A web corpus of Spanish automatically annotated with semantic roles

M. Pilar Valverde Ibáñez
Universidade de Santiago de Compostela

Eckhard Bick
Institute of Language and Communication
University of Southern Denmark

1. Semantic roles

A semantic role is the actual role a participant plays in a situation, irrespectively of its linguistic encoding. 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 1972).

While there is a certain agreement in the linguistic community about the inventory of syntactic categories and functions and their definition, this is not the case at the semantic level, where there is more skepticism about the very definition of semantic role, at least in the more traditional sense.

As the level of analysis scales, so does the abstraction of the description and the gradual nature of language becomes more evident. In the corpus annotation task, this fact results in a low level of agreement among the different projects that deal with role annotation and also in a lower level of inter-annotator agreement or consistency in each annotation. With regard to the available resources for Spanish that contain manually extracted information about semantic roles, 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. In addition, among the three projects, 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, as shown in table 1, following the set of roles used by Bick (2007) for the annotation of Portuguese texts. Those 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. In contrast to the latter, the arguments selected by the verb are not incrementally numbered (ARG0, ARG1, ARG2, ARG-M) according to their degree of proximity in relation to the verb, mainly because the combination of syntactic function and semantic role tags allows the later addition of ARG- attributes on demand, without loss of information. Only ARG-0 and ARG-1 are used internally in the grammar to treat diathesis alternation, as explained in section 2.1.2.

Role	**Definition**	**Example**
§AG	agent	*alguien* come algo
§AGcau	causative agent	*alguien* lo hizo caer
§COG	cognizer	*alguien* se acuerda de algo
§SP	speaker	*alguien* dice algo
§PAT	patient	alguien come *algo*, *algo* se cae
§DON	donor	*alguien* da algo a alguien
§REC	recipient	dar algo a *alguien*
§BEN	benefactive	ayudar a *alguien*
§EXP	experiencer	*alguien* teme algo
§TH	theme	ver *algo*, alguien piensa en *algo*
§TP	topic domain	hablar sobre *algo*
§STI	stimulus	sentir *algo*, *algo* gusta a alguien
§RES	result	producir *algo*
§MES	message	decir *algo*
§SOA	state of affairs, fact	comprobar *algo*, exagerar *algo*
§ROLE	role	trabajar como *guía*
§COM	co-argument	coincidir con *alguien*
§ATR	static attribute	alguien es *alto*
§ATR-RES	resulting attribute	alguien se pone *nervioso*
§MAT	source material	hecho de *hierro*
§POS	possessor	alguien tiene *algo*
§CONT	content	algo contiene *azúcar*
§ID	identity	él se llama *Pedro*
§LOC	location	vivir en la *ciudad*
§ORI	origin, source	venir del *campo*
§DES	destination	ir a la *playa*
§PATH	path	pasear a lo largo del *río*
§LOC-TMP	temporal location	*em 2007, hace un año*
§ORI-TMP	temporal origin	desde *enero*
§DES-TMP	temporal destination	hasta el *domingo*
§EXT-TMP	temporal extension	durante dos *semanas*
§FREQ	frequency	*de vez en cuando, 10 veces*

Table 1a. Semantic roles

Role	**Definition**	**Example**
§EXT	extension, amount	pesar 70 *kilos*
§CAU	cause	porque ..., a causa de...
§COMP	comparation	mejor que nunca
§CONC	concession	aunque ...
§COND	condition	si..., siempre que...
§EFF	effect, consequence	fueron tantos que....
§FIN	purpose, intention	para..., destinado a...
§INS	instrument	cortar con un *cuchillo*
§MNR	manner	de esta manera, lentamente
§COM-ADV	accompanier	junto con..., com algo en la mano
§META	meta adverbial	según..., tal vez, *obviamente*
§ADV	dummy adverbial	many gerund clauses: *admitiendo ...*
§MED	pronominal/unaccus.	*se* acuerda de algo, *se* ha caído
§VOC	vocative	tranquilo, *Juan*!
§FOC	focalizer	*sólo, también*
§NEG	negation	no, sin...
§EV	event, act, process	*algo* termina/comienza
§PRED	(top) predicator	(main verb)
§DENOM	denomination	(lists, headlines)
§INC	verb incorporated	darse *cuenta* de

Table 1b. Semantic roles

The principles followed in the development of the role grammar (and therefore in the annotation of the corpus) are the following:

a) All clause-level complements (valency governed or not)[1], are systematically assigned a semantic role, including relative pronouns as well as subordinate clauses and coordinate clauses that fulfill an adverbial role. A phrase-level complement can receive a role only when a given rule can be applied to clause-level as well as phrase-level complements safely. For example, the noun *sala*, in the noun phrase *la reunión de los profesores en la sala de estudio*, receives the role §LOC, because it's a noun that refers to a place and it is preceded by the preposition *en*, and this type of structure is usually §LOC irrespectively of its syntactic function.

