AnCora-Verb: Two Large-scale Verbal Lexicons for Catalan and Spanish

Juan Aparicio
Mariona Taulé
M. Antònia Martí
Universitat de Barcelona

In this paper, AnCora-Verb is presented: two large-scale verbal lexicons used for the semantic annotation with arguments, thematic roles and semantic class of AnCora corpora (AnCora-Cat for Catalan and AnCora-Esp for Spanish). Each corpus contains 500,000 words with a multilayer annotation in different linguistic fields-from morphology to pragmatics. AnCora-Verb lexicons focuses on syntactic functions, arguments and thematic roles of each verbal predicate taking into account the verbal semantic class and those alternations in diathesis where the predicate can participate. This paper concentrates on the definition and characterization of verb classes and the criteria followed in the assignment of a verb to a specific class.

1. Introduction

It is widely admitted that either lexicons or tagged corpora appear to be a very useful resource for computational and linguistic analysis of languages. Following this line of thought, we present two verbal lexicons, AnCora-Verb-Ca for Catalan and AnCora-Verb-Es for Spanish, which are the basis for the semantic annotation with arguments and thematic roles of AnCora corpora. AnCora (Martí et al. 2007) is currently the largest multilevel annotated corpus of Spanish and Catalan. It consists of two corpora which contain 500,000 words each mostly from newspaper articles. AnCora has been annotated with morphological (PoS), syntactic (constituents and functions) and semantic (argument structure, thematic roles, named entities and *WordNet* senses) information.

In AnCora-Verb lexicons, it is established the mapping between syntactic functions, arguments and thematic roles of each verbal predicate taking into account the verbal semantic class and the diatheses alternations in which the predicate can participate. Each verbal predicate is related to one or more semantic classes. The main goal of this paper is focused on the definition and characterization of verbal classes and the criteria that have been followed for the assignment of a verb to a specific class.

In order to verify the adjustment of this classification two tests have been carried out:

1. The application of tests in expert linguists agreement.
2. The annotation of corpus, which has allowed us to verify the consistency and adjustment of the predicates description.

This paper is organized as follows. In section 2 the theoretical basis for the characterization of verbal predicates are exposed, section 3 introduces AnCora-Verb lexicons, section 4 presents the lexical semantic classes and, finally, section 5, contains our conclusions and focuses on further works.

2. Theoretical basis for the characterization of verbal predicates

The semantic properties used in the characterization of predicates are inspired in the proposal of lexical decomposition of Rappaport-Hovav & Levin (1998) from which the concept of LSS has been taken. LSS as well as the kind of diatheses alternations in which the predicate can participate, determines the number of arguments that a verbal predicate requires and the thematic role of these arguments. In this direction, we follow the lines laid down by Kipper

et al. (2000) and Kingsbury et al., (2002) in the construction of *VerbNet*. For the characterization of the argument structure, we follow *PropBank* annotation system (Palmer et al. 2005), and as far as the diatheses alternations, we follow Vázquez et al. (2000), from which we adopt its diatheses classification.

2.1. *Lexical Semantic Structures*

For the semantic characterization of predicates, our starting points are the four types of LSS, which correspond with the four ontological types of events: (a) accomplishments, (b) achievements, (c) states and (d) activities (Vendler 1967, Dowty 1979):

(a) [*x* CAUSE [BECOME [*y* <*STATE/THING/PLACE*>]]]
(b) [BECOME [*y* <*STATE /PLACE*>]]
(c) [*x* <*STATE* >]
(d) [x ACT <MANNER/INSTRUMENT>]

The lexical decomposition of a predicate in the form of a LSS contains three basic components: the semantic primitives, the constants and the variables. The semantic primitives correspond to the components CAUSE, BECOME and ACT, which determine both the basic meaning of the verb and the event type. The constants (*MANNER, INSTRUMENT, STATE*, etc.) express the idiosyncratic aspect of the verb meaning and are represented in italics. The variables (*x* and *y*) represent the arguments that the verb needs to be syntactically expressed.

