Semantic role annotation: From verb-specific roles to generalized semantic roles

José M. García-Miguel
University of Vigo
Vigo, Spain
gallego@uvigo.es

Gael Vaamonde
University of Vigo
Vigo, Spain
gaelv@uvigo.es

Abstract

This paper aims to present the semantic role annotation carried out on the ADESSE project, an online database with syntactic and semantic information for all the verbs and clauses in a corpus of Spanish. In ADESSE, several subsets of semantic roles have been taken into account, interrelated through different levels of generalization.

1 Introduction

To have at our disposal annotated corpus is an obvious necessity for descriptive or computational purposes. Nevertheless, in carrying out any annotation process, we are required to move between two poles: the consistency of the data and the granularity of the analysis. Undoubtedly, this divergence increases when we have to deal with semantics, and in particular, with semantic role annotation. A factor which plays an important a role on this discrepancy tend to be the procedure adopted: automatic versus manual. The first one ensures a more systematic but coarse-grained product (Gildea & Jurafsky, 2002); the second one allows more accuracy, but it must face greater complexities. From a different point of view, the users of a linguistic resource may need sometimes very broad categories ranging over a wide set of data, and others may more detailed distinctions. Like in other annotation task, also in semantic role annotation the starting point, the design and the intended users determine to a great extent the resulting product (Ellsworth et al. 2002). Nevertheless, there are also some attempts to define a standard based on some existing alternative approaches (cf. Petukhova & Bunt 2008). Some well-known projects of semantic role annotation haven taken different paths in their design: FrameNet (Fillmore et al. 2003) is designed as an ontology of situation types (frames) and participants in those situations (frame elements) [1]. PropBank (Palmer et al. 2005) has a verb-dependent model of description of semantic relations. In this project, arguments are numbered and defined depending on the valency potential of each particular verb sense. VerbNet (Kipper, 2006) approach to meaning is based in an extension of Levin(1993)'s verb classes.

Regarding Spanish language, the Spanish FrameNet[2] project (Subirats 2009) follows exactly the same methodology that the original. But other important resources and projects of semantic role annotation of Spanish corpora use a predefined set a semantic role labels irrespective of situation type. This is the case of AnCora (Martí *et al.*, 2007, Taulé *et al.*, 2008) , and SenSem (Castellón *et al.*, 2006).

In ADESSE, a linguistic resource for Spanish, an intermediary path has been taken trying to combine the specifics of verb-senses, like in PropBank, with some generalizations over process types or verb classes. Fine-grained annotation is achieved by appealing to different subsets of semantic roles, which arise as a result of different levels of generalization. The main design features of ADESSE have been described elsewhere (García-Miguel & Albertuz 2005, García-Miguel et al. 2010) and are briefly summarized in section 2. This paper aims to show a slightly more detailed description of the levels of semantic role annotation in ADESSE, and this is the purpose of section 3.

2 The ADESSE project

ADESSE (*Base de datos de Verbos, **A**lternancias de **D**iátesis y **E**squemas **S**intáctico-*

[1] http://framenet.icsi.berkeley.edu
[2] http://gemini.uab.es:9080/SFNsite/

S**emánticos del **E**spañol*)[3], a project being developed at the University of Vigo, is an online database providing detailed syntactic and semantic information about verbs and clauses from a Spanish corpus. ADESSE is an expanded version of BDS (B**ase de **D**atos **S**intácticos del español actual*), the syntactic analysis of a corpus of Spanish into a relational database. ADESSE takes a syntactically analyzed corpus to semantically annotate all and only the clauses in the corpus. In this respect, ADESSE is partly similar to a Treebank with syntactic and semantic annotation, although limited to argument structure. The manually annotated corpus has 1.5 million words, 159,000 clauses and 3,450 different verb lemmas. BDS contains grammatical features of verbs such as voice, tense and mood, and syntactic features of verb-arguments in the corpus, such as syntactic function, and phrase type. ADESSE has added semantic features such as verb sense, verb class and semantic role of arguments to make possible a detailed syntactic and semantic corpus-based characterization of verb valency. A fundamental goal of the project is to get a corpus-based description of verb valency in Spanish. The database includes, among other things, the syntactic function and the syntactic category for each core argument of each clause in the corpus, and semantic information about verb sense, semantic verb class for each verb sense, and semantic roles for each verb argument.

