ADESSE. A Database with Syntactic and Semantic Annotation of a Corpus of Spanish

Gael Vaamonde, Fita González Domínguez, José M. García-Miguel
Universidade de Vigo
Facultade de Filoloxía e Tradución, Campus Universitario, 36310 Vigo
{ gaelv, fitagd, gallego }@uvigo.es

Abstract

This is an overall description of ADESSE ("Base de datos de verbos, Alternancias de Diátesis y Esquemas Sintactico-Semánticos del Español"), an online database (http://adesse.uvigo.es/) with syntactic and semantic information for all clauses in a corpus of Spanish. The manually annotated corpus has 1.5 million words, 159,000 clauses and 3,450 different verb lemmas. ADESSE is an expanded version of BDS ("Base de datos sintácticos del español actual"), which contains the grammatical features of verbs and verb-arguments in the corpus. 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. Each verb entry in the database is described in terms of valency potential and valency realizations (diatheses). The former includes a set of semantic roles of participants in a particular event type and a classification into a conceptual hierarchy of process types. Valency realizations are described in terms of correspondences of voice, syntactic functions and categories, and semantic roles. Verbs senses are discriminated at two levels: a more abstract level linked to a valency potential, and more specific verb senses taking into account particular lexical instantiations of arguments.

1. Introduction

It is an undeniable fact that computational devices and electronic resources have changed and facilitated our linguistic research. However, it appears to be the case that the simple collection of data may not exhaust the demands and interests of the researcher. Put in other words, raw corpora become insufficient for many linguistic purposes, regardless of the amount of data compiled. Current linguistic research requires the development of detailed syntactic and, most of all, semantic annotation of these corpora, in order to provide the researcher with real useful resources. This paper aims to be an overall description of one of those resources for Spanish: the ADESSE database[1].

ADESSE stands for *Base de datos de verbos, **A**lternancias de **D**iátesis y **E**squemas **S**intáctico--**S**emánticos del **E**spañol* and gives the name to a project developing at the University of Vigo. The main goal of this project is to achieve an online database providing with exhaustive syntactic and semantic annotation about verbs and clauses from a corpus of Spanish.

With this goal in mind, our major purpose is to get a corpus-based database for the empirical study of the interaction between verbs and constructions in Spanish.

2. Initial resources: the corpus and the BDS

All the data annotated in ADESSE come from the ***AR**chivo de **T**extos **H**ispánicos de la **U**niversidad de **S**antiago de Compostela* (ARTHUS), which contains texts in Iberian and American Spanish published from 1980 to 1991. The size of this corpus rises to 1.5 million words, 159.000 clauses and 3500 verbs, and the texts selected cover, on different percentages, essay, narrative, spoken, press and theater.

As part of a project leaded by the University of Santiago de Compostela between 1989 and 2000, the 159.000 clauses of ARTHUS were syntactically annotated and incorporated into a database called BDS -**B**ase de **D**atos **S**intácticos del Español Actual (cf. Rojo 2001) [2].

Specifically, the grammatical features provided by the BDS include, for each clause: clause type and function, mood, tense, modal and phase auxiliaries, polarity, illocutionary force and voice. Moreover, each syntactic argument was enriched with relevant grammatical information, such as syntactic function (Subject, Direct Object, Indirect Object, Oblique Object, Oblique Agent and Predicative), syntactic category (i.e. phrase type), verb agreement or object clitic (if any), preposition (if any), animacy, definiteness and number.

Without denying the obvious usefulness of the BDS, it is equally true that this utility would increase greatly if we could add a full semantic annotation besides the syntactic features above described. ADESSE was primarily designed to tackle that semantic annotation.

Therefore, ADESSE inherits all the syntactic information offered by the BDS and, additionally, provides with specific semantic information, which can be broken down into three main tasks: (1) splitting and defining verb senses, (2) classifying verb senses in semantic verb classes, and (3) annotating verb arguments with semantic roles.

Part of this information can be summarized on the following Table, based on a record (i.e. a clause) extracted from the database:

[1] Other overall introductions about ADESSE can be found in García-Miguel, Costas & Martínez. (2005) and García-Miguel & Albertuz (2005), although different aspects of the database have been updated and expanded since then.

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

Al levantarse, Julián sintió un zumbido en los oídos [JOV:63,41] "When he stood up, Julián felt a ringing in his ears"			
CLAUSE			
PRED.	SENTIR		
Sense	I.1 'to perceive by the senses'		
Class/Domain	Perception		
Modality	Declarative		
Voice	Active		
ARGUMENTS			
Sem. role	A1: Perceiver	A2: Perceived	A-G: Locative
Syn. function	Subject	Direct Obj.	Oblique Obj.
Syn. category	NP	NP	*en* NP
Agreement	3ª sg.		
Lexical head	*Julián*	*zumbido*	*oído*
Animacy	human	concrete	concrete

Table 1: Part of a record in ADESSE

In the following sections, we will try to account for each of these points. Nevertheless, before going into depth, it will be necessary to account for some preliminary information on the strategies used in the building of ADESSE.