2.2. *Argument structure and thematic roles*

The argument structure is determined by the LSS. The semantic relation that each argument maintains with the event denoted by the verb is defined by the thematic roles. For the arguments annotation we have followed the proposal of *PropBank* (Palmer et al., 2005), where the arguments required by the verb are incrementally numbered (ArgA[1], Arg0, Arg1, Arg2, Arg3, Arg4), expressing their degree of proximity in relation to its predicate. The adjuncts are labelled as ArgM.

2.3. *Diatheses alternations*

The diatheses are understood as the syntactic expression of a semantic opposition. Each diathesis focuses specific components of the LSS, such as the causative alternation (1), which focuses the semantic primitive CAUSE, or the inchoative alternation (2), which focuses the primitive BECOME.

(1) [*x* **CAUSE** [BECOME [*y* <*BROKEN*>]]]

The wind-*SUB* broke the window-*DO*

(2) [**BECOME** [*y* <*BROKEN*>]]

The window-*SUB* broke

Furthermore, the expression of most alternations entails an aspectual change, which necessarily implies a change of semantic class. Next we consider, as an example, how a verb moves towards different semantic classes, such as the verb *nadar* 'to swim' when it appears in the extension object alternation:

(3) Maria nadaba / estaba nadando

"*Maria swam / was swimming*"

(4) Maria nadaba / estaba nadando los 100 metros libres

"*María swam / was swimming the 100 meters freestyle*"

The verbs denoting activities as well as those denoting accomplishments accept the periphrasis *to be* + gerund, since they express events that progress in time. However, activities have already happened, even if they are interrupted, whereas accomplishments have to reach an inherent

[1] ArgA is used in *PropBank* to indicate the inductive agent, as in the sentence *Juan paseó a su perro* "Juan walked his dog", where *Juan* is the inductive agent (ArgA) and *his dog* is the agent who takes a walk (Arg0).

point to have happened. As it is observed in (3) *to swim* is an event that happens although it does not have an end point: *Maria has swum* even though the action could be interrupted. However, in (4) the verb *to swim* expresses an event that has a culmination point, that is, the denoted event must be finished to happen. If *Maria swam / was swimming the 100 meters freestyle* and the event is interrupted, it implies that *Maria has not swum the 100 meters freestyle*. We can say, as a conclusion, that in this alternating pair, the predicate *to swim,* in the intransitive construction (3), expresses an activity, whereas in the transitive construction (4), it expresses an accomplishment. As we have just seen, the common property to most diatheses alternations is that verbs belonging to a determined semantic class can move towards other semantic classes, under certain syntactic conditions.

In this proposal of classification, we have only considered productive diatheses, such as causative/inchoative, inchoative/causative, active/passive, resultative, oblique subject, transitive/intransitive, object extension, cognate object, and beneficiary alternation[2].

3. AnCora-Verb Lexicons

AnCora-Verb-Ca and AnCora-Verb-Es lexicons were obtained deriving, for each sense of each verb, all the syntactic schemata in which a verbal predicate appears in AnCora corpora. From this information, the mapping from syntactic functions to thematic roles, and the corresponding argument position, were manually declared in the lexicons.[3]

In AnCora-Verb lexicons, each predicate is related to one or more semantic classes (LSS), depending on its senses, basically differentiated according to the four event classes, and on the diatheses alternations in which a verb can occur. Figure 1 shows the full information associated with the entry *reforzar* "to reinforce" in the AnCora-Verb-Es lexicon.

As it can be seen, the lexical entry of figure 1 includes information about the lemma (*reforzar* "to reinforce"), the different senses associated to their corresponding LSS (in this case LSS1.1 and LSS2.2), the mapping between syntactic function and thematic role (for instance, SUJ Arg0##CAU), and the diatheses alternations in which the verb occurs (in this case, ANTICAUSATIVA "inchoative"). Examples are also included: *la operación refuerza su liderazgo* "the operation reinforces its leadership".

```
reforzar - 01
LSS1.1
SUJ Arg0##CAU
CD      Arg1##TEM
CC      ArgM#por#ADV

EJ:     "sirve para reforzarla por vía de absurdo"
EJ:     "la operación refuerza su liderazgo"
EJ:     "La subida en dos décimas de la tasa de paro reforzó la tendencia al alza"
+ANTICAUSATIVA
LSS2.2
SUJ     Arg1##TEM
CC      ArgM##ADV
EJ:     "Si dos neuronas se activan, sus conexiones se refuerzan"
```

Figure 1: Lexical entry of *reforzar* 'to reinforce' in AnCora-Verb-Es

[2]The specific alternations that are shared by few verbs have not been considered because they do not define general verb classes.