3 Semantic role annotation in ADESSE

Semantic annotation in ADESSE was primarily carried out for descriptive purposes, and follows always a bottom-up approach, starting from the data a trying to define a set of categories that can describe those data. This can explain why the cited project adopts a fine-grained annotation of semantic roles, compared with other similar resources for Spanish, like AnCora or SenSem. Unlike these projects, there is no just one set of roles for annotating arguments in ADESSE. Actually, we do not use any previous list of possible options. The strategy is an inductive one, taking verb meaning as the starting point and describing (types of) participants from each verb sense in an increasingly wide-ranging way. This strategy allows us to cover different levels of granularity and, at the same time, to establish generalizations about argument structure based on lexical verb meaning.

Taking all of this into account, role definition is made at three levels in ADESSE: verb-specific roles, class-specific roles, and generalized semantic roles.

3.1 Verb-specific roles

Verbs categorize types of situations and participants in those situations in a unique way, so at the extreme a distinct set of participant roles must be posited for each verb sense (cf. Langacker, 1991:284). Role definition in ADESSE is initially carried out on this maximally specific level. For each verb sense, we describe its valency potential, that is, the whole set of possible participants accepted with that verb, taking into account all the syntactic patterns recorded in the corpus (its valency realizations). The goal here is, on the one hand, to distinguish roles of participants co-occurring in the same syntactic pattern and, on the other, to trace equivalences between arguments of different syntactic patterns

For example, the verb *contar* ‘to tell a happening’ can be described by considering up to four arguments: A1: ‘the one who tells something’, A2: ‘the thing told’, A3: ‘the one to whom something is told’, and A4: ‘the issue of what is told’. This allows us to describe examples like (1a), where the whole range of participants is expressed in a single clause, as well as (1b) or (1c), where only a subset of them is selected. (In these examples 1-2-3-4 stand for A1-A2-A3-A4)[4]:

(1) a. [1] *Cuénta*[*nos* 3] [*algo* 2] [*de Madrid* 4]
‘Tell [_1] [us 3] [something 2] [about Madrid 4]’
b. [*El viejo* 1] *cuenta* [*su última treta* 2]
‘[The old man 1] tells [his last ruse 2]’
c. *¡Ah, si* [*yo* 1] [*le* 3] *contara!*
‘¡Oh, if [I 1] told [you 3]!’

The main problem in this process is to decide about the semantic equivalence between arguments of different syntactic patterns, and to decide if the examples are instances of the same verb sense. The general strategy has been to make as few verb sense distinctions as possible, reducing lexical entries are to a minimum.

Verb-specific description of semantic roles is also adopted in PropBank (Palmer et. al., 2005),

[3] http://adesse.uvigo.es/

[4] Note in passing that the database registers as arguments, not only full noun phrases and pronouns, but also clitics (*le*) and referents evoked by verb agreement like the A1 argument of (1a).

a project who aims to annotate a syntactically parsed corpus with information about argument structure. In this project, verbal arguments are labeled as numbered arguments, from Arg0 on.

Following the PropBank style, ADESSE also assigns a sequential number to each verbal argument: A0, A1, A2, … Nevertheless, there exist two important differences. The first one has to do with the scope of numbered arguments (we will turn to this question in section 3.3.). A second difference has to do with role labels. In PropBank, there is no semantic role label associated with each incrementally numbered argument, but only a brief description (generally, a formula of the type: ‘V-er’, ‘thing V-ed’) and, sometimes, the corresponding thematic role used in VerbNet (cf. Kipper et al., 2002).

In ADESSE, we usually do not suggest specific role labels on this level (but see Figure 2). If so, we would have to admit as many labels as existing slots for each verb recorded in the corpus[5]. However, our description of valency potential actually includes semantic role labels for each argument. In ADESSE, this information is directly inherited from the following more abstract level of representation, where types of situations and their corresponding types of participants must be considered.

3.2 Class-specific roles

Assuming that each situation is unique, the verbal lexicon of any language allow us to abstract commonalities from those partially different situations. With this idea in mind, one of the goals in ADESSE is to get a semantic classification of Spanish verbs by delimiting a set of possible conceptual classes or types of events. This is also a bottom-up process of grouping lexical entries. ADESSE’s classification has an ontological basis and a hierarchical structure, with up to four levels at the present stage[6]. Each semantic class is associated with a set of semantic roles which are prototypical for the conceptual domain evoked, so that verbs belonging to the same class will share the same subset of semantic roles.