3. Beyond syntactic annotation

The data of the BDS provide us with the range of syntactic patterns which a verb can be combined with. However, we must assume that mere syntactic information is not enough if one wants to get a detailed description about argument structure of verbs. A pair of examples may illustrate this fact.

If we search for the verb *vestir* 'dress' in the BDS, we will get the following syntactic schemas, among others:

(1) a. Active Subject – Direct Object
Juan vistió a su hijo
'John dressed his son
b. Active Subject – Direct Object - Oblique Object
Juan vistió a su hijo de soldado
'John dressed his son as a soldier'
c. Reflex. Subject – Direct Object
Juan se vistió
'John got dressed'

In Spanish, this verb can also be used with the meaning of 'wear', as in the following sentence:

(2) a. Active Subject – Direct Object
Juan viste una camisa blanca
'John wears a white shirt'

Therefore, the same basic transitive pattern (i.e. Active Subj-DObj) is used to express a different subset of participants involved in the situation. On the one side, we get "the one who dresses" (let's name it [0]), functioning as Subj, as well as "the the one who gets dressed" (let's name it [1]), functioning as DObj. On the other side, we get "the one who gets dressed" ([1]) as Subj and "the thing/garment worn" ([2]) as DObj:

(3) a. Active Subject – Direct Object
Juan [0] *vistió a su hijo* [1]
'John dressed his son
b. Active Subject – Direct Object
Juan [1] *viste una camisa blanca* [2]
'John wears a white shirt'

The same syntactic pattern can be mapped with different configurations of semantic arguments ([0-1] or [1-2]). Regarding the verb *vestir*, it could be still possible to use the BDS in order to discriminate each set of patterns, by appealing to additional grammatical features (e.g. "animacy"). That is, we could get all the examples of *vestir* in the BDS which are similar to (3a) but different to (3b), since only examples which pattern like (3a) have an animate DObj. But obviously, this strategy only works as an *ad hoc* arrangement, not as a systematic method for refined searches, regardless of the verbs and arguments involved.

Consider now the following instance. With the verb *elegir* 'choose', the same syntactic pattern (Active Subj-DObj) corresponds again to two semantic schemas. The Subj expresses "the person who chooses" (let's name it [1]) in both cases, while the DObj includes two semantic arguments: "the person or thing which is chosen", ([2]), as well as "the role carried out by that person/thing", ([3]):

(4) a. Active Subject – Direct Object
Los españoles [1] *eligieron a Zapatero* [2]
'The Spanish chose Zapatero'
b. Active Subject – Direct Object
Los españoles [1] *eligieron presidente* [3]
'The Spanish chose president'
c. Active Subject – Direct Object – Predicative
Los españoles [1] *eligieron a Zapatero* [2] *presidente* [3]
'The Spanish chose Zapatero as president'

Animacy arises now as a useless criterion to separate examples like (4a) from examples like (4b), since there is no clear correspondence between that feature and the set of arguments involved in each semantic schema. Therefore, without additional semantic annotation we have no chance to discriminate and recover such information from a corpus.

On the other side, if cases such as *vestir* or *elegir* show how a single syntactic form can involve two (or more) semantic schemas, a discrepancy between syntactic and semantic schemas can runs along the opposite direction. That is, the same set of semantic arguments can be expressed by means of different syntactic schemas. The verb *arrojar* 'throw' can illustrate this fact:

(5) a. Active Subject – Direct Object – Indirect Object
Ellos [0] *le arrojaron piedras* [1] *al ladrón* [2]
'They threw stones at the thief"
b. Active Subject – Direct Object – Oblique Object
Ellos [0] *arrojaron piedras* [1] *a la ventana* [2]
'They threw stones at the window'

In (5), we have two ways to express a same set of participants associated with *arrojar* ("the one who throws" ([0]), "the thing thrown" ([1]) and the "goal of the throwing" ([2])). In (5a), the IObj is used for the third participant, while in (5b) an oblique stands for the expression of the "goal".
Be it as it may, what cases such as (3), (4) and (5) suggest is that we need additional semantic annotation in order to approach the interaction between verbs and constructions. The basic strategy applied in ADESSE to annotate that information starts from a distinction between valency potential and valency realizations (Agel 1995).