[3] The list of thematic roles consists of 20 different thematic labels: AGT (Agent), AGI (Induced Agent), CAU (Cause), EXP (Experiencer), SCR (Source), PAT (Patient), TEM (Theme), ATR (Attribute), BEN (Beneficiary), EXT (Extension), INS (Instrument), LOC (Locative), TMP (Time), MNR (Manner), ORI (Origin), DES (Goal), FIN (Purpose), EIN (Initial State), EFI (Final State) and ADV (Adverbial).

In order to guarantee the coherence and consistency of the data, the building of AnCora-Verb lexicons was carried out in three consecutive stages. The aims of the first stage were to define the basic criteria and to test agreement among expert linguists for a sample of 470 predicates. In a second stage, the rest of the predicates were analyzed and in a third stage, we proceeded to the readjustment of the classification.

The analysis of the first 470 verbal predicates[4] was carried out in two phases. First, five linguists considered 70 verbs of upper-middle frequency (between 4 and 40 occurrences) and analyzed them in parallel. Secondly, four linguists proceeded to the analysis of 400 predicates in parallel, which reached an agreement next to 95%.[5] The remaining 5% of disagreement was related to the differentiation of senses and to the identification of the verbal form, for example, when it had to be decided whether a certain structure corresponded to a verb and its complements or to an idiom (*dar* + *un susto* 'to give + a fright' vs *dar_un_susto* "to give_a_fright").

In the second stage, the rest of entries that compose the verbal lexicon AnCora-Verb-Es (1.495 verbs) were analyzed in an independent way. The problematic cases were discussed in weekly meetings. In the third stage, we proceeded to the readjustment of the semantic classes giving as result the final version, which is presented in this paper.

The semantic classes used to characterize the verbal predicates are organized in a hierarchy of two levels. In the first level, it is expressed the information about the LSS that is directly related to the event structure. In a second level, it is expressed the information related to the argument structure and the thematic roles. Thus, the association of a determined semantic class to a verb allows us to infer its argument structure and thematic roles.

The Spanish lexicon, AnCora-Verb-Es, contains a total of 1965 verbs corresponding to 3,671 senses and the Catalan lexicon, AnCora-Verb-Ca, contains 2,151 verbs and 4,513 senses. These figures correspond to the total amount of verbal predicates, which appears in AnCora-Es and AnCora-Ca corpora (500,000 words each).

Currently, a consulting interface for these lexicons is available at *http://clic.ub.edu/ancora*.

4. Spanish and Catalan Semantic Classes

Now we present the 13 semantic classes that we have compiled. They are grouped around the four main LSS types: accomplishments (A), achievements (B), states (C) and activities (D). These general classes are further split into subclasses, depending on argument structure, thematic roles and diatheses alternations. These classes correspond with general syntactic distinctions, those of the class (A) are transitive and ditransitive predicates and those of the class (B) and (D) are intransitive predicates.

Accomplishments (A)

LSS1.1 [*x* CAUSE [BECOME [*y* <*STATE/PLACE*>]]]
(causative accomplishment)
LSS1.2 [[*x* DO-SOMETHING] CAUSE [BECOME [*y* <*STATE/PLACE*>]]]
(agentive accomplishment)

The LSS corresponding to an accomplishment (LSS1.1, LSS1.2) are composed by the combination of the semantic predicates CAUSE, DO and BECOME. Accomplishments are basically associated with a complex event structure, which involves a causing subevent and a change of state or location subevent. Each argument of the verb is associated with a distinct subevent. The semantic roles assigned to each argument can be identified with the particular argument position associated to the semantic predicates composing the LSS and to the diatheses alternations in which a verb participates. Thus, accomplishments give rise to three verbal classes:

[4] The 470 selected verbs correspond with 4,585 occurrences.

[5] The agreement index was obtained confronting the results of each one of the linguists in the assignment of the LSS, of the argument structure and of the thematic roles for each one of the predicates.