[1] In the HISPAL parser, the main clause-level syntactic functions are the following: subject (@SUBJ), accusative object (@ACC), dative object (@DAT), prepositional object (@PIV), adverbial complement (@ADVL), subject adverbial complement (@SA), object adverbial complement (@0A), subject complement (@SC), object complement (@OC), passive agent complement (@PASS), adjunct predicative (@PRED), vocative (@VOK), top node noun phrase (@NPHR) and main verb. For more details about the symbol set, see http://beta.visl.sdu.dk/visl/es/info/symbolsopen.es.html and http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html.

b) As exemplified in the example of appendix A, the role tags (at the end of line, and preceded by §) are assigned to the semantic head of the implicit semantic dependency relation, i.e. to the noun in a noun phrase, or to the dependent of a preposition. The notational convention is that semantic role tags are assigned to tokens, Constraint Grammar-style, alongside syntactic and dependency tags.

The semantic head is not necessarily equivalent to the syntactic head, and therefore the constituent's syntactic function and role function may be on different tokens. That happens mainly in the following cases:

- Prepositional phrases, where the syntactic head is the preposition, but the semantic head is the dependent.
- The semantic head of a (sub)clause (or any clause-level function) is the verb. If there is a verbal chain, the syntactic head is the first verb, while the semantic head is the last one. In addition, in clauses introduced by *el hecho de que* (like in *El hecho de que te haya llamado no significa que quiera venir a la fiesta*) the syntactic head is the noun *hecho*, while the semantic head is the verb of the subordinate clause.

c) Only one role is allowed for each token, and thus ambiguity is not allowed. Only the main verb can receive two role tags. The main verb always receives and "internal" role, as the semantic head of the clause (§PRED), and can also receive an "external" one, when the clause as a whole plays a role (§TP, §CAU, §EFF, etc.), as in the example of appendix A.

d) The rules of the grammar use syntactico-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. One of our main sources to obtain semantic information about verbs has been the ADESSE database (García-Miguel and Albertuz 2005). As for the syntactic information, we have used the Spanish CorpusEye corpora (http://corp.hum.sdu.dk).

2. The grammar

We have developed a role grammar made up by 568 Constraint Grammar handwritten rules that exploit syntactic and semantic information to assign role tags to the clause-level complements. The input to the semantic role grammar is provided by the HISPAL parser (Bick, 2006a).

2.1. The linguistic approach

Linguistically, the three main difficulties in the assignment of semantic roles are the following. First, the relative lack of lexico-semantic information, that is essential for the correct assignment of semantic roles (section 2.1.1). Second, the fact that there is no clear correspondence between syntactic functions and semantic roles, for which some systematic diathesis alternations have been estab-

lished (section 2.1.2). And third, the behaviour of the particle *se* (section 2.1.3), that can receive in Spanish several syntactic and therefore semantic interpretations.

2.1.1. Semantic information

The ADESSE database, that contains syntactico-semantic information about the clauses and verbs of a Spanish corpus of 1.5 million words, allows us to study the relationship between the syntactic function and semantic role in valency-governed clause-level complements.

All in all, 96 sets of verb lexemes that tipically allow a given role with a given syntactic function have been defined, moving part of the lexical information into 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" "explicar" "hablar" "preguntar";

With the list in (a) and the following MAP rule, the grammar assigns the role "speaker" (§ SP) to any subject (or agent complement in the passive voice) (§ARG0&) whose dependency-parent (p) is one of the verbs of the list.

(b) MAP (§ SP) TARGET §ARG0& (p V-SP-SUBJ);

In addition to that, the semantic features of the head (whether this is an animate entity, a human, a place, a tool, etc.) can also determine the role played by the complement. In certain occasions, the semantic propotype information that is available in the HISPAL lexicon in an experimental way (Bick 2006b), has also been used: for example, 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*, *en dirección a*, etc.).

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

2.1.2. Diathesis alternation

Secondary tags (§ARG0&, §ARG1&) are used internally in the grammar to systematize the most frequent diathesis alternation constructions, allowing a more straightforward mapping between syntactic functions and semantic roles.

The tags §ARG0& and §ARG1& are assigned respectively to two types of arguments: 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 (with participle or with *se*)
- the subject of unaccusative verbs (inherently unaccusative or with *se*)
- the accusative object of the rest of verbs (including impersonal constructions with *se*).

