AnCora 2.0: Argument Structure Guidelines for Catalan and Spanish

Working paper 4: TEXT-MESS 2.0 (Text-Knowledge 2.0)

Taulé, M., Martí, M.A., Borrega, O.
2011

FFI2009-06497-E/FILO TIN2006-15265-C06-06

Contents

1 Introduction

This manual presents the guidelines for the annotation of argument structure of verbal predicates and their semantic class of the Spanish and Catalan AnCora 2.0 corpora. The semantic annotation of verbal predicates implies the systematic mapping between syntax and semantics, basically expressed in the argument structure. This mapping ultimately motivates the semantic classes. In this proposal, each verbal predicate was assigned to a specific semantic class and every syntactic function was tagged with both arguments and thematic roles [Aparicio et al., 2008] and [Taulé et al., 2008]. The semantic properties used were defined assuming lexical decomposition [Levin and Rappaport-Hovav, 1995] and [Rappaport-Hovav and Levin, 1998] from which the concept of Lexical Semantic Structure (LSS) was taken. The 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 line, we followed the guides laid down by [Kipper et al., 2002] and [Kingsbury et al., 2002] in the construction of *VerbNet*.

We consider the proposal presented by Levin and Rappaport-Hovav to be appropriated for our work mainly for two reasons. First, because in their model converge lexical semantic, event and argument information and diathesis alternations. And, second, because similar works in corpus and computational linguistics have been carried out following this approach, such as *VerbNet*, a lexicon with lexical semantic, argument and diathesis information for English predicates. *VerbNet* follows Levin's semantic classification and adopts *PropBank* semantic annotation [Palmer et al., 2005].

We characterize predicates by means of a limited number of LSS and Event Structure Patterns, according to the four basic event classes: states, activities, accomplishments, and achievements [Vendler, 1967, Dowty, 1991]. These general classes can be split into subclasses, as we will see in section 4. Semantic roles are determined by the event class the predicate belongs to and by the type of diathesis alternation the predicate presents. Thus, not only thematic roles are assigned, but also predicates are characterized both from the aspectual and from the argumental perspective. In fact, the semantic classes determine the mapping between syntactic functions and semantic roles.

This information is currently being stored in the lexicon AnCora-Verb 2.0 [Aparicio et al., 2008] for both languages, which is, in practice, our annotation guide.

Section 2 deals with the definition of the four basic event classes. In section 3, we discuss the thematic roles adopted and in section 4 we talk about the LSS adopted and the possible diathesis alternations they can participate in. Section 5 explains in greater detail the annotation criteria followed, and illustrates with examples the composition of AnCora-Verb 2.0 lexicon and the annotation of AnCora 2.0 corpora.

2 Basic Semantic Event Classes

Four general semantic classes have been laid down which can be further subclassified depending on thematic roles and diatheses. In the definition of these main classes only Argument0 and Argument1 have been taken into consideration, because they are the basic arguments involved in the definition of predicate structures (they typically correspond to the syntactic Subject and Direct Object functions). These criteria give raise to a coarse grained classification which has been later on specified by splitting each general class into subclasses (LSS). This subclassification has not been developed in as much detail as the thematic role assignment, since, although it may be very useful, mapping thematic roles into syntactic functions is not the main goal of this methodology.

Following [Vendler, 1967] and [Dowty, 1991], we assume there are four ontological event classes: states, activities (or processes), accomplishments and achievements.

1. [x<STATE>]
2. [x ACT<MANNER/INSTRUMENT>y]
3. [x CAUSE[BECOME[y<STATE/THING/PLACE>]]]
4. [BECOME[y<STATE>]]

The general frame depicted by 1 corresponds to the ontological class *state*, with just one entity involved in the event, which focuses in the state. The frame in 2 corresponds to *activities* (or *processes*), and usually presents agentive subjects and passive objects, thus allowing passive constructions. The frame in 3 corresponds to *accomplishments* that refer to resulting states in external cause processes, and usually presents causative subjects and allows anti-causative constructions. Finally, the frame in 4 corresponds to *achievements* that refer to a resulting state in processes without external cause.[1]

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. As we will see, LSS determines the number of arguments a verbal predicate requires and the thematic role of these arguments, and restricts the set of all possible diatheses.

