Verbs, semantic classes and semantic roles in the ADESSE project

José M GARCÍA-MIGUEL **Francisco J ALBERTUZ**

ADESSE project
Facultade de Filoloxía e Tradución
University of Vigo
E-36200 Vigo, Spain
{gallego, albertuz}@uvigo.es

Abstract

This paper contains an overall description of ADESSE (http://webs.uvigo.es/adesse/), a project whose main goal is to manually provide definitions and information about semantic roles and semantic class membership for all the verbs in a syntactic database of nearly 160,000 clauses retrieved from a Spanish corpus of 1,5 million words.

1 Introduction

In this paper we outline the ADESSE (*Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español*) project, developed at the University of Vigo. The goal of the project is to achieve a database with syntactic and semantic information about verbs and clauses from a corpus of Spanish. The main final outcome of ADESSE will be a corpus-based syntactic-semantic database including for each verb and each clausal construction in the corpus a pattern of arguments characterized in terms of syntactic function, phrase type, semantic features, and semantic role. This will be accompanied by absolute and relative frequencies for each constructional alternative.

The starting point is a syntactic database of contemporary Spanish (BDS)[1], containing the syntactic analysis of almost 160,000 clauses from a corpus of 1,5 million words. The main tables of the BDS contain a register for each clause, including general grammatical features of the clause (verb form, polarity, modality, voice, etc.) and related fields for any core syntactic argument. For each syntactic argument, the following features are offered:

- [SynFunc]Syntactic Function: Subject, Direct Object, Indirect Object, Oblique Object, Locative, Manner, Oblique Agent, Attribute
- [Agr/Clit] Verb agreement or object Clitic (if any)
- [SynCat] Syntactic Category, i.e. phrase type
- Preposition (if any)
- Animacy: Human, Concrete, Abstract, Propositional
- Definiteness
- Number

Table 1 shows an example from the BDS with some of the syntactic information that has been annotated, namely, the syntactic features that we consider more relevant for ADESSE.

Cuando estaba en la universidad me escribía canciones de amor [TER:127] 'When he was at the University, he used to write love songs for me'			
SynFunc	Subj	DObj	IObj
Agr/Clit	3sg		*me*
SynCat		NP	
Animacy	Human	Concrete	Human

Table 1. Basic syntactic information about a clause in the BDS

One of the most evident benefits of the BDS is that we can get detailed information about the syntactic constructions of the verbs registered in the corpus. However, the utility of the database would increase greatly if we could also add some semantic features, a task that is also being developed independently by other semantic annotation projects (Ellsworth et al 2004; Sgall et al 2004). So, the goal of ADESSE is to keep all the syntactic information from BDS, and to create new tables and fields for the introduction of relevant semantic information: semantic roles, verb senses, and verb classes.

Our theoretical background assumes the independence and semantic compatibility of verb meaning and construction meaning (García-Miguel 1995:24-25, Goldberg 1995). We think that the global meaning of a sentence combines the meaning of lexical items and the meaning of grammatical constructions in a non deterministic way, but in a process of partial compositionality (Langacker 2000:152). We also adhere to some tenets of frame semantics, and particularly to

[1] BDS is partly accessible at http://www.bds.usc.es/

some practices of the FrameNet project[2], although there are also some important differences that will be commented on below. Put simply, we think that the syntactic structure of the clause must be explained through semantics. The verb evokes a complex conceptual representation that includes some basic participants in a scene. The syntactic alternations with the same verb provide alternate construals of the scene focusing on different facets of the situations. With this problems in mind, ADESSE aims to become a data base for the empirical study of the interaction between verb meaning and construction meaning.