The tag §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&.

In the following examples, the complements with the head *manifestantes*, that depend on the same verbal lemma in the three cases, receive the tag §ARG1& (and therefore the same semantic role, §BEN), by means of only one rule, despite displaying different "surface" syntactic functions (accusative object, subject and accusative object, respectively) in different syntactic constructions:

El presidente ha multado a los *manifestantes*
Los *manifestantes* han sido multados por el presidente.
Se ha multado a los *manifestantes*

2.1.3. The particle se

The particle *se* constitutes one of the main sources of ambiguity in the automatic syntactic analysis of Spanish, and at the same time its interpretation is crucial for the correct assignment of roles. In addition, the particle itself can receive several roles:

- Pronominal (§MED), with intrinsically pronominal verbs.
- "Subject" roles. The *se* in passive, impersonal or unaccusative constructions receives the role that the subject would receive in the active version of the same constructions.
- In reflexive and reciprocal constructions the *se* functions as accusative or dative object and therefore receives the corresponding role, in the same fashion as a full phrase. It should be noted that in reflexive sentences with a subject and accusative object, both functions refer to the same entity, despite receiving different roles (§AG and §PAT, respectively, for example, like in *ella se asea*).
- *se* as a form of *le* (in *se lo, se la,* etc.) functions as dative object, and therefore receives the corresponding role, in the same fashion as a full phrase.

2.2. CG3

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 designed to operate 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 relations. Together with the use of arbitrary window sizes –rather than a strict sentence window– the new structural links also support anaphora and discourse markup. In therms of parsing strategy, CG3 provides for the hybridisation with other major parsing paradigms, by integrating corpus-derived statistical information and feature-attribute unification. Finally the new formalism 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 a CG rule, a context condition (in parenthesis) contains an obligatory position marker, consisting of a number indicating relative distance in tokens. The default (positive number) is a right context, while a negative number indicates a left context. For example, rule (d) maps the role §META to constituents like *en su opinión*, *en nuestra opinión*, etc, using distance between tokens. Specifically, the rule maps the role §META to a dependent of preposition (@P<) if the dependent itself (0) corresponds to the lemma “opinión,” immediately preceded (-1) by a possessive pronoun, immediately preceded by preposition *en*.

(d) MAP (§META) TARGET @P< (0 ("opinión") LINK -1 (<poss.*>r) LINK -1 PRP-EN);

In CG3’s direct use of dependency links, as we have seen in rule (c), topological position markers 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).

To understand the workings of a CG system, it is important to note that it is not a one-pass method. First, different layers of analysis are treated progress-

ively, allowing higher levels (such as semantic role annotation) to use already established "safe" information from previous levels (e.g. part of speech disambiguation or syntactic function). Second, rules –though in sequential and/or heuristic order–, can be rerun in a second or third pass in a kind of bootstrapping iteration, first providing new context for later rules, then themselves profiting from the tagging or disambiguation performed by these same later rules. In a wider sense, we exploit a similar bootstrapping effect by tagging a corpus with an immature Constraint Grammar in order to extract, for instance, verb complementation patterns or preposition statistics, which can then be fed into the system as lexical valency information or word sets in the grammar.

3. The web corpus

An Internet corpus of 11.2 million words has been automatically annotated at the dependency-syntactic level with the HISPAL parser, and at the semantic role level with our role grammar. The evaluation of the results is described in section 4.1, and the relationship between semantic roles, syntax and lexical categories found after the annotation is commented in section 4.2.

The corpus has been compiled from the Internet, using a method described in Sharoff (2006) and also used for the Leeds Internet corpora (http://corpus.leeds.ac.uk/internet.html). The method is designed to ensure a uniform source and domain spread, as well as good general lexical coverage of the harvested data. First, from a corpus-based frequency list, we manually chose the top 500 words that were specific to Spanish (i.e. not at the same time Portuguese, English, etc.) and seeded an URL collection with Google lookups of fifty thousand random 3-tuples and five thousand 4-tuples from this word list. All in all we ended up with 127,500 unique URLs. Of these only 31% where classical .htm(l) addresses. About 12% were marked as pdf or WORD documents which were discarded for the current corpus project for technical reasons, but on the other hand might offer a source of more coherent or more carefully added text in the future.