In order to finish this section, it must be pointed out that each one of these general semantic classes is defined also in terms of telicity and dynamicity. Telicity is the property of a verb or verb phrase that presents an event as being complete or completable in some sense. Dinamicity is the property of a verb or verb phrase that presents an event as implying activity in some sense. Thus, *states* are defined as [–dynamic] and [–telic]; *activities* are defined as [+dynamic] and [–telic]; *accomplishments* are defined as [+dynamic] and [+telic]; and, finally, *achievements* are defined as [–dynamic] and [+telic]. We have not brought under consideration the trait [± punctual].

3 Thematic Roles and Argument Structure

The semantic relation that each argument maintains with the event denoted by the verb is defined by the thematic roles. We have adopted a set of 20 different thematic roles, each one of them able to mapping to several syntactic functions and argument positions. A complete list of thematic roles with their corresponding label is shown in Table 1.

adverbial (*adv*), agent (*agt*), attribute (*atr*), beneficiary (*ben*), cause (*cau*), cotheme (*cot*), destination (*des*), final state (*efi*), initial state (*ein*), experiencer (*exp*), extension (*ext*), purpose (*fin*), instrument (*ins*), location (*loc*), manner (*mnr*), origin (*ori*), patient (*pat*), source (*src*), theme (*tem*), time (*tmp*)

Table 1: Thematic roles and labels in AnCora 2.0

For the arguments, we have followed the proposal of PropBank [Palmer et al., 2005], where the arguments required by the verb sense are incrementally numbered, expressing their degree of proximity in relation to its predicate. Thus, we distinguish between seven possible argumental slots: arg0, arg1, arg2, arg3, arg4, argM and argL. The first five tags are numbered from less to more obliqueness with respect to the verb. ArgM corresponds to adjuncts, and argL codes lexicalized complements of light verbs[2].

Argument structure is determined by LSS. Depending on event structure and diathesis alternations, arguments might appear in different syntactic positions, and thematic roles might appear in different argument slots. Appendix A shows all possible combinations of argument, thematic role and syntactic function, with examples in Spanish.

[1] In Catalan and Spanish there are two types of passive constructions: passives with auxiliary verb *ser* (to be) plus the main verb in participle form, and passives with the pronoun *se* (*Esta mañana han sido vendidos cinco libros – Esta mañana se han vendido cinco libros* 'Five books have been sold this morning'). The pronoun *se* may as well be used in anti-causative constructions (*La puerta se abrió* 'The door opened').

[2] *PropBank* uses one more tag (ArgA) to encode the inductive agent. We have decided to leave it out and, instead, resolve these cases syntactically.

4 Lexical Semantic Structures and Diathesis Alternations

LSS determines the number of arguments that a verbal predicate requires and the thematic role of these arguments, and describes the syntactic function of said arguments. In our model, each LSS restricts the set of all possible diatheses[3] it can incur in, and each verb sense is associated to one LSS. Diathesis alternations are the result of focusing different components of the LSS they belong to. That is to say, diatheses are surface structures that result from focusing different components of the predicate LSS. Diatheses must be understood as the syntactic expression of a semantic opposition.