The conceptual basis adopted in ADESSE to characterize types of events and participants is reminiscent of FrameNet (Fillmore et al., 2003). However, there are important differences between both projects (García-Miguel & Albertuz 2004). ADESSE classes and subclasses are much more schematic than frames in FrameNet: the 63 verb classes of ADESSE (for approximately 4000 verb entries) cannot reflect the fine-grained distinctions of the more than 1000 frames defined in FrameNet. Nevertheless, FrameNet has frames at different levels of schematicity, and an elaborated system of inheritance relations between frames. More schematic frames, inherited or used by more specific ones, are most similar to ADESSE classes and subclasses.

Some of the labels used for these class-specific roles may fit with traditional thematic roles (e.g. agent, patient, instrument, location, etc.). Nevertheless, role labels in ADESSE where chosen by aiming at two factors: specificity (depending on the verbal class) and transparency (descriptive adequation). Some of them are stated in the following table:

Class	**A0**	**A1**	**A2**
Feeling		Emoter	Emoted
Perception	Causer	Perceiver	Perceived
Cognition	Causer	Cognizer	Content
Possession		Possessor	Possessed
Transfer	Donor	Final-poss.	Possessed
Change	Agent	Patient	

Table 1. Some class-specific roles in ADESSE

Verb-specific arguments inherit by default the labels from class-specific roles. For example, the valency potential of *prestar* ‘to lend’, which is classified as a verb of ‘transfer’, is semantically described by making reference to the set of roles associated with that class, that is: A0: ‘Donor’, A1: ‘Final-Possessor’, A2: ‘Possessed’ (see Figure 1). The same set of labels is used to semantically annotate the arguments of verbs like *dar* ‘to give’, *pagar* ‘to pay’, *vender* ‘to sell’, etc:

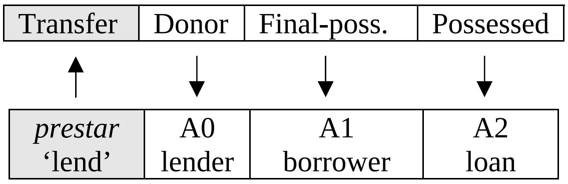

Figure 1. Verb-specific roles of *prestar*, a verb of Transfer.

Up to now, ADESSE comprises a total of 196 class-specific roles spread over 63 different se-

[5] So far, there are 4,016 verb meanings and 9,758 verb-specific arguments in ADESSE, giving an average of 2,4 arguments per verb.

[6] The whole semantic classification can be consulted in http://adesse.uvigo.es/data/clases.php.

mantic classes[7]. Given that the semantic classification is hierarchical, with up to four levels of more general and more specific process types, class-specific roles allow us to cover and define types of participants at different levels of generalization. So, for example, the class of ‘change’ is subdivided in three subclasses: a) verbs of creation (e.g. *crear* ‘create’, *producir* ‘produce’), b) verbs of modification (*abrir* ‘open’, *romper* ‘break’), and c) verbs of destruction (*destruir* ‘destroy’, *eliminar* ‘erase’). Each subclass is associated with a different set of semantic roles: a) Creator and Creation, b.) Agent and Affected, c) Destroyer and Destroyed. But the more schematic class of ‘change’ neutralizes these semantic contrasts, abstracting the common properties of the mentioned roles into an Agent and a Patient. Likewise, the class 'Mental process' includes the classes Feeling, Perception, and Cognition so that the semantic roles Experiencer and Stimulus, associated to the Mental class must be seen as generalizations over the participant roles of the more specific process types. These and other similar cases of generalizations concerning class-specific roles are summarized in figure 2:

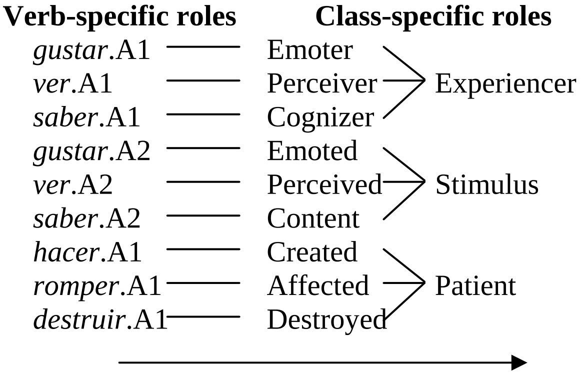

Figure 2. Semantic roles and levels of generalization

The set of relations between classes and class-specific roles in ADESSE is reminiscent of the network of inheritance relations between Frames and Frame Elements in FrameNet, although not as much fine-grained.

One might think that, by following this line of generalization, a maximally schematic level of representation could be achieved, so that we could get a limited set of semantic roles independently of process types.