Even after source selection, the downloaded data still contained considerable amounts of binary data and other non-textual data, such as unmarked (and hence undetected) pdf data and xml meta data that had to be filtered out. Other post-processing steps involved encoding normalisation (to ISO-latin-1) and sentence splitting. As the most sophisticated filter we wrote a program (*hisponly*) designed to separate Spanish text from text in other languages, in particular English text on Spanish web sites. The program computes language probabilities for each line, using specific character combinations and frequent trigger words, and discards lines with more foreign language than Spanish triggers. As a default, neutral lines without triggers (especially: short lines) inherit their language attribute from the preceding line. Since this heuristic method worked reasonably well, we did not resort to the safer, but much slower, method of looking all

words up in a monolingual Spanish dictionary. Note that a certain language tolerance threshold is actually desirable, because Spanish sentences often contain minor amounts of foreign material, such as names, loan words, quotes and –nowadays– English technical terms, not least from the internet domain.

After cleaning, our data was distributed across encoding types as shown in table 2 (counting only function-carrying words, not punctuation tokens, etc.):

iso-8859-1	13.072.056 words	36,6 %
utf-8	7.374.397 words	20,7 %
Windows	2.182.745 words	6,1 %
unknown	13.079.055 words	36,6 %
all	35.708.253 words	100,0 %

Table 2. Encoding types

Since character recognition is crucial to the HISPAL parsers lexical lookup, and since its native mode is ISO-latin, a higher error percentage can be assumed even for utf-8 and windows-1252, since not all characters can be successfully filtered into iso-8859-1 (e.g. dashes and certain quotes). This is especially true for the data with unknown encoding, that for this reason was kept apart in the corpus search interface. For our semantic role annotation, the safe (known-encoding) part was further cleansed by excluding lines starting in '*' (lists) or '<', or containing "words" with double CamelCase (e.g. aBcD, typical of binary data), underscores or internal colon (URLs), resulting in a subcorpus of 11.16 million words.[2]

We decided to make the annotated corpus accessible at VISL's Corpuseye site (http://corp.hum.sdu.dk). The interface is based on the CQP engine and allows us to do normal and regular expression text searches as well as "refined" menu-based searches, where it is possible to look for any combination and sequence of lemma, word category, syntactic function or semantic role. Results can be evaluated in terms of frequency statistics and cooccurrence strengths. To accommodate for the semantic role categories, the search engine as such had to be changed since no other corpora at the site previously offered this type of information.

[2] Though in principle relevant in a web corpus, we did not so far –apart from ensuring URL-uniqueness– try to treat text duplication. The phenomenon is by no means infrequent, due to extensive copying or repetition between related web sites (as well as possible "plagiarism" between unrelated websites), but it is not a trivial task to remove duplicates while at the same time provide for natural repetition of shorter text strings such as individual sentences and sub-sentence segments.

4. Results

In section 4.1 we provide some preliminary evaluation measures for the automatic annotation with semantic roles, and in 4.2 we examine the relation between semantic roles, syntax and lexical categories in the resulting annotation.

4.1. Evaluation

A soft evaluation has been carried out by manually revising the role labels in a fragment of 5,000 running words. The standard evaluation measures recall, precision and F_1 are provided. Overall, the automatic role annotation achieves a 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, a promising 91.4% recall, 88.6% precision and 90.0% F_1 are achieved (tp=1249, fp=160, fn=117).

In addition, this first preliminary evaluation allows us to detect some source of errors in the role grammar that can be fixed in a second developmental 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 only includes the passive clitic *se* as a target when this is placed to the right of the verb, and not when it is place to the left as well.

Among false positives (fp) that receive a wrong role tag, more linguistic issues arise. The revision shows that (manually) annotating a corpus with semantic roles (or other type of semantic information) is a difficult task, given the inherent gradual nature of semantic concepts. In some cases, it is not clear which semantic role a given complement should receive because there is no clear-cut division between them. For example, 32 errors correspond to the role §BEN and 12 to §DES, that conflict with §REC. 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). Therefore, an accurate definition of the most conflictive roles, or systematic criteria for unclear cases needs to be established to reach a higher level of consistency.

On the other hand, 15 errors correspond to relative pronouns (*el que, el cual*, etc), for which we need to have information about the antecedent to improve the precision of its annotation.

The new annotated corpus has shown that there is still room for improvement in our grammar, which at the same time will be facilitated by the great amount of data now available. In addition, we need to carry out a more precise

---

[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)

evaluation in order to be able to compare our results more reliably with other role labelling systems for Spanish that are based on machine learning (Márquez et al. 2007, Morante et al. 2007, Surdeanu et al. 2008), which achieve an F_1 of around 86%.