Furthermore, the expression of most alternations entails an aspectual change, which necessarily implies a change of semantic class. As an example, let us consider the following sentences:

- Juan **corre** (*Inergative-agentive, activity*)
  - *Juan runs* - 1 argument
- Juan **corre** los cien metros (*Transitive-agentive, accomplishment*)
  - *Juan runs the 100-meters* - 2 arguments
- Se **corren** los 100 metros (*Unaccusative-state, achievement*)
  - *The 100-meters are run* - 2 arguments (1 is elliptic)
- Se **corre** los 100 metros (*Impersonal-state, achievement*)
  - *The 100-meters are run* - 1 argument
- **Hice correr** a Juan los 100 metros (*Transitive-causative, accomplishment*)
  - *I made Juan run the 100-meters* - 3 arguments
- Se le **hizo correr** los 100 metros (*Unaccusative-impersonal, achievement*)
  - *He was forced to run the 100-meters* - 2 arguments
- **Corridos** los 100 metros, Juan descansó (*Resultative-attributive, state*)
  - *Once the 100-meters were over, Juan relaxed* - 1 argument

All sentences in the list above denote the same event (Someone –*Juan*– ran the 100-meters, maybe forced by someone else –*me*, in the example- or by something). But each one of the sentences focuses in a different component of the basic event, thus entailing both aspectual changes (from inergativity to passivity, resultativity, etc.) and semantic class (we can see at least one sentence belonging to each one of the four main semantic event classes described in section 2). This example is very illustrative about the degree in which a verb belonging to a determined semantic class may move towards other semantic classes under certain syntactic conditions.

As we have already said, our proposal of classification is coarse grained. We have only considered productive diatheses. Specific alternations shared by few verbs have been left out because they do not define general classes. This improves general robustness and coherence.

Next, we will present the specific LSS derived from the general semantic event classes discussed above. These LSS are the result of combining the general class with the argument structure and the thematic roles that can fulfill each argument slot. Each verbal class is also characterized for admitting certain diatheses alternations. For each verbal sense, its semantic class is established, and the mapping between syntactic functions[4], argument structure and thematic roles is declared.

There are 24 LSS compiled and described, grouped around the 4 general event classes. These 24 LSS derive from the analysis of the 5079[5] verbs in AnCora 2.0 corpora. On the basis of a draft of the annotation guide,

[3] We follow in essence the diathesis classification of [Vàzquez et al., 2000]. It must be said, nevertheless, that causative and impersonal alternations are possible for all LSS, as explained in section 4.5.

[4] We extracted the verbal syntactic frames from the corpus as it has been described in [Taulé et al., 2005, Civit and Martí, 2005]

[5] 2830 verbs for AnCora-ES (Spanish) and 2249 verbs for AnCora-CA (Catalan).

annotator agreement tests have been carried out. In a first step, only Spanish verbs were taken into account. 70 verbs have been studied and tagged by five annotators in parallel, and in three phases (10, 30 and 30 verbs in each phase). After annotating each group an agreement discussion was carried out in order to revise and update the annotators guide. Once the guidelines were established, in a second step, 400 verbs were annotated by two pairs of annotators, each pair working in parallel with the same set of verbs. For these pairs of annotators the observed agreement rate was of 95% and 96%, respectively. This agreement rate has been obtained by confronting the results of the mapping between functions and thematic roles of one member of the pair against the other. The remaining 4% and 5% of disagreement has been discussed and the annotator guide modified when necessary. Almost all cases of disagreement are related to sense discrimination (assignment of LSS) and the identification of verbal forms, for instance, when it is necessary to decide if a given structure corresponds to a verb and its complements or to an idiom (*dar + un susto* vs. *dar_un_susto*, -to fright). In a second step, the analysis of the remaining Spanish verbs and of all Catalan verbs has been done by the annotators independently, with constant feedback among them, leaving out the analysis and annotation of Prepositional Objects due to their large variation in thematic role assignment. In this second step, a subset of 13 coarse-grained classes was used. Finally, in the third step, Prepositional Objects were analyzed and their thematic roles determined and incorporated to the annotation of both the Lexicon and the Corpora by two experienced annotators. This step yielded a further subdivision of previous classes, giving raise to the final 24 LSS currently existant.