As an equivalent of what is labeled ArgM in PropBank, we consider a small group of semantic roles for additional or secondary participants. These general roles (AG) are possible with verbs belonging to different semantic classes and allow to fully describe the valency potential of many verbs for which the inherited class-specific roles are not enough. The set labels used so far for these additional participants is: *Beneficiary, Location, Manner, Matter, Purpose, Reference, Attribute, Final State, Object, Means, Possessor, Facet, Company, Cause, Source, Role* .

However, for the more nuclear arguments, at the higher level of abstraction we must face a heterogeneous set of variables reflecting features of completely different semantic domains. Therefore, it is necessary to take into account the syntactic-semantic commonalities observed among the whole set of semantic roles.

3.3 Generalized semantic roles

There exist several linguistic theories which have dealt with a maximally schematic representation of argument linking (cf. Dowty 1991, Van Valin & LaPolla 1997, Croft 1998). Although different in many respects, all these proposals must be based on some kind of template or scale on which relative positions of arguments could be accounted for.

A usual way to do that is by starting from a logical decomposition of predicates based on Aktionsart distinctions, as proposed in RRG (cf. Van Valin & LaPolla, 1997). What these authors suggest is that all possible thematic relations can be summarized in only five distinctions, corresponding to the argument positions allowed by logical structure templates (Figure 3)[8]. As a result, a hierarchy is obtained from which two macro-roles are posited, Actor and Undergoer:

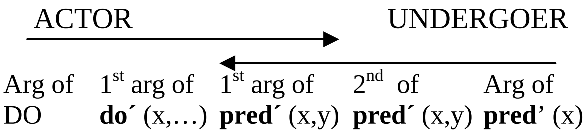

Figure 3. Actor-Undergoer hierarchy in RRG

Briefly, Actor macro-role fits with the subject of transitive and unergative verbs, while Under-

[7] An inventory much more bigger than the one used by AnCora (20 semantic role labels) or SenSem (32 semantic role labels)

[8] x and y are arguments of a predicate **pred'**. **do'** is a generalized activity predicate, and DO is a generalized causative predicate.

goer macro-role fits with the object of transitives and the subject of unaccusatives.

	Actor	**Undergoer**	**[other]**
KNOW	knower	thing known	
LEARN	learner	thing learned	
TEACH	teacher	thing learned learner	learner thing learned

Table 2. *Know, learn* and *teach* in RRG

A strategy based on correlative pointers to annotate predicate argument structures is used in PropBank: "An individual verb's semantic arguments are numbered, beginning with zero. For a particular verb, Arg0 is generally the argument exhibiting features of a Prototypical Agent (Dowty 1991), while Arg1 is a Prototypical Patient or Theme. No consistent generalizations can be made across verbs for the higher-numbered arguments, though an effort has been made to consistently define roles across members of VerbNet classes." (Palmer et al. 2005: 75). Therefore, in this project Arg0 is generally applied to the subject of transitive and unergative verbs, establishing similar correspondences to RRG (see Table 3).

	Arg0	**Arg1**	**Arg2**
KNOW	knower	thought	attributive
LEARN	learner	subject	teacher
TEACH	teacher	subject	learner

Table 3. *Know*, *learn* and *teach* in PropBank

Regarding ADESSE, we have already mentioned how verb arguments are incrementally numbered. However, beyond describing the valency potential of each verb, these numbered arguments can serve to represent generalizations from argument positions, in the way of variables in logical templates. In ADESSE, default pointers for arguments are chosen taking into account the following correspondences: A0=initiator or causer, A1=1st argument of **pred´**, A2=2nd argument of **pred´**. Schematically, we could trace the parallelisms between ADESSE hierarchy and the Actor-Undergoer hierarchy as follows:

A0	**A1**		**A2**
Arg of DO	1st arg of **do**´(x,…)	1st arg of **pred´**(x,y) or **pred´**(x)	2nd arg of **pred´** (x,y)

Figure 4. ADESSE hierarchy versus Actor-Undergoer hierarchy

As can be deduced from Figure 4, in ADESSE A0 is reserved for the first argument of causatives, so that we can see more easily the correspondences between causatives and their non-causative counterpart (Table 4).