4.2. The relation between semantic roles, syntax and lexical categories

With the information available in the corpus, and keeping in mind that we base our observations on non-revised, automatically annotated data, that necessarily implies a certain error margin, we can study some tendencies about the relationship between syntactic information and semantic roles. In table 3, typical correspondencies are listed for some major roles in the order of corpus frequency, covering (a) syntactic function, (b) part of speech and (c) semantic prototype (nouns only). Arrows in the function column indicate dependency attachment direction (i.e. @<ACC has its verb to the left).

As expected, we find that every role can be fulfilled by multiple syntactic functions. The less ambiguous roles in this respect are the §AG, §SP and §COG roles, that correspond always to the subject (in the active voice) or the agent complement (in the passive voice). The §TH, which is the most frequent role, corresponds always to the subject or to the accusative object. However, while the roles §SP and §COG are primarily determined by the head verb, the roles §AG and §TH cover a wide range of verbs and semantic features. In fact, they are the default roles for the functions §ARG0& and §ARG1&, respectively. In addition, since the functions of accusative object and subject are the most frequent (clause-level and valency-governed) functions in Spanish, they are also the most ambiguous in the semantic role level. They can potentially play many roles[4], especially when they are a phrase or pronoun (i.e. not a clause): both can correspond to around 20 different roles out of the 52 roles in our list, 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 the lowest subject/object ratio[5], as table 4 shows.

Our data also permit to judge the markedness of pre- or postverbal position. Thus, typically human roles (§AG, §PAT, §BEN, §REC) typically 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. The roles §PRED, §DENOM, §MED

---

[4] Cf. appendices A and B for corpus example sentences for every role of the subject and accusative object, respectively.

[5] The ratio is here defined as subject frequency divided by the sum of subject and object frequency.

and §VOK are dummy roles, since they always correspond to the same and unique syntactic tags.

Role	Syntactic function[6]	Part of speech[7]	Semantic prototype class[8]
§TH (14.6.%)	@<ACC (51%), @SUBJ>, @<SUBJ	N (57%), PERS, INF, VFIN	<sem-c> (10%), <f>, <f-psych>
§AG (6.6%) (~§SP, §COG)	@SUBJ> (77%), @<SUBJ, @<PASS	N (45%), PROP, INDP, PERS	<Hprof> (7%), <sem-c>, <HH>, <H>
§ATR (6%)	@<SC (65%), @PRED>, @<PRED	N, ADJ, PCP	<act> (7%), <sem-c>, <H>
§BEN (5%)	@ACC> (37%), @<DAT, @DAT>	INDP (35%), PERS, PRP-N, N	<HH> (13%), <H>, <Hprof>
§LOC-TMP (4%)	@<ADVL (37%), @ADVL>, @N<	ADV (34%), PRP-N, N, VFIN, PRP-INF	<per> (31%), <temp>, <dur>
§EV (3.7%)	@<ACC (54%), @SUBJ>, @<SUBJ	N (85%), INF, VFIN	<act> (33%), <act-s>, <act-d>
§LOC (3%)	@<ADVL (39%), @N<, @ADVL>, @<SA	PRP-N (55%), PRP-PROP, ADV, VFIN	<L> (10%), <Lh>, <Lcountry>
§REC (1.6%)	@<DAT (40%), @DAT>, @<ADVL	PERS (41%), PRP-N, PRP-PROP	<H> (9%), <act>, <Hprof>, <HH>
§TP (1.5%)	@FS-<ACC (34%), @<ACC, @<ADVL, @<PIV	VFIN (33%), N, PRP-N	<sem-c> (14%), <f>, <act>, <sem-s>, <ac>
§PAT (0.4%)	@SUBJ> (65%), @<ACC, @<SUBJ	N (55%), INDP, PROP	<sem-c> (7%), <f>, <H>, <HH>

Table 3. Relation between semantic roles, syntax and lexical categories

6 SUBJ=subject, ACC=direct object, PASS=passive agent, DAT=dative, ADVL=adverbial, PIV=prepositional object, N<=postnominal, FS-ACC=finite subclause (que/interrogative clause).

7 N=noun, PROP=proper noun, PERS=personal pronoun, INDP=independent pronoun, VFIN=finite verb, INF=infinitive, PRP-N=prepositional phrase (pp) with noun, PRP-PROP=pp with proper noun.

8 <H>=human, <Hprof>=professional, <HH>=human group/organisation, <sem-c>= cognitive semantic product, <f>=feature, <f-psych>=psychological feature/emotion, <act>= action, <act-s>=speech act, <act-d>="do"-act, <L>=location, <Lh>=human/functional place, <Lcountry>=country, <per>=period, <temp>=point in time, <dur>=duration, <ac>=abstract countable. Due to unresolved polysemy or subspecifications, more than one tag may occur with each noun.