4.1 LSS1 (A): accomplishments ([+dynamic],[+telic])

This general event structure is further subdivided into three subclasses: transitive-causative (A1), transitive-agentive, (A2) and ditransitive-agentive (A3), each one with a number of possible LSS. All verbs falling within this general event structure share the *resultative* and *passive* alternations, apart from *impersonal* and *causative*, which are common to all verbs (see footnote 3 in section 4).

LSS corresponding to accomplishments are composed by the combination of the semantic predicates CAUSE, DO and BECOME in a complex event structure involving a causing subevent and a change of state or location subevent.

4.1.1 LSS A1: transitive-causative

Transitive-causative verbs associate the external causer argument (x) with the semantic predicate CAUSE and the internal participant that undergoes the change with the argument (y). Since these verbs participate in the *anticausative* alternation, x is referred to as *arg0-cause* and y as *arg1-theme*. Arg0 is syntactically the subject, while arg1 is syntactically the direct object. The presence and nature of a second participant (arg2) defines the final classes A11, A12 and A13.

A11.transitive-causative
[x CAUSE[BECOME[y<STATE>]]]
Arg0=CAU
Arg1=TEM
Diatheses[6]: [+anticausative(B21)[7]] [+causative(=)] [+impersonal(=)][+resultative(C21)] [+/-passive(B22)] [+benefactive(=)]
Spanish verbs: *abrir* 'to open', *romper* 'to break', *cerrar* 'to close', *hundir* 'to sink', ...
Catalan verbs: *afectar* 'to affect', *espantar* 'to frighten', *activar* 'to activate', ...

Verbs in this class are defined by lacking a third argument and associating arg0 to a *cause* θ-role and arg1 to a *theme* θ-role. Broadly speaking, the change in state of the internal argument is lexically encoded in the verb stem.

A12.transitive-causative-state
[x CAUSE[BECOME[y<STATE>z]]]
Arg0=CAU

[6] We indicate the new class the verb will belong to when it undergoes the alternation between round brackets. An equal sign means no change in class.

[7] *Anticausative* alternation is also known as *ergative* or *inchoative* alternation.

Arg1=TEM
Arg2=EFI
Diatheses: [+anticausative(B21)] [+causative(A11)] [+impersonal(=)] [+resultative(C21)] [+/-passive(B12)] [+benefactive(=)]
Spanish verbs: *habituar* 'to get used', *disfrazar* 'to disguise', *promocionar* 'to promote', ...
Catalan verbs: *adaptar* 'to adapt', *incitar* 'to incite', *tenyir* 'to dye', ...

Verbs in this class are defined by the presence of a participant z, arg2, which bears the *final state* θ-role. Like all other verbs in this general A1 LSS, arg0 bears *cause* θ-role and arg1 bears *theme* θ-role.

A13.transitive-causative-instrumental
[x CAUSE[BECOME[y<STATE>*(with)* z]]]
Arg0=CAU
Arg1=TEM
Arg2=INS
Diatheses: [+anticausative(B21)] [+causative(A11)] [+impersonal(=)] [+resultative(C21)] [+/-passive(B12)] [+benefactive(=)]
Spanish verbs: *encharcar* 'to waterlog', *enguarrar* 'to get dirty', *plagar* 'to plague', ...
Catalan verbs: *inundar* 'to flood', *omplir* 'to fill', ...

This class is defined by the presence of a third argument, arg2, associated to the θ-role *instrument*. arg0 and arg1 remain the same as for the other two classes in A1 LSS. Broadly speaking, the change in state is lexically encoded in the verb stem.

4.1.2 LSS A2: transitive-agentive

Transitive-agentive verbs associate the external causer argument (x) with the semantic predicate DO and the internal participant that undergoes the change with the argument (y). Since they allow the passive alternation but not the inchoative one, participant x is referred to as *Arg0-agent*. There is always an internal argument (y), whose semantic and syntactic behavior determines de three subclasses available.