4. Basic strategies: valency potential and valency realizations

The valency potential of a verb is the set of potential arguments which can be selected by that verb, while the valency realizations refer to the set of argument which are really expressed by each syntactic form.
Consider the verb *regresar* 'return'. The valency potential of *regresar* can be described by making use of four semantic roles: Theme [1], Source [2], Goal [3] and Path [4]. It could be the case that the whole set of potential verb arguments is expressed by means of a single syntactic realization, as in (6a). However, generally each syntactic realization selects only a subset of the potential arguments a verb can combine with ((6b), (6c), (6d)):

(6) a. *El buque* [1] *regresó* [2] *a Vigo desde Malta* [3] *por el estrecho* [4]
'The ship returned to Vigo from Malta through the strait'
b. *El buque* [1] *regresó desde Malta* [3]
'The ship returned from Malta'
c. *El buque* [1] *regresó a Vigo* [2] *por el estrechó* [4]
'The ship returned to Vigo through the strait
d. *El buque* [1] *regresó*
'The ship returned'

With this problem in mind, a basic strategy in ADESSE is to define the valency potential of each verb, i.e. the whole range of participants which are possible with that verb, and to register in the corpus all the valency realizations which are actually expressed. Returning to the verb *regresar*, this task leads us to get in ADESSE the following information:

Valency potential of of REGRESAR 'return'			
A1 (Theme)	A2 (Source)	A3 (Goal)	A4 (Path)
Valency realizations (some of them in active voice)			
A1:Subj		A3:Loc(a)	
A1:Subj			
A1:Subj	A2:Loc(de)		
A1:Subj		A3:Loc	
A1:Subj			A4:Loc(por)
A1:Subj		A3:Loc(con)	

Table 2: Valency potential and valency realizations of *regresar*

The main task we are concerned with is to annotate which of the potential participants is selected in each syntactic scheme of each verb. This information is additionally accompanied by absolute and relative frequencies about different quantitative data (e.g. arguments, syntactic-semantic patterns, verb classes, …):

Potencial valencial

Indice	Etiquetas	Frecuencia
A1	MÓVIL (MOV)	200 (100 %)
A2	ORIGEN (ORI)	27 (13.5 %)
A3	DIRECCIÓN (DIR)	104 (52 %)
A4	TRAYECTO (TRA)	2 (1 %)

Table 3: Valency potential of *regresar* , with frequencies about semantic roles (adapted from ADESSE's Web site)

Realizaciones valenciales

Voz	Argumentos	N_ejemplos
$REGRESAR_{act}$	**A1**:*MOV* = SUJ; **A3**:*DIR* = **a** LOC	95
$REGRESAR_{act}$	**A1**:*MOV* = SUJ	67
$REGRESAR_{act}$	**A1**:*MOV* = SUJ; **A2**:*ORI* = **de** LOC	26
$REGRESAR_{act}$	**A1**:*MOV* = SUJ; **A3**:*DIR* = LOC	5
$REGRESAR_{act}$	**A1**:*MOV* = SUJ; **A4**:*TRA* = **por** LOC	2
$REGRESAR_{act}$	**A1**:*MOV* = SUJ; **A3**:*DIR* = **con** LOC	1
$REGRESAR_{act}$	**A1**:*MOV* = SUJ; **A3**:*DIR* = **junto a** LOC	1
$REGRESAR_{act}$	**A1**:*MOV* = SUJ; **A2**:*ORI* = **desde** LOC	1
$REGRESAR_{act}$	**A1**:*MOV* = SUJ; **A3**:*DIR* = OIND	1
$REGRESARSE_{impers}$	**A1**:*MOV* = SUJ; **A3**:*DIR* = **a** LOC	1

Table 4: Valency realizations of *regresar*, with absolute frequencies of each syntactic-semantic pattern (adapted from ADESSE's Web site)

Taking this basic strategy into account, we will focus from now onto the primary semantic information which can be accessed through ADESSE.

5. The lexicographic task. Defining verb senses

In spite of the existing lexicographic tradition, the attempts to establish criteria for identification, definition and order of senses are limited. Moreover, well-known Spanish dictionaries such as *Diccionario de Uso del Español (DUE), Diccionario del Español Actual (DEA)*

and *Diccionario de Construcción y Régimen (DCR)* show significant differences among them in semantic analysis.
This proves that each analysis depends on the level of granularity meant by the researcher. Since meaning emerges from context, there is a continuum of contextualized uses, and the most difficult lexicographic decisions is the selection of an appropriate level of granularity. If we split a verb in many different senses, we lose generalizations. If we don't split, we get categories which are too heterogeneous. Trying to avoid the disadvantages of any of those two approaches, in ADESSE we have proceeded in two steps: a first level of verb meaning, associated with a semantic domain and a set of participant roles, an a second level of particular specific verb 'senses'
In the first level of analysis, we have put together the examples making a minimum distinction of meaning (homonyms and quasi-homonyms). The 3436 verb lemmas of the corpus have become over 4000 verb entries. For example:

(7) Perder-I: 'To lose, to leak'
¿Tú sabes cómo perdí mi pierna?
'You know how I lost my leg?'
No ha perdido su sonrisa
'She has not lost his smile'
Perdía de siete a ocho kilos todas las Semanas
'He lost 7 or 8 kilos every weeks'

(8) Perder-II: 'To lose [a competition], to be defeated'
El equipo de Serra Ferrer perdió el pasado domingo en Valladolid
'Serra Ferrer's team lost on Sunday in Valladolid'
Arancha Sánchez y Helena Sukova perdieron ante Martina Navratilova y Pam Sriver
'Arancha Sanchez and Helena Sukova lost to Martina Navratilova and Pam Sriver'

(9) Perder-III: 'To miss, to waste'
Se marchó antes de lo habitual para no perder el avión
'He left earlier than usual to not miss the plane'
No pierdas más tiempo
'Don't waste more time'

The verb entries correspond with the most general meaning. In those entries we have included examples that share some semantic features and have a common *valency potential*.
Therefore, each verb entry include all the meanings related either literally or figuratively.
The second level corresponds with the identification of specific uses. At present, in the database there are 5059 dictionary entries, because in this second phase, 584 lemmas have been analyzed, of which 226 are of the most common (with over 110 examples). This means that have been reviewed over 150,000 clauses to identify its meaning. For instance, some uses of *perder-I* are (senses and microsenses):

(10) Perder-I:
1.-Dejar de tener [una parte o una propiedad material 'To lose'
2.-Dejar de tener [a alguien] por muerte, desaparición o desamor 'To miss [someone] for death, disappearance or lack of love'
3.-Dejar de tener [cualidades, un estado sentimientos] 'To miss [qualities, states, feelings]'
4.-Resultar perjudicado [en un negocio o acción] donde se persiguen unos beneficios 'Don't make profits in a business or action'
5. - Adelgazar. Bajar [de peso] 'To slim down'. 'To lose weight'

As a result, a hierarchical description of verb meanings is obtained.
So, delimiting meanings is to look for differences and resemblances among the verb uses, always in two directions: from general to particular and vice versa. This strategy allows us to establish the level of independence or unification among the examples and if it is appropriate to differentiate new meanings.
Nowadays, in addition to the identification and annotation of verb senses and microsenses, we are considering lexical realizations of arguments. It is important to bear in mind that the annotation of lexical arguments helps us to distinguish senses and meanings. Lexical features of the arguments, (i.e concrete, abstract, animate or inanimate) are a useful information for lexicographic task. For instance, in Spanish it is not the same *perder: las llaves* 'keys', *ocho kilos* 'eight kilos', *el avión* `plane', *media hora* 'half hour'. The lexical information makes the study of Verb+N combinations and support verbs easier.
The syntactic and semantic information of the corpus we are adding show and complete the *lexical or behavioural profile* (Hanks 1996) of the verb in every level of analysis, that is, the range of constructions and other lexical items with which a particular verb regularly co-occurs. As a result, the syntactic-semantic annotation presented in the database aims to provide a whole characterization of meaning and lexical profile to every single verb.

6. Types of situations. Delimiting semantic domains

There exist basically two main criteria of semantic classification of verbs. One of them lies on the notion of lexical aspect, frequently known as 'aktionsart', and allow us to distinguish between 'states', 'activities', 'achievements', 'accomplishments', … The second one adopts a more ontological perspective and allow us to establish conceptual classes, like 'verbs of perception', 'verbs of cognition', 'verbs of contact', … Some Spanish resources like AnCora (Taulé et al. 2008) and SenSem (Vázquez et al 2006) have given priority to the first criterion of classification; others like FrameNet (Fillmore et al 2003; and, for Spanish, Subirats 2009) is akin to the second.
The understanding of verb meaning in ADESSE fits well

with the ontological criterion. We think that each individual verb evokes a conceptual frame, that is, a complex conceptual representation which includes some basic participants in a scene (the valency potential, as was described above). The main goal of ADESSE verb classification is to represent generalizations over these types of conceptual frames evoked by individual verbs.