2 Verbs and Semantic Arguments

As it has been observed, each verb evokes a conceptual scenario which can be accounted for by describing the set of potential semantic arguments which that verb can be combined with. For example, the conceptual frame of *escribir* 'write' can be described by making use of four semantic roles: 0-Writer, 1-Text, 2-Recipient and 3-Topic. Though sometimes it is possible to express the whole set of semantic arguments, as in (a), syntactic constructions usually select a subset, profiling them in different ways and leaving the rest unexpressed, as in (b) or (c):

(a) *Juan* [0] *le escribió una carta* [1] *a su madre* [2] *sobre sus recuerdos de infancia* [3]
'John wrote a letter to his mother about his childhood remembrances'
(b) *Juan* [0] *escribió una carta* [1]
'John wrote a letter'
(c) *Juan* [0] *le escribió a su madre*[2]
'John wrote to his mother'

What definitively proves that syntax is not enough is that, sometimes, the same syntactic construction can be mapped with different configurations of semantic arguments. Compare examples (b) and (c) below, from the verb *sustituir* 'substitute, replace', [0-Agent / 1-Substituted (Old Entity) / 2-Substitute (New Entity)], where the syntactic pattern Subj DObj corresponds to two semantic schemas (0-1 and 2-1):

(a) *Rijkaard* [0] *sustituyó a Xavi*[1] *por Deco*[2]
'Rijkaard replaced Xavi with Deco'
(b) *Rijkaard* [0] *sustituyó a Xavi*[1]
'Rijkaard replaced Xavi'
(c) *Deco*[2] *sustituyó a Xavi* [1]
'Deco replaced Xavi'

[2] See http://www.icsi.berkeley.edu/~framenet/ and, for an overview, Fillmore et al. (2003).

Finally, it is possible that (what is at first considered) one verb evokes, in different instances, frames corresponding to different semantic domains. For example, the verb *enseñar* admits uses as the following ones:

(a) *Ella* [0] *le* [2] *enseñaba su idioma* [1]
'She taught him her language'
(b) *Ella* [0] *le* [2] *enseñaba las fotos* [1]
'She was showing him the pictures'
(c) *Ella* [0] *enseñó al niño* [2] *a caminar* [1]
'She taught the baby how to walk'

It seems clear that we must distinguish two frames, one corresponding to the domain of Cognition (examples a and c, roughly equivalent to English *teach*, despite the differences in syntactic construction) and the other to Perception (example b, English *show,* despite the fact that the constructions is similar to that in a). In cases such as this one, we need different sets of semantic roles for labelling verb arguments [0-Teacher, 1-Thing taught, 2-Learner vs. 0-Shower, 1-Thing shown, 2- Seer], so we postulate two different verb senses.

In order to account for these and other similar facts, the design of our database takes a structure, whose main tables and relations are depicted in Figure 1

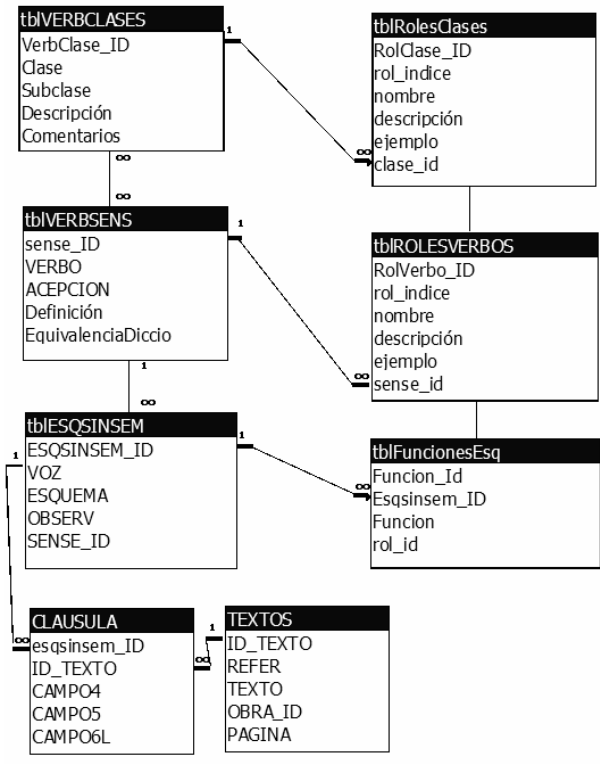

Figure 1: ADESSE database (partial) structure

Each record of the BDS ("Clausula" in fig. 1) is linked to a table of syntactic schemata ("tblEsq-

SinSem") where we map each syntactic function with a participant index (the equivalent of "0", "1", and "2" in the examples above). Each schema, in turn, is linked to a verb sense (tblVerbSens), associated with a set of participant roles, and ascribed to one or more semantic classes (tblVerbClases). The following sections explain the process in more detail.