Role	Frequency	Subject/object ratio	L/R ratio (Left or right of verb)
§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, subject/object ratio and left to right ratio

Finally, with regard to the relationship between semantic roles and syntactic categories, the following tendencies stand out: the roles §EFF and §COND are always played by a clause, while the roles §DES, §ORI, §ID, §COM-ADV, §COM, §ROLE and §MAT are always introduced by a preposition.

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

5. Conclusions and future work

Automatically assigning role tags to clause-level complements is a difficult task. First, it relies on a previous syntactic analysis, that in the case of automatic analysis systems imply a certain error margin. Second, the mapping between syntactic functions and semantic roles needs a kind of linguistic information that is not widely available from current corpora, or is difficult to use because of the disagreement on the inventory of roles in the different resources. And third, semantic annotation (of roles or other type of semantic information), automatically or manually driven, is a difficult task because of the inherent gradual nature of semantic concepts, that conflicts with the discrete categories used in corpus annotation.

However, as our system is rule-based, we expect a positive bootstrapping effect from the construction of corpora automatically annotated with semantic roles, such as our Spanish web corpus. As opposed to manually annotated data, there are no volume constraints, and the sheer size of such corpora will allow the statistical identification of patterns and structures that in turn can be used to improve the valency and semantic class lexica underlying both the syntactic and

role parsers, improving the scope and performance of the rules at each new iteration of the bootstrapping process described. Future work will also have to explore the possibility of integrating our semantic role annotator in an applicative context such as Question Answering or Information Extraction.

Appendix A. Tagging sample

"<Los>"
"el" <fr:83> <f:1835430> <artd> <exdem> DET M P @>N #1->2
"<fabricantes>"
"fabricante" <cjt-head> <fr:0> <f:17> <'producer'> <HHorg> N M/F P @SUBJ> §ARG0& **§AG** #2->7
"<y>"
"y" <fr:100> <f:1904300> <Rare> <co-subj> KC @CO #3->2
"<productores>"
"productor" <cjt> <fr:98> <f:9998> <'producer'> <HHorg> N M P @SUBJ> §ARG0& **§AG** #4->2
"<tienen>"
"tener" <fr:100> <f:68529> <se:12> <vq> <x+PCP> <xque> <vdb> <vr> <vt> <'have'> <mv?> <aux> <mv> V PR 3P IND VFIN @FS-STA #5->0
"<que>"
"que" <clb> <fr:38> <f:1720876> <komp> <corr> KS @PRT-AUX< #6->5
"<probar>"
"probar" <fr:100> <f:1881> <se:45> <a^xp> <vdb> <vr> <vq> <vt> <'prove'> <mv?> <mv> V INF @ICL-AUX< **§PRED** #7->5
"<que>"
"que" <clb> <clb-fs> <fr:38> <f:1720876> <komp> <corr> KS @SUB #8->11
"<sus>"
"su" <fr:100> <f:335709> <poss 3S/P> <si> DET M/F P @>N #9->10
"<productos>"
"producto" <fr:84> <f:39275> <'product'> <ac-cat> N M P @SUBJ> §ARG0& **§TH** #10->11
"<son>"
"ser" <fr:97> <f:260305> <x+PCP> <vK> <vk+ADJ> <vk+N> <'be'> <mv?> <mv> V PR 3P IND VFIN @FS-<ACC **§PRED §TP** #11->7
"<seguros>"
"seguro" <fr:100> <f:6442> <'sure'> ADJ M P @<SC **§ATR** #12->11
"<$.>"
"$." #13->0

```
"<$¶>"
      "$¶" #1->0
</s>
```

Appendix B. Semantic roles of the subject (in alphabetical order)
§AG: Con ellos el **usuario** puede buscar y visualizar fragmentos encontrados y descargarlos o enviarlos por e-mail usando el navegador web habitual y sin necesidad de estar en casa o en la oficina.

§AGCau: Y la **evolución** hace aflorar eso que está implícito en las entrañas de las cosas.

§BEN: Si el PP se refunda **UPN** se va a llevar una gran sorpresa.

§COG: El **presidente** de EA de Gipuzkoa Iñaki Galdós consideró ayer que el pacto con el PNV y Ezker_Batua-Aralar resulta <<insuficiente para avanzar en un proceso de normalización política e incluso para un escenario de gobernabilidad y se mostró partidario de implicar al PSE.

§COMP: Madeleine Slade quien estuvo durante 23 años al lado de el Mahatma indiao hasta que éste murió asesinado en 1948 está caracterizada en el cine por la nueva estrella Geraldine James que es mucho más guapa que **yo** según reconoció la interesada cuando le fueron presentadas las primeras fotografías poco antes de su muerte.