Transitive-causative verbs (A1 class), such as *romper* "to break", associate the causer argument (x) with the semantic predicate CAUSE and the participant that undergoes the change with the argument (*y*). Since these verbs participate in the inchoative and resultative alternations, *x* is referred to as *Arg0-Cause* and *y* as *Arg1-Theme*.

Unlike causative verbs, transitive-agentive verbs (A2 class), such as *escribir* "to write", associate the causer argument (*x*) with the semantic predicate DO, and since they allow the passive alternation, the argument *x* is referred to as *Arg0-Agent* and the argument *y* as *Arg1-Patient.*

Finally, ditransitive-agentive verbs (A3 class) involve three participants in the LSS: an acting agent (*x*) does something that causes (*y*) to become in another location or spatial configuration (*z*). Since these verbs allow the passive alternation, *x* is represented as *Arg0-Agent* and *y* as *Arg1-Patient.* The third argument involved in the event (*z*) can appear as an *Arg2-Locative* in verbs such as *poner* "to put" or as an *Arg2-Beneficiary* in verbs such as *enviar* "to send", giving raise to the locative (A3.1 subclass) and to the beneficiary ditransitive-agentive verbs (A3.2 subclass).

A1: transitive-causative class

LSS1.1 [*x* CAUSE [BECOME [*y* <*STATE* >]]]
Arg0##CAU
Arg1##TEM
Diatheses: [+Inchoative] [+Resultative]
Spanish verbs: *romper* "to break", *abrir* "to open", *cerrar* "to close", *hundir* "to sink"…
Catalan verbs: *afectar* "to affect", *convertir* "to turn into", *omplir* "to fill"...

A2: transitive-agentive class

LSS1.2 [[*x* DO-SOMETHING] CAUSE [BECOME [*y* <*STATE*>]]]
Arg0##AGT
Arg1##PAT
Diatheses: [−Inchoative] [+/−Resultative] [+Passive] [+/−Beneficiary] [+/−Intransitive]
Spanish verbs: *escribir* "to write", *barrer* "to sweep", *leer* "to read", *visitar* "to visit"…
Catalan verbs: *afirmar* "to affirm", *decidir* "to decide", *guanyar* "to win"…

A3.1: ditransitive-agentive locative class

LSS1.3.1 [[*x* DO-SOMETHING] CAUSE [BECOME [*y* <*PLACE*> *z*]]]
Arg0##AGT
Arg1##PAT
Arg2##LOC
Diatheses: [−Inchoative] [+/−Resultative] [+Passive] [+/−Oblique subject]
Spanish verbs: *poner* "to put", *almacenar* "to store", *publicar* "to issue"…
Catalan verbs: *incorporar* "to include", *moure* "to move", *trasladar* "to transfer"…

A3.2: ditransitive-agentive beneficiary class

LSS1.3.2 [[*x* DO-SOMETHING] CAUSE [BECOME [*y* <*PLACE*> *z*]]]
Arg0##AGT
Arg1##PAT
Arg2##BEN
Diatheses: [-Inchoative] [+/-Resultative] [+Passive]
Spanish verbs: *enviar* "to send", *dar* "to give", *decir* "to say", *robar* "to rob"…
Catalan verbs: *explicar* "to explain", *permetre* "to allow", *vendre* "to sell"…

Achievements (B)

Achievements are associated with a specified resulting state or location. Unlike accomplishments, achievements have a simple event structure (LSS2), which lacks the causing subevent that characterizes accomplishments.

LSS2 [BECOME [*y* <*STATE/PLACE*>]]

Achievements are related to unaccusative verbs, a set of verbs that in terms of their LSS are basically monadic and in terms of their argument structure take a single internal argument (*Arg1*).

The representation of the argument structure allows some distinctions to be made between the internal and the external argument of a verb. Internal arguments are expressed in the syntax projected inside the verb phrase (VP), whereas, external arguments are expressed external to the VP headed by the verb selecting those arguments. Unacccusativity is related to the fact that the grammatical subject of an unaccusative verb behaves as the direct object of a transitive verb, consequently, the subject of an accusative verb and the object of a transitive verb bear the same semantic role: *Theme*, and occasionally *Patient*.