3 Defining and unifying Verb Senses

Since our starting point is a database that contains very little semantic information, our first task has been to identify and define verb senses. This includes, among other things, a rough definition, a pointer to dictionary entries, and the splitting of a lemma into several verb senses when a unitary definition is not possible.

With respect to the distinction of verb senses, it must be remembered that our main interest is clause structure and not lexicology or lexicography, so we have not applied most of the criteria used in the lexicographical work. According to our theoretical background and our practical aims, we only distinguish verb senses when they are associated with different sets of semantic roles (see *enseñar* above). For example, the verb *escribir* has in ADESSE a single entry encompassing different subsets of a unique role set, despite the fact that some Spanish dictionaries distinguish up to three senses

Other lexical databases such as WordNet (Fellbaum 1998) follow a completely different way and admit a highly polysemic structure. That is, each possible group of synonyms ("synset") gives a new sense and then a new verb entry. So for example, WordNet 2.0 distinguishes 9 senses of *write*, 4 senses of *replace*, 13 senses of *show* and 2 senses of *teach*. ADESSE recognizes just one sense in each case for the Spanish equivalents of that English verbs.

Among the typical cases that do not imply different verb entries in ADESSE, one finds the following ones:

(a) *Constructional alternations,* whose meaning differences can be attributed rather to the constructional schema than to the verb. Under a single verb entry we can find voice alternations (active, middle, passive), causative/inchoative alternation, locative alternation, or some other rearrangement of arguments. In fact, the corpus recording of constructional alternations is the main goal of the ADESSE project.

(b) *Paradigmatic alternatives* inside an argument slot. Many verbs adjust their meaning depending on the nature of their more central arguments. For example, Spanish dictionaries distinguish about 15 senses of the verb *montar* 'mount', correlating with the nature of the thing mounted: a horse ('ride'), a concrete object ('assemble'), a business ('found', 'start'), an egg ('whip'), etc. Nevertheless, the schematic features of the argument structure do not vary very much and ADESSE contains just two senses of *montar*: 'ride' vs. 'assemble, set up'.

(c) *Metaphoric and metonymic uses* that can be extended or mapped from the basic sense of the verb. Nevertheless, although metaphoric uses do not suppose a new verb entry, they are identified and annotated in the corpus.

4 Verb Classes

In ADESSE each verb (in each sense) is given one (sometimes more) semantic class label(s). We use a hierarchical classification with two main levels: class and subclass. At the present we recognise 12 verb classes which reflect large semantic domains. Some classes can be grouped altogether into larger macroclasses, similar to some extent to Halliday's (2004) types of process:

MACROCLASS	CLASS	VERBS
1 Mental	11 Feeling	186
	12 Perception	72
	13 Cognition	122
2 Relation	21 Attribution	132
	22 Possession	117
3 Material Processes	31 Space	513
	32 Change	394
	33 Other facts	205
	35 Behavior	152
4 Communication		258
5 Existence		115
6 Causative and dispositive		57
TOTAL VERBS		

Table 2. *Top-level classes in ADESSE*

However, our basic and more useful category is subclass. Verb classes are therefore divided into 51 subclasses, associated with more concrete conceptual frames, each of which provides a (partially) specific set of semantic roles for labelling verb arguments (see below).

For example, the verb class Change splits into 5 subclasses as shown in Table 3:

SUBCLASS	VERBS
3200 General	14
3210 Creation	30
3220 Destruction-Consumption	35
3230 Modification	298
3231 Personal Care	17

Table 3. *Change subclasses in ADESSE*

In each verb class there is a General subclass including verbs with a more schematic content. For example, Change verbs such as *pintar* 'paint' or *cocinar* 'cook' are considered General Change verbs because they admit both Modification and Creation readings[3]. On the other hand, some subclasses are actually verbal groups inside a subclass, and identify more specific sets of verbs for further study. Thus Grooming or Personal Care verbs (3231), such as *lavar* 'wash' or *cepillar* 'brush', constitute a subtype of Modification verbs (3230), as reflected in the numerical index.