§CONT: Estos **criterios** sobre apertura se contenían también en el informe Cashman que aprobamos con amplia mayoría hace dos semanas en Estrasburgo.

§DON: En 1963 la **CIA** le envió de regalo un traje de buzo impregnado con bacilo de tuberculosis, intentó plantar un caracol explosivo en un arrecife donde Castro hacía pesca submarina y hasta probó con un bolígrafo envenenado.

§EV: Las **investigaciones** de el Hubble de enero de 1997 han mostrado interacciones interesantes entre las jóvenes estrellas de el cúmulo de el Trapecio con los discos protoplanetarios.

§EXP: A su juicio los doce **consejeros** de el Ejecutivo foral de UPN-CDN se han sentido también atacados en su propia dignidad cuando han visto que al presidente de su comunidad se le insultaba de manera personal y se ponía en duda su dignidad personal como la dignidad del cargo.

§EXT-TMP: Ya sólo quedan 13 días de Bush y **13 días** pasan rápido, afirmó Sebastián en unas declaraciones.

§INC: La misma **pregunta** se podría hacer en otro aspecto si, dice John Edwards o „Tubby Banerjee“, dos brillantes jóvenes ejecutivos con buenos grados, y que conocen bien el negocio, hubieran sido promovidos al tablero.

§MES: Asimismo habilitará un sistema eficaz para recibir y responder las **consultas,** los **reclamos** y las **denuncias** de los beneficiarios de las instituciones educativas y de las empresas de transporte.

§NEG: El presidente del Ejecutivo foral hizo hincapié en que su formación no quiere romper el pacto con el PP pero advirtió de que tampoco quiere que **nadie** pueda coartar nuestra libertad para adoptar nuestras propias decisiones.

§PAT: **Cinco** de nuestros compañeros fueron detenidos en la 26 comisaria de Pudahuel luego de constatar lesiones fueron vejados maltratados y torturados.

§POS: **Fidel** tiene un raro olfato primero para conocer a las personas y segundo para presentir las emboscadas.

§REC: Aparentemente todos estos descubrimientos se perdieron por algún tiempo de modo que fue Christian_Huygens **quien** recibió principalmente el crédito de su redescubrimiento independiente en 1656 por ejemplo por Charles Messier cuando la agregó a su catálogo el 4 de marzo de 1679.

§RES: El ex espía cree que aquella **cápsula** de veneno fue fabricada por la CIA como parte de un plan tramado en 1960 con la mafia de Chicago que fue expuesto la semana pasada cuando la agencia desclasificó más de 700 páginas de antiguos documentos.

§SOA: Abundan problemas como la basura, la **destrucción** de los manglares y la contaminación por agroquímicos.

§SP: En concreto la **coalición** de izquierdas ha señalado que no formará parte del futuro ejecutivo de la Diputación si no se paraliza la construcción de la futura incineradora de residuos sólidos urbanos.

§STI: **James Dean** gustó mucho entonces y sigue haciéndolo en la actualidad en todo el mundo.

§TH: Precisamente la **bandera** de la defensa de estos proyectos constituye un elemento clave en la negociación de un programa de gobierno foral.

§TP: No hay financiaciones indiscriminadas, sino que se comprueba la **calidad** de las acciones, se insiste en ella y las propias acciones deben ser objeto de un control y seguimiento y sus resultados han de hacerse públicos.

Appendix C. Semantic roles of the accusative object

§BEN: La posibilidad de una negociación entre EA y el PSE en Gipuzkoa podría abrir un debate en el primer partido entre quienes defienden la autonomía de la ejecutiva que preside Galdós para trazar su política de alianzas y quienes consideran que afecta a la dirección nacional porque temen que un pacto de esta naturaleza desestabilice el **Gobierno Vasco**.

§CONT: Un hombre devuelve un bolso extraviado que contenía un **cheque** de 30.000 euros.

§DES: Las Asambleas Continentales de Europa, África y América nos han señalado **caminos** nuevos, senderos inexplorados, temas apenas balbuceados, incertidumbres e interrogantes propios de nuestro tiempo.

§EV: El candidato de la unidad regional exhortó a los partidos nacionales que eviten las **imposiciones** y **maniobras** desde Caracas que tanto daño le han causado a la unidad en las regiones.

§EXP: La que más **me** fascina es precisamente aquella de la que menos sé.

§EXT-TMP: Tras nuestra negativa de retirar la instrumentaria mientras ellos los pacos no explicaran la razón de la parcial suspensión de la actividad a los vecinos que esperaban dicho acto los pacos procedieron a llamar refuerzos que demoraron **menos** de 5 minutos en presentarse en el lugar.