A21.transitive-agentive-patient
[[x DO-SOMETHING]CAUSE[BECOME[y<STATE>]]]
Arg0=AGT
Arg1=PAT
Diatheses: [+causative(A11)] [+benefactive(=)] [+/-passive(B22)] [+impersonal(=)] [+/-resultative(C21)] [+/-intransitive(D11)] [+oblique subject(=)]
Spanish verbs: *hacer* 'to do', *querer* 'to want', *ver* 'to see', ...
Catalan verbs: *voler* 'to want', *guanyar* 'to win', *saber* 'to know', ...

In this class, the internal argument y is assigned *patient* θ-role. Its syntactic function is always direct object. The external argument x is syntactically the subject, and is assigned *agent* θ-role. This is the most common LSS in both the corpora and the lexicon.

A22.transitive-agentive-theme
[[x DO-SOMETHING]CAUSE[BECOME[y<STATE>]]]
Arg0=AGT
Arg1=TEM
Diatheses: [+causative(A11)] [+benefactive(=)] [+passive(B22)] [+impersonal(=)] [+/-resultative(C21)] [+/-intransitive(D11)]
Spanish verbs: *participar* 'to participate', *insistir* 'to insist', *luchar* 'to fight', ...
Catalan verbs: *parlar* 'to talk', *col·laborar* 'to collaborate', *confiar* 'to trust', ...

The internal argument y in A22 class is assigned the *theme* θ-role. Its syntactic function is always prepositional object, and may be introduced by a variety of prepositions. As for the rest of the verbs in A2 class, the external argument x is syntactically de subject, and is assigned *agent* θ-role.

A23.transitive-agentive-extension
[[x DO-SOMETHING]CAUSE[BECOME[y<STATE>]]]
Arg0=AGT
Arg1=EXT
Diatheses: [+causative(A11)] [+benefactive(=)] [+passive(B22)] [+impersonal(=)] [+/-resultative(C21)] [+/-intransitive(D11)]
Spanish verbs: *recorrer* 'to cover (a distance)', *tardar* 'to take (time)', *correr* 'to run', ...
Catalan verbs: *trigar* 'to take (time)', *facturar* 'to invoice', *recórrer* 'to cover (a distance)', ...

The internal argument y in A22 class is assigned the *extension* θ-role. Its syntactic role may be either direct object or prepositional object. The external argument x is syntactically the Subject, and gets *agent* θ-role. Constituents marked with the *extension* θ-role refer to quantities and rarely allow for pasivization.

4.1.3 LSS A3: ditransitive-agentive

Ditransitive-agentive verbs associate the causer argument (x) with the semantic predicate DO and the participant which undergoes the change with the internal argument y. All verbs in this class have a third argument (z) associated to a PLACE in space (broadly speaking), or to a STATE. Since these verbs allow passive alternation, the argument x bears the *agent* θ-role. The syntactic and thematic nature of the two remaining participants further subdivides this general class into five subclasses.

A31.ditransitive-patient-locative
[[x DO-SOMETHING]CAUSE[BECOME[y<PLACE>z]]]
Arg0=AGT
Arg1=PAT
Arg2=LOC
Diatheses: [+causative(A11)] [+passive(B12)] [+oblique subject(=)] [+/-resultative(C21)] [+impersonal(=)] [+benefactive(=)]
Spanish verbs: *señalar* 'to point out', *colocar* 'to place', *registrar* 'to register', ...
Catalan verbs: *incloure* 'to include', *publicar* 'to publish', *acollir* 'to take in', ...

The internal argument y in A31 class is assigned *patient* θ-role and is always the direct object. The third argument might be syntactically a prepositional object or an (argumental) adjunct, and it is always assigned the *location* θ-role. The semantic interpretation of arg2 is, therefore, bounded to a physical location in space. The *oblique subject* diathesis is typical of this group of verbs.

A32.ditransitive-patient-benefactive
[[x DO-SOMETHING]CAUSE[BECOME[y<PLACE>z]]]
Arg0=AGT
Arg1=PAT
Arg2=BEN
Diatheses: [+causative(A11)] [+passive(B12)] [+/-resultative(C21)] [+impersonal(=)]
Spanish verbs: *decir* 'to say', *dar* 'to give', *explicar* 'to explain', ...
Catalan verbs: *demanar* 'to ask', *presentar* 'to introduce', *oferir* 'to offer', ...