For example, consider the verb *ver* 'see'. This verb evokes a situation where we must assume two basic participants: 'someone who sees' and 'someone or something seen'. A verb like *escuchar* 'listen to' also evokes two participants: 'the listener' and 'the listenee'. Similar cases are illustrated by cases such as *mirar* 'look', *observar* 'observe', *advertir* 'notice', *oír* 'hear' and so on. Therefore, we can generalize over all of them and stipulate a type of 'perceptual' situations, where two basic participants are involved: a perceiver and a perceived.

Applying the same strategy, we can suggest another category including verbs like *gustar* 'like, *sentir* 'feel', *sufrir* 'suffer' or *disfrutar* 'enjoy', since all of them can be seen as verbs of 'feeling', that is, verbs involving a relationship between an experiencer and an some kind of stimulus.

Finally, following this generalization process we can abstract the common aspects shared by verbs of perception, verbs of feeling and even other type of verbs, like verbs of cognition (*pensar* 'think', *entender* 'understand') or verbs of election (*decidir* 'decide', *elegir* 'choose'). All these kind of processes share the feature of being mental activities where someone 'experience' something. Therefore, all of them are included in a more abstract verbal category: the mental class.

We can summarize this idea in table 4, which includes part of the semantic classification developed in ADESSE:[3]

Semantic Classes			Ex.
MENTAL	Feeling		*gustar*
	Perception		*ver*
	Cognition	Knowledge	*saber*
		Belief	*creer*
RELATIONAL	Attributive	…	*ser*
	Possession	…	*tener*
MATERIAL	Space	Displacement	*ir*
		Location	*poner*
		…	
	Change	Creation	*crear*
		Modification	*romper*
		Destruction	*destruir*
	Other facts	Contact	*tocar*
		Emission	*emitir*
		Meteorology	*llover*
		…	

Semantic Classes			Ex.
VERBAL	Communic.		*decir*
	Judgement		*criticar*
EXISTENTIAL		...	*existir*
MODULATION	Causative	…	*obligar*
	Dispositive		*tratar*
	Support V		*dar*
	…		

Table 4: Part of the semantic classification in ADESSE

As we can see from the table above, ADESSE classification is not only basically conceptual but structurally hierarchical.

At top level, we distinguished six main groups or macroclasses, similar to some extent to Halliday's (1985) types of process. Each of these macroclasses are split into different classes (e.g. space or change, within the material processes), which reflect large semantic domains. Most of them are in turn subdivided in several subclasses, thus producing a third hierarchical level. This level is associated with more specific conceptual frames, so that the semantic domain reflected by the previous level can be appropriately refined. For example, the general set of verbs of Space is organized in ADESSE taking into account six subclasses:

Subclasses	Examples	Nº
Displacement	*salir* 'leave'	232
Location	*poner* 'put'	230
Union	*incluir* 'include'	138
Posture	*sentar* 'sit'	46
Manner-of-movement	*temblar* 'tremble'	42
Orientation	*volver* 'turn'	10

Table 5: 'Space' subclasses in ADESSE

Finally, a fourth level is considered in some cases where further semantic subdivisions can be established within subclasses. For example, within the Relational processes two main classes are considered: attributive verbs and possession verbs. The first one is in turn subdivided in three subclasses: verbs of relation (e.g. *representar* 'represent, *relacionar* 'relate'), verbs of property (e.g. *resultar* 'be', *quedar* 'be') and verbs of naming (e.g. *definir* 'define', *calificar* 'describe'). However, within the 'property' set we can specify two further groups of verbs, depending on the nature of the property assigned to an entity: verbs of measure (e.g. *costar* 'cost', *pesar* 'weigh'), if the property refers to a measurable quantity; and verbs of appearance (e.g. *oler* 'smell', *saber* 'taste'), if the property refers to something directly perceived.

The following Table summarizes the arrangement of Relations processes in ADESSE and serve to illustrate the four levels of generalization considered in our semantic classification:

[3] The whole semantic classification can be consulted in http://adesse.uvigo.es/data/clases.php. More information can be found in Albertuz (2007).

Level 1	Level 2	Level 3	Level 4
(2) Relational	(2.1) Attributive	(2.1.1) Relation	
		(2.1.2) Property	(2.1.2.1) Measure
			(2.1.2.2) Appearance
		(2.1.3) Naming	
	(2.2) Possession	(2.2.1) Acquisition	
		(2.2.2) Transference	

Table 6: Relational processes in ADESSE

As it can be inferred from the several tables above, ADESSE's ontological and hierarchical classification of verbs is clearly not exempt from problems. Delimiting semantic domains and assigning verbs into them entails obvious complexities, which arise from the nature of the semantics itself.

These complexities increase if we take into account that (macro)sense distinctions in ADESSE have been limited to a minimum, so each lexical entry receiving semantic classification actually includes a set of related senses (v. section 5).

