5.0 % | 3.2 % | 59.2 % |
| **§LOC-TMP** | 4.0 % | 23.7 % | 42.6 % |
| **§EV** | 3.7 % | 43.4 % | 30.0 % |
| **§LOC** | 3.0 % | 0.0 % | 23.0 % |
| **§REC** | 1.6 % | 87.8 % | 44.7 % |
| **§TP** | 1.5 % | 4.0 % | 7.5 % |
| **§PAT** | 0.4 % | 80.0 % | 68.5 % |

Table 4: frequency, function and position ratios

Our data also permit to judge the markedness of pre- or postverbal position. Thus, as expected, typically human roles (§AG, §PAT, §BEN, §REC) often occur left of the verb, while non-human roles (§TH, §LOC, §TP) are more frequent to the right. Interestingly, the rightward tendency is less marked in temporal (§LOC-TMP) than in spatial (§LOC) constituents.

For annotation examples, we refer to the grammatical search interface we established for our internet corpus, with both concordances and statistics (http://corp.hum.sdu.dk).

## 6 Conclusion and future work

We have shown that it is possible to use a rule-based approach for the semantic-role annotation of Spanish. However, given problems like (a) the interdependence between syntactic and semantic annotation, (b) the scarcity of necessary linguistic and corpus information and (c) a certain gradual nature of role definitions, more work has to be done. Here, we expect a positive bootstrapping effect from the construction of corpora automatically annotated with semantic roles, such as our Spanish web corpus.

## References


Alonso, L.; Capilla, J.; Castellón, I.; Fernández, A. and Vázquez, G. (2007): "The Sensem Project: Syntactico-Semantic Annotation of Sentences in Spanish", *Recent Advances in Natural Language Processing IV. Selected papers from RANLP 2005. Current Issues in Linguistic Theory,* John Benjamins Publishing Co, , pp. 89--98.

Bick, E. (2006): "A Constraint Grammar-Based Parser for Spanish", *Proceedings of TIL 2006 - 4th Workshop on Information and HLT.*

Bick, E. (2007): "Automatic Semantic Role Annotation for Portuguese", *Proceedings of TIL 2007 - 5th Workshop on Information and Human Language Technology / Anais do XXVII Congresso da SBC*, pp. 1713--1716.

Fillmore, C. (1968): "The case for case", *in* E. Bach and R. Harms (ed.), *Universals in linguistic theory*, Holt, Reinehart and Winston, New York.

García-Miguel, J. and Albertuz, F. (2005): "Verbs, semantic classes and semantic roles in the ADESSE project", in K.Erk; A. Melinger and S. Schulte im Walde (ed.), *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*.

Gildea, D. and Palmer, M. (2002): "The necessity of parsing for Predicate Argument Recognition", *ACL 2002*.

Hajicova, E.; Panenova, J. and Sgall, P. (2000): *A Manual for Tectogrammatic Tagging of the Prague Dependency Treebank*, Technical report, UFAL/CKL Technical Report TR-2000-09, Charles University, Chzech Republic.

Jackendoff, R. (1972): *Semantic interpretation in Generative Grammar*, The MIT Press, Cambridge.

Karlsson et al. (1995): *Constraint Grammar - A Language-Independent System for Parsing Unrestricted Text.* Natural Language Processing, No 4. Berlin & New York: Mouton de Gruyter.

Morante, R. and den Bosch, A. V. (2007): "Memory-based semantic role labelling", *Proceedings of RANLP-2007*, pp. 388-394.

Màrquez, L.; Villarejo, L. and Martí, M. (2007): "Semeval-2007 Task 09: Multilevel semantic annotation of Catalan and Spanish", *Proceedings of SemEval 2007*, pp. 42-47.

Taulé, M.; Martí, M. and Recasens, M. (2008): "AnCora: Multilevel Annotated Corpora for Catalan and Spanish", *Proceedings of LREC 2008.*

Vaamonde, G. (2008): "Algunos problemas concretos en la anotación de papeles semánticos. Breve estudio comparativo a partir de los datos de AnCora, SenSem y ADESSE", *Procesamiento del lenguaje natural*, nº 41, pp. 233--240.



ISSN 1736-6305 Vol. 4
http://hdl.handle.net/10062/9206