The internal argument y in A32 class is assigned *patient* θ-role and is always the direct object. The third argument is syntactically the indirect object, and it is always assigned the *beneficiary* θ-role. The semantic interpretation of arg2 as a PLACE is therefore less literal than in A31 class: it must include traits such as [+human] or [+animate]. Most verbs of physical and verbal transfer belong to this class.

A33.ditransitive-theme-locative
[[x DO-SOMETHING]CAUSE[BECOME[y<PLACE>z]]]
Arg0=AGT
Arg1=TEM
Arg2=LOC
Diatheses: [+causative(A11)] [+impersonal(=)] [+resultative(C21)] [+oblique subject(=)] [+benefactive(=)] [+/-passive(B12)]

Spanish verbs: *competir* 'to compete', *implicar* 'to imply', *coincidir* 'to agree', ...
Catalan verbs: *informar* 'to inform', *fonamentar* 'to base', *competir* 'to compete', ...

The internal argument y in A33 class is assigned *theme* θ-role and is always the prepositional object. The third argument is syntactically an adjunct, and it is always assigned the *locative* θ-role. Again, as in A31, the semantic interpretation of arg2 as a PLACE is literal. This class has no direct object, syntactically. Only a small, restricted set of verbs belong to this class.

A34.ditransitive-patient-theme
[[x DO-SOMETHING]CAUSE[BECOME[y<STATE>z]]]
Arg0=AGT
Arg1=PAT
Arg2=TEM
Diatheses: [+causative(A11)] [+impersonal(=)] [+resultative(C21)] [+passive(B12)] [+oblique subject(=)] [+benefactive(=)]
Spanish verbs: *acusar* 'to accuse', *ayudar* 'to help', *condenar* 'to condemn', ...
Catalan verbs: *denunciar* 'to report', *obtenir* 'to obtain', *amenaçar* 'to threaten', ...

The internal argument y in A34 class is assigned *patient* θ-role and is always the direct object. The third argument is syntactically a prepositional object, and it is always assigned the *theme* θ-role. In this class, arg2 is no longer associated to a semantic interpretation bound to a locative. It is, broadly speaking, closer to depicting an amalgam of final state/instrument/comitative θ-roles.

A35.ditransitive-theme-cotheme
[[x DO-SOMETHING]CAUSE[BECOME[y<STATE>z]]]
Arg0=AGT
Arg1=TEM
Arg2=COT
Diatheses: [+causative(A11)] [+impersonal(=)] [+resultative(C21)] [+cotheme(B23)] [+passive(B12)] [+benefactive(=)]
Spanish verbs: *separar* 'to sepparate', *conectar* 'to connect', *aliar* 'to ally', ...
Catalan verbs: *unir* 'to join', *pactar* 'to agree on', *combinar* 'to combine', ...

The internal argument y in A35 class is assigned *theme* θ-role and is always the direct object. The third argument is syntactically a prepositional object, and it is always assigned the *cotheme* θ-role. In this class, arg2 should be interpreted close to a comitative.

4.2 LSS2 (B): achievements ([-dynamic],[+telic])

This general event structure is subdivided into two subclasses: *unaccusative-motion* (B1) and *unaccusative-state* (B2), depending on the constant they associate with (either PLACE or STATE). Unnacusative verbs are basically monadic in terms of their LSS and in terms of their argument structure, taking a single internal argument (Arg1). However, the presence of a second argumental constituent is usual.

Achievements are associated with a simple event structure which lacks the causing subevent that characterizes accomplishments.

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. Unaccusativity 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 unaccusative verb and the object of a transitive verb bear the same semantic role: *theme*, and *patient* for passives.