Unlike other verbal typologies, which use a fixed inventory of top-level categories, or which introduce the typology as the final outcome of a complete analysis of verbs, our classification is still provisional and its current structure represents working hypotheses about semantic organization that are always tested (and corrected, if necessary) for usefulness and empirical adequacy. As a point of departure, we have reviewed other semantic classifications, from the more lexically oriented (as WordNet) to the syntactical-semantic ones based on diathesis alternations (Levin 1993), though our premises fit better with proposals such as Dixon's (1991), Halliday's (2004) and FrameNet's (but see below). In fact, most of our clasess and subclasses are present in most classifications, but often with important differences in extension and hierarchical position. Some similarities between our system and WordNet high-level categories are evident.

Semantic verb classes in ADESSE are not empirically well-defined sets; rather, they represent generalizations over types of conceptual frames evoked by individual verbs in their specific instances, so problems of conceptual overlapping and fuzzy borders are expected, especially if, unlike WordNet, we are reluctant to divide verbal senses. Verbal meanings are multidimensional and highly flexible, and the classification of verbs is only possible by identifying the basic dimension(s) of meaning they profile and by keeping them apart from contextual influence. As an example, *frotar* 'rub' designates a Manner of Movement (without displacement, as *acunar* 'rock (to sleep)'), but it sems to profile a contact (as *tocar* 'touch') made by exerting force (as *presionar* or *pulsar* 'press') that can cause a modification/displacement of an entity (as *limpiar* 'clean'). Therefore, *frotar* has been classified as an Other Facts:Contact verb. Sometimes, however, verbs seem to equally profile more than one semantic dimension (and equally evoke more than one conceptual frame), so ADESSE allows multiple classification: *escribir* 'write' belongs to Change: Creation and Communication:General subclasses (as *crear* 'create' and *decir* 'say' respectively); *durar* 'last' is a verb of Existence:Time and also of Attribution:Value (as *tardar* 'delay' and *costar* 'cost' respectively), etc.

[3] Compare *No había cocinado espárragos desde que ella llegó a casa* 'She had not cooked asparagus since she had arrived home' [BAIRES:493, 21] with *Podríamos pasar las veladas* […] *cocinando "escudellas del Ampurdán"* 'We could spend the evenings […] cooking *escudellas del Ampurdán*' [a typical Catalonian dish] [AYER:24, 5].

5 Semantic Roles: Between Verb Senses and Verb Classes

The identification and annotation of semantic roles is a fundamental task of the project, given that the basic goal is to document empirically the linking of syntactic functions and semantic roles. This goal should be achieved at any predefined level: semantic class, verb senses, syntactic schemata, and clauses of the corpus. In order to simplify a bit the manual process of annotation and to achieve a greater coherence within the database, we assume that each level inherits by default the semantic information established in the higher levels; that is, in principle, we do not annotate each clause in the corpus, but the syntactic schemas that they instantiate. Syntactic schemas, in turn, point to roles that are defined for each verb sense. And verb participant roles can inherit features and labels from class-defined participant roles. In any case, we account for the possibility that each lower level contradicts or increments the information inherited from the higher levels.

First, each conceptual (sub)class is associated with a set of semantic roles prototypical for the cognitive domain denoted by the verbs belonging to it. Role labels are created by aiming at specificity (with class-specific labels) and transparency (descriptive adequation), trying to use, as far as possible, widely used traditional labels. Here are the role labels associated with some classes:

Change:Modification:
A0:Agent; A1:Affected

Communication:
A1:Sayer; A2:Message; A3:Addressee; A4:Topic

Feeling:
A1:Experiencer; A2:Stimulus

Possession:Belonging:
A1:Possessor; A2:Possessed

Space:Displacement
A0:Causer; A1:Theme; A2:Source; A3:Goal

Secondly, each verb entry is associated with a set of semantic roles embracing any possible core participant in the scenes designated by the verb in any syntactic schema (see above examples with *escribir, sustituir,* and *enseñar*). In general, a set of explicit inheritance relations makes a verb inherit by default the roles considered basic for the class to which it belongs, although some verbs need some additional arguments in order to account for any syntactic construction with such verbs. For example, the verb *sustituir*, a member of the class Other facts:Substitution, inherits a set of roles that is common to other verbs of the same class (*reemplazar, cambiar2, suplir*, etc):