In achieving the classification of verbs, the strategy in ADESSE is primary to focus on the more prominent aspects of the meaning of each verb and, as far as possible, to compare it with the prototypical cases of each class.

However, many verbs seem to allow or even require multiple categorization. With this problem in mind, we consider the possibility of a single verb being assigned to two semantic classes.

For example, the verb *clavar* 'hammer (a nail)' belongs simultaneously to Contact (as *golpear* 'hit', *tocar* 'touch', *morder* 'bite, *chocar* 'collide' …) and Location (as *poner* 'put', *colocar* 'place', *cargar* 'load', *colgar* 'hang', …). The verb *unificar* 'unify' belongs simultaneously to Modification (as *cambiar* 'change', *romper* 'break, *limpiar* 'clean', *organizar* 'organize', …) and Union (as *reunir* 'collect', *incluir* 'include', *agregar* 'add', *incorporar* 'join'). And the verb *discutir* 'argue' belongs simultaneously to Communication (as *decir* 'say', *preguntar* 'ask', *hablar* 'talk', *exclamar* 'exclame', …) and Competition (as *luchar* 'fight', *vencer* 'defeat', *competir* 'compete', *atacar* 'attack', …).

7. Semantic role annotation

One basic goal in ADESSE is to document empirically the linking of syntactic functions and semantic roles, so semantic role annotation arises as a fundamental task of the project. In fact, we can see the remaining semantic information provided by ADESSE (i.e. verbs senses separation and semantic classification) as the necessary steps to deal with the identification and annotation of semantic roles.

Delimiting a useful close list of semantic roles is a complex work which has been dealt with in many occasions and with really different results. This discrepancy may respond basically to the degree of generalization which has been adopted. At one end, each particular verb "defines a distinct set of participant roles that reflect its own unique semantic properties (e.g. the subject of *bite* is a slightly different kind of agent from the subject of *chew*)." (Langacker, 1991: 284). At the other end, some proposals appeal to maximally generalized semantic roles, such as the macroroles of Actor and Undergoer (Van Valin & LaPolla, 1997) or the thematic protoroles of Agent and Patient (Dowty, 1991). On an intermediate level, we find usual labels like agent, patient, instrument, beneficiary, location, theme, …

Regarding semantic roles in ADESSE, the strategy of annotation lies in the consideration of different levels. First of all, each verbal class is associated with a set of semantic roles, which are prototypical for the cognitive domain evoked by the verbs belonging to it. Some of the labels used to denotate these sets may fit with traditional thematic roles as suggested above. Nevertheless, role labels linked to semantic classes were chosen in ADESSE by aiming at two factors: specificity (depending on the verbal class) and transparency (i.e. descriptive adequation). Examples of role labels used in ADESSE are shown in the following Table:

Class	0	1	2	-
Feeling		Senser	Stimulus	
Perception	Causer	Perceiver	Perceived	
Cognition	Causer	Cognizer	Content	
Possession		Possessor	Possessed	
Transfer	Ini-poss	Final-poss	Possessed	
Localization	Causer	Theme		Locative
Change	Agent	Patient		
Communic.		Sayer	Message	Receiver
Existential	Causer	Existent		

Table 7: Some class-specific roles in ADESSE

Secondly, each verb sense is also associated with a set of semantic roles which allows to describe the whole range of participants selected by that verb (i.e. its valency potential, as described above). Generally, the set of labels used for verb-specific roles are inherited by default from the roles associated with the semantic class to which that verb belongs. Put simply, labels for verb-specific roles are directly taken from class-specific roles.

For example, in order to describe the valency potential of *dotar* 'provide', we need to take into account three participants (therefore, three role labels): 'the provider', 'the person who is provided with something' and 'the thing provided'. Because of classifying *dotar* as a verb of 'Transfer', this verb inherits the set of role labels associated with that semantic class, namely: 'initial possessor', 'final possessor', 'possessed':

	A0	A1	A2
TRANSFER	**Ini-poss**	**Final-poss**	**Possessed**
Dotar 'provide'	Provider	Receiver	Thing

Table 8: Semantic roles labels for *dotar*

Arguments receive a correlative number, as in PropBank (Palmer et al. 2005). Nevertheless, PropBank generally applies Arg0 to the subject of transitive and unergative verbs, as in table 9:

	Arg0	**Arg1**	**Arg2**
KNOW 'understand'	*Knower*	*Thing known or thought*	*attributive*
LEARN 'learn'	*student*	*subject*	*teacher*
TEACH '(try to) make learn"	*Teacher*	*subject*	*student(s)*

Table 9: Arguments of *know, learn and teach* in PropBank

In ADESSE, we have reserved A0 for the first argument of causatives, so that we can see more easily the correspondences between causatives and their non-causative counterpart.