4.2.1 LSS B1: unaccusative-motion

Unaccusative-motion verbs associate the internal argument y with a change of location predicate. There are two subclasses of verbs within this LSS: *lexically* unaccusative verbs (especially, motion verbs such as *llegar* 'to arrive', or *venir* 'to come'), which belong to the B11 class, and verbs resulting from a passive diathesis, which belong to the B12 class. The main difference between them is the θ-role associated to the internal argument: *theme* in the former case and *patient* in the later.

B11.unaccusative-motion
[BECOME[y<PLACE>]]
Arg1=TEM
Arg2=LOC
Diatheses: [+impersonal(=)] [+causative(A11)] [+resultative(C21)]
Spanish verbs: *ir* 'to go', *salir* 'to exit', *parar* 'to stop', ...
Catalan verbs: *entrar* 'to go in', *arribar* 'to arrive', *caure* 'to fall down', ...

Arg1 is syntactically the subject and is associated to the *theme* θ-role. Arg2 may syntactically be an adjunct or a prepositional object and is associated to the *location* θ-role, or to the more specific *destination* or *origin*.

B12.unaccusative-passive-ditransitive
[BECOME[y<PLACE>]]
Arg1=PAT
Arg2=LOC/BEN/TEM
Arg0=AGT
Diatheses: *see below*
Spanish verbs: *preguntar* 'to ask', *consultar* 'to consult', *dirigir* 'to address/direct', ...
Catalan verbs: *traslladar* 'to carry', *donar* 'to give', *recollir* 'to pick up', ...

This LSS is specific for encapsulating the passivization of ditransitive verbs. All arguments and θ-roles are shared with its active counterpart, although there is a change both in the lexical structure (from *accomplishment* to *achievement*) and in its syntactic counterpart. The first internal argument (arg1) is the subject and keeps its *patient* θ-role, and the external argument (arg0) may appear as an oblique agent complement, keeping its original *agent* θ-role too. The second internal argument (arg2) encompasses the list of possibilities given in A3 class: it may be a prepositional object or an adjunct, as syntax goes, and it can be associated to the *beneficiary*, *location* or *theme* θ-roles.

Being itself the result of a diathesis, class B12 is considered to participate in no other alternations.

4.2.2 LSS B2: unaccusative-state

Unaccusative-state verbs associate the internal argument y with a change of state predicate. There are three subclasses of verbs within this LSS: *lexically* unaccusative verbs, which belong to the B21 class, verbs resulting from a passive diathesis, which belong to the B22 class, and verbs resulting from a cotheme diathesis, which belong to the B23 class. The main difference between them is the θ-role associated to the internal argument: *theme* in the first and last cases and *patient* in the second.

B21.unaccusative-state
[BECOME[y<STATE>]]
Arg1=TEM
Arg2=EFI
Diatheses: [+causative(A11)] [+impersonal(=)]
Spanish verbs: *producir* 'to cause/produce', *convertir* 'to convert/become', *pasar* 'to pass', ...
Catalan verbs: *morir* 'to die', *pujar* 'to climb/go up', *créixer* 'to grow', ...

Arg1 is syntactically the subject and is associated to the *theme* θ-role. Arg2 may syntactically be an adjunct, a prepositional object or a predicative complement and is associated to the *final state* θ-role or, alternatively, to the *attribute* or *initial state* θ-roles. Other roles are possible (especially *beneficiary* or *extension*) when the

frame comes from an anticausative alternation.

B22.unaccusative-passive-transitive
[BECOME[y<STATE>]]
Arg1=PAT
Arg0=AGT
Diatheses: *see below*
Spanish verbs: *ver* 'to see', *considerar* 'to consider', *utilizar* 'to use', ...
Catalan verbs: *fer* 'to do/make', *produir* 'to make/produce', *celebrar* 'to celebrate', ...