	A0	A1	A2
SUBSTITUTION	Agent	Substituted	Substitute
Sustituir	Agent	Substituted	Substitute

However, verb-specific role labels are used whenever there is a total or partial mismatch between a verb argument and class-specific role labels. For example, the verb *escribir* 'write' is both a Creation verb and a Communication verb. Its argument roles are inherited from Creation (Agent – Effected – Beneficiary) and from Communication (Sayer – Message – Addressee – Topic); but for the sake of clarity, the first two participants are labelled as Writer and Text.

	A1	A2	A3
COMMUNICATION	Sayer	Message	Addressee
Escribir	Writer	Text	Recipient

Third, the syntactic constructions of each verb are annotated simply with a pointer from each syntactic argument to one of the roles defined for the verb entry. This pointer allows us to trace the correspondences between arguments of different syntactic schemas (the pointer being identical for the equivalent arguments of diathesis alternations such as active / passive, causative / inchoative and so on). For example, in Figure 2, both the active voice object [D] and the passive voice subject [S] get the pointer "1", indicating the Text written[4]. Given that syntactic functions are linked to a pointer, we could change the labels or the details of the classification without touching the essential aspects of the diathesis alternations.

Multiplying syntactic schemas by verb senses, we get about 12500 syntactic-semantic schemas that constitute the main target of our annotation. Given that each clause of the corpus is being linked to a syntactic-semantic pattern, we think

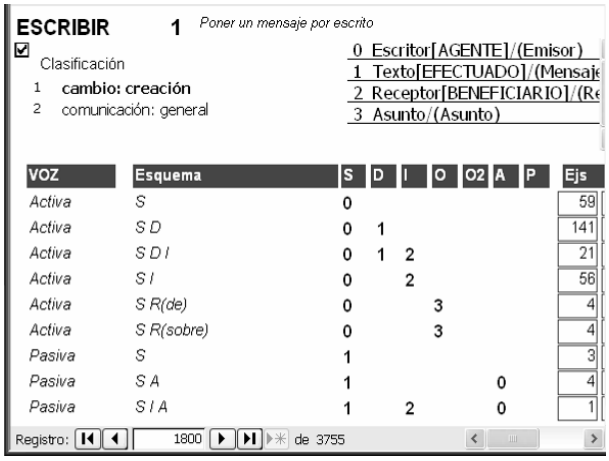

Figure 2. Patterns of *escribir* in ADESSE

that this strategy will allow us to characterize semantically the 159,000 clauses of the corpus in a relatively short time. This way, each clause is receiving an annotation similar to Table 4, which expands the example in Table 1.

Me escribía canciones de amor [TER:127] 'He used to write love songs for me'			
Escribir	Writer	Text	Recipient
CREATION	Agent	Effected	Benefactive
COMMUNIC.	Sayer	Message	Addressee
SynFunct	Subj	DObj	IObj
Agr/Clit	3sg		*me*
SynCat		NP	
Animacy	Human	Concrete	Human

Table 4. Syntactic and semantic annotation of arguments in a clause of BDS+ADESSE

6 Comparing with FrameNet

Our classification has a clear conceptual basis, which makes it very similar in some respects to FrameNet. Nevertheless, there are some important differences, beginning with the fact that we use a syntactically analyzed corpus to semantically annotate all and only the clauses in the corpus, not a set of selected sentences that illustrates frames and lexical units.[5]

Moreover, in FrameNet, the basic unit is obviously the *Frame*, so that Frame Elements and Lexical Units are defined in relation to the frame they belong to. In ADESSE, by contrast, the basic unit is the verb. Classes and subclasses represent generalizations over argument configurations in an attempt to get a set of role labels applicable by default to the verbs of the same class.

On the other hand, and more relevant in practice, ADESSE classes and subclasses are much

[4] This strategy has many similarities with PropBank annotation procedure (Kingsbury-Palmer 2002).