	A0	**A1**	**A2**
SABER 'know'		*Knower* [Cognizer]	*Thing known* [Content]
APRENDER 'learn'		*Learner* [Cognizer]	*Subject* [Content]
ENSEÑAR 'teach'	*Teacher* [Initiator]	*Learner* [Cognizer]	*Subject* [Content]

Table 10: Arguments in ADESSE

Additionally, we consider a small group of semantic roles which are generally independent of verb classes (causer, beneficiary, purpose, manner…). 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.

For example, the verb *barrer* 'sweep' belongs to Modification and inherits from this class the role labels of 'agent' and 'affected'. On the other side, the verb *sentir* 'feel' belongs to the class of Feeling and, therefore, inherits a 'senser' and an 'stimulus'. However, in order to annotate examples like (11) we must consider additional arguments, with roles like beneficiary, manner or locative, which are not directly associated with the semantic domain evoked by the verb:

(11) *Ni para barrerme la casa sirve* [CAI:025,21]
'He serves neither to sweep the house [for me]'

	A0	A1	AG
MODIFIC.	**Agent**	**Affected**	
Barrer	Agent	Affected	Beneficiary

Table 11: Semantic roles labels for *barrer*

We have seen class-specific roles, verb-specific roles and inheritance relationships between them. Additionally, we have illustrated cases where another kind of roles, general roles, must be considered. Finally, we have to deal with a last level which is relevant in the process of annotation of semantic roles in ADESSE: the level of syntactic-semantic schemas or valency realizations (as defined above).

Once the valency potential of each verb sense has been described and class-specific roles labels have been inherited, it only remains to annotate each syntactic pattern recorded for that verb in the corpus. This process is done by means of pointing for each syntactic argument the corresponding semantic role previously defined for the verb entry (see Tables 2 and 3 above).

This strategy has an obvious advantage. Given that each clause in the corpus is linked to a syntactic-semantic pattern, we do not need to apply the annotation of semantic roles to the 159.000 clauses but only to the 12.500 schemas, assuming that all the clauses linked to a given scheme will inherit the semantic information applied to the last one.

So, for example, the annotation of the approximately 2.000 clauses recorded for the verb *ver* 'see' in the basic transitive pattern (Subj-DObj) is done by annotating directly the correspondences between the semantic roles and the syntactic arguments of the transitive pattern itself, namely: Subj = A1 Perceiver / DObj = A2 Perceived. Further information is added in order to point out figurative senses and other specific comments for each example.

8. Current state and future work

Broadly speaking, the core of the information which has been accounted for in this work is currently done, although the permanent review we undertake on the data lead us to treat them as provisionally completed.

Other relevant goals of the project are at this time in progress. This is the case of the lexicographic task, that is, the definition and hierarchical organization of verb (micro) sense. Moreover, we are currently performing in ADESSE the annotation and treatment of lexical instantiations of arguments for the study of Verb+N combinations and support verbs. Obviously, the annotation of lexical realizations of each argument in the corpus will expand the search options in ADESSE. Besides the frequencies about verb meaning and construction meaning, the database will provide aspects of lexical combinatory for each verb, that is, frequencies about the type of entities with which a verb appears. In fact, this information is already available for 125.000 verb arguments in the corpus (a 70 % of the total, roughly)

Immediate future work will include the definition and annotation of semantic relations between verb senses (synonymy, hyponymy, antonymy…) and information about the argument structure of deverbal nouns.

Regarding the first goal, our idea is to establish different semantic relationships (synsets), not only between different verbs (e.g. *besar* 'kiss' or *acariciar* 'caress' as hyponyms of *tocar* 'touch'), but above all between different related senses of the same verb (e.g. the hyponymic relationship between the meanings of *beber* 'drink' as 'take liquid' and 'take alcoholic liquid).

Regarding the second goal -and taking into account the annotation already available about argument structure of

verbs-, we have in mind to expand that information to deverbal nouns in the corpus. That is, besides the set of arguments recorded for *destruir* 'destroy, *sentir* 'feel' or *regresar* 'return (v)', we will get equivalent information for *destrucción* 'destruction', *sentimiento* 'feeling' and *regreso* 'return (n)', to name some examples.

All the information provided by ADESSE can be freely consulted at http://adesse.uvigo.es/data/. The website offers reports by verb, by syntactic pattern, by semantic class. Moreover, it is possible to perform advanced searches (e.g. specifying different features of one or more arguments, like category, lexical item, semantic role…).