Unaccusative verbs can be split into two classes, depending on the constant associated with the LSS: unaccusative motion verbs (B1 class), such as *llegar* "to arrive", that are associated with the constant *PLACE* and the unaccusative state verbs (B2 class), such as *crecer,* "to grow", that are associated with the constant *STATE*.

B1: unaccusative-motion class

LSS2.1 [BECOME [*y* <*PLACE*>]]
Arg1##TEM/PAT
Diatheses: [−Passive]
Spanish verbs: *llegar* "to arrive", *ir* "to go" *salir* "to go_out", *venir* "to come"…
Catalan verbs: *baixar* "to go_down", *caure* "to fall", *entrar* "to go_in", *pujar* "to go_up"...

B2: unaccusative-state class

LSS2.2 [BECOME [*y* <*STATE*>]]
Arg1##TEM/PAT
Arg2##EFI
Diatheses: [−Passive] [+Causative]
Spanish verbs: *crecer* "to grow"*, florecer* "to bloom"…
Catalan verbs: *créixer* "to grow", *trencar-se* "to get broken", *enfonsar-se* "to collapse"…

States (C)

The LSS corresponding to states (LSS3) denotes the stative notion of being in a state.

LSS3 [*x* <*STATE*>*y*]

In terms of their argument structure, stative verbs take two arguments. On the one hand, they take an internal argument (*Arg1*), which appears as syntactic subject bearing the semantic role *Theme.* On the other hand, they take an *Arg2,* of which thematic role gives rise to four verbal classes: existence verbs (C1 class), such as *estar* "to be" map *Arg2* into *Locative;* attributive verbs (C2 class), such as *ser* 'to be', map *Arg2* into *Attribute;* scalar verbs (C3 class), such as *pesar* "to weight", map *Arg2* into *Extension;* and beneficiary state verbs (C4 class), such as *gustar* "to like", which *Arg2* maps the thematic role *Beneficiary* or *Experiencer.*

C1: Existence-state class

LSS3.1 [*x* <*STATE*>*y*]
Arg1##TEM
Arg2##LOC
Diatheses: [−Passive]
Spanish verbs: *estar* "to be", *existir* "to exist"…
Catalan verbs: *haver-hi* "there_is/are", *existir* "to exist"…

C2: attributive-state class

LSS3.2 [*x* <*STATE*>*y*]
Arg1##TEM
Arg2##ATR
Diatheses: [−Passive]

Spanish verbs: *ser* "to be", *parecer* "to seem" *tener* "to have"…
Catalan verbs: *estar* "to be", *tenir* "to have"…

C3: scalar-state class

LSS3.3 [*x* <*STATE*>*y*]
Arg1##TEM
Arg2##EXT
Diatheses: [−Passive]
Spanish verbs: *medir* "to measure", *pesar* "to weight", *valer* "to cost"…
Catalan verbs: *costar* "to cost", *durar* "to last", pesar "to weight"…

C4: beneficiary-state class

LSS3.4 [*x* <*STATE*>*y*]
Arg1##TEM
Arg2##BEN/EXP
Diatheses: [−Passive]
Spanish verbs: *gustar* "to like", *doler* "to hurt"…
Catalan verbs: *agradar* "to like", preocupar "to worry"…

Activities (D)

Activities have a simple event structure:

LSS4 [x ACT <MANNER/INSTRUMENT>]

The semantic predicate ACT denotes an acting entity (*x*) that does something. Activities are related to inergative verbs, a set of verbs that in terms of their LSS are basically monadic and in terms of their argument structure take a single external argument (*Arg0*). Depending on the thematic role assigned to *Arg0*, activities give raise to three verbal classes: agentive-inergative verbs (D1 class), such as *correr* "to run", take an *Arg0-Agent;* experiencer-inergative verbs (D2 class), such as *dormir* "to sleep", take an *Arg0-Experiencer;* and, source-inergative verbs (D3 class), such as *sudar* "to sweat", that take an *Arg0-Source*.

