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/

Al levantarse, Julián sintió un zumbido en los oídos [JOV:63,41] “When he stood up, Julián felt a ringing in his ears”			

(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).

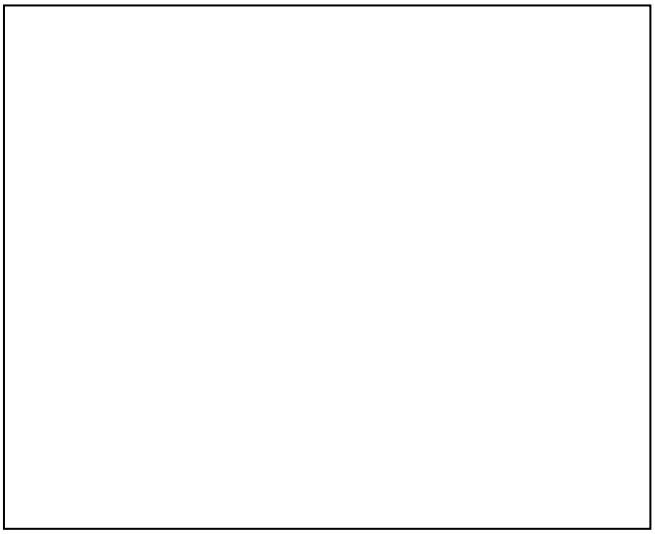
4. Basic strategies: valency potential and valency realizations

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'			



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.

Level 1	Level 2	Level 3	Level 4

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**

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/