§EXT: La encuestadora Zogby le otorga un 50 por **ciento** contra un 43 por ciento de su rival republicano.

§INC: Señor presidente, la comunicación de la Comisión constituye, creo, un verdadero hito en la historia de la política social, porque pone fin a un período en el que la **mención** a la tercera edad se hacía desde la consideración de estas personas como la generación de los retirados.

§LOC-TMP: La idea lleva ya un **tiempo** en su cabeza.

§NEG: Habló durante 4,51 minutos hasta que el presidente del tribunal le cortó porque lo que estaba diciendo **nada** tenía que ver con el juicio.

§ORI-TMP: Vinieron amigos que habían estado con él hace cinco **años**, otro que estaba en Asturias.

§PAT: El programa puede eliminar todos los **rastros** de aplicaciones monitorizadas o borrar igualmente aplicaciones no monitorizadas.

§PATH: Miembros del aparato militar cruzaron hace tres semanas la **frontera** para entregar los explosivos al comando.

§REC: Para apoyar la creatividad y el avance de los conocimientos hace falta identificar las nuevas generaciones de investigadores dotándo**las** de los medios de independencia necesarios para que desarrollen sus propias ideas, estima Bertil Anderson, director ejecutivo de la Fundación Europea para la ciencia FEC y miembro del Comité Nobel.

§RES: Galdós subrayó en una conferencia de prensa en San Sebastián las claves políticas que entiende que se derivan de las últimas elecciones una vez que en Gipuzkoa el socialista Buen ha mostrado su preferencia por formar un **gobierno** de coalición entre el PSE y EA en la Diputación foral.

§SOA: Pero también sé, porque a mi me suele ocurrir, que a un presidente que tiene la cabeza en una cuestión que reclama toda la **atención** social y mediática de el país no le vayas con un tema regional.

§STI: El engorde con melamina podría ser frecuente en otros sectores, lo que ha hecho a las autoridades ordenar el análisis de los piensos para animales pero no habría causado entre la población fallos renales masivos como los provocados en bebés y mascotas estadounidenses gracias a que los adultos gozan una **dieta** más variada.

§TH: Galdós consideró que para avanzar en un proceso de normalización política en Euskal Herria hay que tener en cuenta al Partido Socialista y citó como referencia a tener en cuenta el **diálogo** entre el PSN y Nafarroa Bai.

§TP: Amigos soy Adulfo Pistarino el Moderador del Blog en el cual sus eminencias Pablo Fattini y Martin Von Pazzz discutirán variados **temas** sin tal vez llegar a una conclusión ya que estos prohombres tienen cada uno una manera de ver la vida [...].

References

Alonso, L., J. Capilla, I. Castellón, A. Fernández, and G. Vázquez, G. (2007): "The Sensem Project: syntactico-semantic annotation of sentences in Spanish," in *Recent Advances in Natural Language Processing IV. Selected papers from RANLP 2005. Current Issues in Linguistic Theory,* John Benjamins, 89-98.

Bick, E. (2006a): "A constraint grammar-based parser for Spanish," in *Proceedings of TIL 2006 - 4th Workshop on Information and Human Language Technology*.

Bick, E. (2006b): "Noun sense tagging: semantic prototype annotation of a Portuguese treebank," in J. Hajic and J. Nivre (eds.), *Proceedings of TLT 2006*, 127-138.

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

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

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

Gildea, D. and M. Palmer (2002): "The necessity of parsing for predicate argument recognition," in *ACL 2002*.

Hajicova, E., J. Panenova and P. Sgall (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*, Cambridge, MIT Press.

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 A. V. den Bosch (2007): "Memory-based semantic role labelling," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2007*), 388-394.

Màrquez, L., L. Villarejo and M. Martí (2007): "Semeval-2007 Task 09: Multilevel semantic annotation of Catalan and Spanish," in *SemEval 2007. Proceddings of the 4th International Workshop on Semantic Evaluations*, 42-47.

Sharoff, S. (2006): "Open-source corpora: using the net to fish for linguistic data," in *International Journal of Corpus Linguistics* 11, 4, 435-462.

Surdeanu, M., R. Morante and L. Màrquez (2008): “Analysis of joint inference strategies for the semantic role labelling of Spanish and Catalan,” in *Lecture Notes in Computer Science*, 4919, 206-218.

Taulé, M., M. Martí and M. Recasens (2008): “AnCora: multilevel annotated corpora for Catalan and Spanish,” in *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*, 41, 233-240.