D1: agentive-inergative class

LSS4.1 [x ACT <MANNER/INSTRUMENT >]
Arg0##AGT
Diatheses: [−Passive] [+/-Object Extension]
Spanish verbs: *correr* "to run", *caminar* "to walk", *nadar* "to swim"…
Catalan verbs: *jugar* "to play", *navegar* "to sail", *treballar* "to work"…

D2: experiencer-inergative class

LSS4.2 [x ACT <MANNER/INSTRUMENT >]
Arg0##EXP
Diatheses: [−Passive] [+/−Cognate Object]
Spanish verbs: *dormir* "to sleep", *soñar* "to dream"...
Catalan verbs: *dormir* "to sleep", *respirar* "to breath"…

D3: source-inergative class

LSS4.3 [x ACT <MANNER/INSTRUMENT >]
Arg0##SRC
Diatheses: [−Passive] [+/−Cognate Object]
Spanish verbs: *roncar* "to snore", *sudar* "to sweat"…
Catalan verbs: *cridar* "to shout", *plorar* "to cry"…

4. Conclusions and further work

In this paper we have presented the lexicons AnCora-Verb-Ca and AnCora-Verb-Es, focusing specially on the verbal semantic classes that determine the mapping between syntactic functions and thematic roles. We have set four main Lexical Semantic Structures (LSS): accomplishments, achievements, states and activities. For each LSS several subclasses are defined taking into account the argument structure, the thematic roles and the diathesis alternations that a predicate accepts. All this information is represented in the lexicons, where verbal predicates are semantically characterized. This lexicon is used for the semiautomatic semantic tagging of the AnCora corpora.

As future lines of research, we can consider the linking of AnCora lexicons with other lexical resources, such as *VerbNet, FrameNet* and *WordNet*. These lexical resources codify different type of linguistic knowledge and the creation of a common base that links all them together will allow that one benefit from the other.

We also intend to analyze and describe the semantic of verbs participating in complex predication, as it happens with verbal periphrasis and light verbs, and its effects on the event structure.

Acknowledgements

This research has been supported by the projects CESS-ECE (*Corpus Etiquetados sintáctica y semánticamente para el Español, Catalán y Euskera*) (HUM 2004-21127-E) PRAXEM (HUM2006-27378-E), Lang2World (TIN2006-15265-C06-06) and by the government of the Generalitat de Catalunya.

We would also like to thank you all the annotators: Oriol Borrega, Núria Bufí, Joan Castellví, Maria Jesús Díaz, Silvia García, Difda Monterde, Aina Peris, Lourdes Puiggròs, Marta Recasens, and Bàrbara Soriano Bautista.

References

Dowty, D. (1991). "Thematic proto-roles and argumental selection". *Language* 67 (3). 547-619.

Ruppenhofer, J.; Ellsworth, M.; Petruck, M.; Johnson, C.; Sheffczyk, J. (2006). *Framenet II: Extended Theory and Practice*. [on line]. Berkeley. *http://www.icsi.berkeley.edu/framenet* [Acces date: 2008].

Kingsbury, P.; Palmer, M.; Marcus, M. (2002). "Adding semantic annotation to Penn TreeBank". In *Proceedings of the 2002 Conference on Human Language Technology*. San Diego.

Kipper, K.; Dang, H. T.; Palmer, M. (2000). "Class-Based Construction of a Verb Lexicon". *AAAI-2000 Seventeenth National Conference on Artificial Intelligence*. Austin.

Martí, M. A.; Taulé, M.; Márquez, Ll.; Bertrán M. (2007). "Anotación semiautomática con Papeles Temáticos de los corpus CESS-ECE". In *Procesamiento del Lenguaje Natural-TIMM*. Alicante: SEPLN. 67-76.

Palmer, M.; Kingsbury, P.; Gildea, D. (2005). "The Proposition Bank: An Annotated Corpus of Semantic Roles". *Computational Linguistics* 21 (1). 71-105.

Rappaport Hovav, M.; Levin, B. (1998). "Building Verb Meanings". In Butt, M.; Geuder, W. (eds.). *The Projection of Arguments: Lexical and Compositional Factors*. Stanford: CSLI Publications. 97-134.

Vázquez, G.; Fernández, A.; Martí, M. A. (2000). *Clasificación verbal. Alternancias de diátesis*. Lleida: Edicions de la Universitat de Lleida.

Vendler, Z. (1967). *Linguistics in Philosophy*. New Cork: Cornell University Press.