[5] In this respect, our goal is similar to that of PropBank and SALSA (Ellsworth et al 2004).

more schematic than frames in FrameNet[6]. This appears to be self-evident if we compare our 52 classes with the more than 300 frames containing verbs. Therefore, in ADESSE verbs such as *ver* 'see' and *mirar* 'look at' or *oír* 'hear' and *escuchar* 'listen' are included in the Perception class, disregarding semantic features as intentionality or attention which justify the FrameNet distinction between Perception_Experience and Perception_Active frames.

In line with our theoretical background, in ADESSE we try to keep apart verb meaning and construction meaning, and consequently we do not delimit verb senses simply on the basis of constructional alternations. FrameNet dissociates in different frames, for example, any verb participating in the locative alternation. Therefore, *load* in *John loaded the wagon with hay* is assigned to the frame Filling, whereas *load* in *Betty load the stuff in the car* is included in the frame Placing. By contrast, ADESSE unifies the spatial senses of *cargar* 'load' under just one verb sense under the class Localization. The meaning differences observed as a consequence of the 'locative alternation' are attributed to the meaning of the respective argument-structure constructions (in line with Goldberg 1995).

Moreover, ADESSE classes allow a variable degree of correspondence between a verb's argument structure and the pattern of participant roles prototypical for the class it belongs. For example, *mentir* 'lie' and *callar* 'be silent' are Communication verbs although *mentir* does not combine with a Message nor *callar* with a Recipient.

Last, apart from class-specific role labels, ADESSE can use verb-specific role labels. By default, verb-specific role-labels are inherited from class-specific role-labels, even though a verb can have a set of roles partly different from the class to which it is ascribed. This is the case of the verb *escribir* 'write' commented above. The use of verb-specific role-labels does away with the need to create new frames whenever the class or subclass is too wide.

7 Conclusion

At the time of writing this paper, the ADESSE project contains a provisional semantic classification of about 1700 verb senses, and an index of semantic role for each argument of about 4000 syntactic-semantic schemas, which correspond to more than 50000 clauses of the corpus. There is a lot of work to be done, but we aim to achieve a useful database for descriptive studies of the interaction between verbs and constructions in Spanish. So that we can obtain, for example, the diathesis alternations for any verb, the syntactic realizations of a participant role, or the syntactic constructions for a semantic domain (and vice versa).

[6] Nevertheless, FrameNet has frames at different levels of schematicity. More schematic frames, inherited or used by more specific ones, are most similar to ADESSE classes and subclasses. In fact, FrameNet I grouped specific frames into semantic 'domains'.

8 Acknowledgements

ADESSE is being supported by the University of Vigo, the Spanish Ministry of Science and Technology (BFF2002-01197) and the Galician Autonomous Government (PGIDIT03PXIC-30201PN). We also acknowledge a lot of people in Vigo and Santiago that have contributed to the building of the BDS for many years.

References

Dixon, Robert M. W. 1991. *A New Approach to English Grammar, on Semantic Principles*, Oxford University Press, Oxford.

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

Fellbaum, Christiane. 1998. "A Semantic Network of English Verbs". In *WordNet: An Electronic Lexical Database*, Fellbaum, Christiane, ed., pages 69-104, MIT Press, Cambridge (MA).

Fillmore, C.J. / C. Johnson / M. Petruck. 2003. Background to FrameNet. In *International Journal of Lexicography,* 16/3: 235-250.

García-Miguel, José M. 1995. *Transitividad y complementación preposicional en español.* Universidade de Santiago de Compostela.

Goldberg, Adele. 1995. *Constructions. A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago

Halliday, M.A.K. 2004. *An Introduction to functional grammar*. E. Arnold, London (3rd edition)

Kingsbury, P. and M. Palmer. 2002. From TreeBank to PropBank. In *Proceedings of LREC-2002*. Las Palmas.

Langacker, Ronald. 2000. *Language and Conceptualization.* Mouton de Gruyter, Berlin.

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

Sgall, P. / J. Panevová / E. Hajičová. 2004. Deep Syntactic Annotation: Tectogrammatical Representation and Beyond. In *Proceedings of HLT-NAACL-2004*. Boston.