	A0	**A1**	**A2**
SABER 'know'		knower [Cognizer]	thought [Content]
APRENDER 'learn'		learner [Cognizer]	subject [Content]
ENSEÑAR 'teach'	teacher [Causer]	subject [Cognizer]	learner [Content]

Table 4. *Saber, aprender* & *enseñar* in ADESSE

That way, a greater coherence with lexical meaning and lexical relations is achieved, while linking of semantics and syntax is understood in terms of relative positions in the argument scale. As can be seen in Table 5, Subject is almost always higher than DObj in the hierarchy of GSRs

Subj - DObj (+ oblique) in Active Voice		
Subj=**A1**	DObj=**A2**	61%
Subj=**A0**	DObj=**A1**	25 %
Subj=**A0**	DObj=**A2**	3 %
Other		10%

Table 5. Linking of grammatical relations and arguments. Frequency in ADESSE

4. Conclusion

We have outlined a system for describing semantic roles at different levels of granularity. About 326K arguments of 159K clauses have been given annotation at one or more levels in the database. The frequency of each role index is given in Table 6.

index	**more common class-specific role labels**	**N**
A0	Causer, Agent, Donor, Assigner, …	31521
A1	Theme, Cognizer, Communicator, Perceiver, Affected, Possessor, …	156958
A2	Content, Perceived, Possessed, …	103103
A3	Goal, Addressee, Perceived-2, …	16414
A4/A5	Path, Content-2, Activity, Code, …	4566
AG	Beneficiary, Location, Reference, ..	13312

Table 6. Frequency of arguments in ADESSE

Acknowledgements

This work has been supported by the Spanish Ministry of Science and Innovation (BFF2002-01197, HUM2005-01573, FFI2008-01953) the Galician Government (PGIDIT03PXIC-30201PN), and the University of Vigo.

References

Irene Castellón, Ana Fernández-Montraveta, Gloria Vázquez, Laura Alonso Alemany & Joan Antoni Capilla. 2006. The SenSem Corpus: A Corpus Annotated at the Syntactic and Semantic Level. *Fifth International Conference on LREC*, 355-359

William Croft. 1998. Event Structure in Argument Linking. In Miriam Butt & Wilhelm Geuder (eds.), *The Projection of Arguments: Lexical and Compositional Factors*, Standford, Center for the Study of Language and Information

David Dowty. 1991. Thematic Proto-roles and Argument Selection. *Language*, 67(3), 547-619

M. Ellsworth / K. Erk / P. Kingsbury / S. Padó. 2004. PropBank, SALSA, and FrameNet: How Design Determines Product. In *Proceedings of LREC-2004*, Lisbon.

Charles J. Fillmore, Christopher R. Johnson & Miriam R. L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16 (3):235-250.

José M. García-Miguel & Francisco Albertuz. 2005. Verbs, Semantic Classes and Semantic Roles in the ADESSE project. In K. Erk, A. Melinger & S. Schulte im Walde (eds): *Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, Saarbrücken, pp. 50-55.

José M. García-Miguel, Fita González Domínguez & Gael Vaamonde. 2010. ADESSE, a Database with Syntactic and Semantic Annotation of a Corpus of Spanish. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. (Valletta, Malta: Mai 2010). European Language Resources Association (ELRA)

Daniel Gildea & Daniel Jurafsky. 2002. Automatic Labeling of Semantic Roles. *Computational linguistics*, 28(3):245-288.

Karin Kipper, Martha Palmer & Owen Rambow. 2002. Extending PropBank with VerbNet Semantic Predicates. *Workshop on Applied Interlinguas*, Tiburon, CA

Ronald W. Langacker. 1991. *Foundations of Cognitive Grammar, Vol. II: Descriptive Application*, Standford, Standford University Press.

Beth Levin. 1993. *English Verb Classes and Alternations: a Preliminary Investigation*. Chicago / London: University of Chicago Press

Maria Antònia Martí, Mariona Taulé, Manu Bertran & Lluís Màrquez. 2007. *AnCora: Multilingual and Multilevel Annotated Corpora*, Draft. http://clic.ub.edu/ancora/ancora-corpus.pdf

Martha Palmer, Daniel Gildea & Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles, *Computational Linguistics*, 31(1), 71-106.

Volha Petukhova & Harry Bunt. 2008. LIRICS Semantic Role Annotation: Design and Evaluation of a Set of Data Categories. *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. (Marrakech, Morocco: Mai 2008). European Language Resources Association (ELRA).

Carlos Subirats. 2009. Spanish Framenet: A frame-semantic analysis of the Spanish lexicon. In Hans Boas, ed. *Multilingual FrameNets in Computational Lexicography. Methods and Applications*. Berlin/New York: Mouton de Gruyter, pp. 135-162.

Mariona Taulé, Mª Antonia Martí & Marta Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. (Marrakech, Morocco: Mai 2008). European Language Resources Association (ELRA).

Robert D. Van Valin & Randy J. LaPolla. 1997. *Syntax. Structure, meaning and function*. Cambridge, Cambridge University Press