It must be pointed out that ADESSE is currently not available for purposes other than their use as a reference tool. Because of copyright restrictions the corpus can be consulted but not downloaded. In the immediate future a verbal lexicon will be freely downloadable from the website itself.

9. Conclusions

ADESSE is, above all, an online database for the empirical study of the interaction between verbs and constructions in Spanish. Through the browser provided by the website, we can get information on various aspects about argument structure in Spanish: e.g. constructional alternatives for a verb, a syntactic function or a semantic role (with frequencies in the corpus), verbs and syntactic constructions for a semantic domain, verbs and semantic domains for a particular construction …

However, we have seen that ADESSE is even more than that. Additionally, it allows the search and study of multiple correlations between syntactic and semantic features (case, person, number, definiteness, tense, mood, …). At present, the database is also being enriched with lexical and lexicographic information, which besides the future annotation of semantic relationships and argument structure of deverbal nouns, will considerably increase the possibilities of this linguistic resource.

Taking all of this in mind, we think that ADESSE is a useful corpus-based database for descriptive studies on Spanish. Ultimately, it represents a response to the current need for annotated corpora including detailed syntactic and semantic annotation.

10. Acknowledgements

The devolopment of ADESSE has been supported by the projects ADESSE (BFF2002-01197), ADESSE-II (HUM2005-01573) and ALEXSYS (FFI2008-01953) from the Spanish Ministry of Science and Innovation. Besides the authors of this article, the research team of ADESSE is currently made up of Inmaculada Anaya, Ana Caíño, Vanessa Dacosta, María Gómez, Amelia Huzum, María del Carmen Méndez, and Antonio Rifón. Significant contributions to the development of ADESSE were made also by former team members Fran Albertuz, Susana Comesaña, Lourdes Costas, Iago Crespo, and Susana Martínez.

11. References

Ágel, V. (1995), "Valenzrealisierung, Grammatik und Valenz", *Zeitschrift für germanistische Linguistik*, 23, 2-32

Albertuz, F. (2007). Sintaxis, semántica y clases de verbos. Clasificación verbal en el proyecto ADESSE. In P. Cano López (Coord.), *Actas del VI Congreso de Lingüística General*, Santiago de Compostela, May 3-7 of 2004, Vol. 2-2., pp- 2015-2030.

Dowty, D. (1991). Thematic Proto-roles and Argument Selection, *Language*, Vol. 67, nº 3, pp. 547-619.

Fillmore Ch. J., Johnson Ch., Petruck M. 2003. "Background to FrameNet", *International Journal of Lexicography* 16/3, pp. 235-250

García-Miguel, J.M. & F. Albertuz (2005). Verbs, semantic classes and semantic roles in the ADESSE project. In K. Erk, A. Melinger & S. Schulte im Walde (eds.), *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of verb Features and Verb Classes*. Saarbrüken, February 28-March 1.

García-Miguel, J.M., L. Costas & S. Martínez (2005). Diátesis verbales y esquemas construccionales. Verbos, clases semánticas y esquemas sintáctico-semánticos en el proyecto ADESSE. In G. Wotjak & J. Cuartero Otal (eds.), *Entre semántica léxica, teoría del léxico y sintaxis.* Frankfurt am Main: Peter Lang, pp. 373--384.

Halliday, M.A.K. (1985). *An Introduction to Functional Grammar*, London: Edward Arnold, 2004, 3th ed.

Hanks, Patrick. 1996. Contextual dependency and lexical sets. *International Journal of Corpus Linguistics* 1/1: 75–98.

Langacker, R. (1991). *Foundations of Cognitive Grammar, Vol. II. Descriptive Application*, Stanford: Stanford University Press.

Palmer, M.; P. Kingsbury; D. Gildea (2005): The Proposition Bank: An Annotated Corpus of Semantic Roles, *Computational Linguistics*, 31:1 , pp. 71-105

Rojo, G. (2001): La explotación de la Base de Datos Sintácticos del español actual. In De Kock, J. (ed.): *Lingüística con corpus*, Salamanca: Universidad de Salamanca.

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

Taulé, M., M.A. Martí, M. Recasens (2008) Ancora: Multilevel Annotated Corpora for Catalan and Spanish. *Proceedings of 6th International Conference on Language Resources and Evaluation*. Marrakesh (Morocco).

Van Valin, R. D & R. J. LaPolla (1997): *Syntax. Structure, meaning and function*. Cambridge: Cambridge University Press.

Vázquez, G., L. Alonso, J.A. Capilla, I. Castellón, A. Fernández (2006). "SenSem: sentidos verbales, semántica oracional y anotación de corpus", *Procesamiento del Lenguaje Natural*, 37, pp. 113-120.