This LSS is specific for encapsulating the passivization of transitive verbs, as B12 did so with ditransitive ones. All arguments and θ-roles are shared with its active counterpart, although there is a change both in the lexical structure (from *accomplishment* to *achievement*) and in its syntactic counterpart. The internal argument (arg1) is the subject and keeps its *patient* θ-role, and the external argument (arg0) may appear as an oblique agent complement, keeping its original *agent* θ-role too.

Being itself the result of a diathesis, class B22 is considered to participate in no other alternations.

B23.unaccusative-cotheme
[BECOME[y<STATE>]]
Arg1=TEM
Arg2=COT
Diatheses: *see below*
Spanish verbs: *separar* 'to sepparate', *asociar* 'to associate', *desconectar* 'to disconnect', ...
Catalan verbs: *barallar* 'to fight', *unir* 'to join', *casar* 'to marry', ...

This is the last of the classes which result from diatheses. In this case, the class is formed by verbs from the A35 class which have undergone the cotheme alternation, and shares arguments and θ-roles with it: arg1 is the subject, with *theme* θ-role, and arg2 is a prepositional object with *cotheme* θ-role. It is not possible to have the external argument expressed in this LSS.

Being itself the result of a diathesis, class B23 is considered to participate in no other alternations.

4.3 LSS3 (C): states ([-dynamic],[-telic])

This general event structure is subdivided into four classes: *state-existential* (C1), *state-attributive* (C2), *state-scalar* (C3) and *state-benefactive* (C4). 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, whose thematic role gives rise to the four verbal classes and their subclasses.

4.3.1 LSS C1: state-existential

C11.state-existential
[x<PLACE>y]
Arg1=TEM
Arg2=LOC
Diatheses: [+impersonal(=)] [+causative(A11)] [+oblique subject(=)] [+cognate object(=)] [+resultative(C21)] [+benefactive(=)]
Spanish verbs: *haber* 'to be/have', *estar* 'to be', *existir* 'to exist', ...
Catalan verbs: *haver* 'to be/have', *començar* 'to start', *acabar* 'to end', ...

Arg1 is syntactically the subject. Arg2 maps to a locative and is syntactically an adjunct or a prepositional complement, if expressed at all. It is very common for verbs belonging to this class to show just one explicit argument. Auxiliary verb *haber* (Spanish) /*haver* (Catalan) in its use as main verb (existential 'to be' or 'to

have') belongs to this class.

4.3.2 LSS C2: state-attributive

C21.state-attributive
[x<STATE>y]
Arg1=TEM
Arg2=ATR
Diatheses: [+causative(A11)] [+impersonal(=)] [+benefactive(=)] [+/-resultative(=)]
Spanish verbs: *ser* 'to be', *tener* 'to have', *estar* 'to be', ...
Catalan verbs: *ser* 'to be', *tenir* 'to have', *estar* 'to be', ...

Arg1 is syntactically the subject, while arg2 is syntactically an attribute, typically, or a direct object, for such verbs as *tener* (Spanish) / *tenir* (Catalan) –'to have'. As for θ-roles, arg2 maps to an *attribute*. Copulative verbs such as *ser* when used as a main verb or *estar* in its most common sense belong to this class.

4.3.3 LSS C3: state-scalar

C31.state-scalar
[x<STATE>y]
Arg1=TEM
Arg2=EXT
Diatheses: [+causative(A11)] [+impersonal(=)]
Spanish verbs: *valer* 'to cost', *pesar* 'to weigh', *durar* 'to last', ...
Catalan verbs: *costar* 'to cost', *durar* 'to last', *tardar* 'to take (time)', ...

Arg1 is syntactically the subject while arg2 may either be a direct object, an adjunct or a prepositional object. Arg2 maps with *extension* θ-role, an argument referring to some sizable and measurable magnitude such as length, weight, time, price, etc. Notice that, even if verbs belonging to this class appear to be transitive –they may accept a direct object complement–, passive alternation is not possible.

4.3.4 LSS C4: state-benefactive

Though few in number, verbs belonging to C4 class are among the most frequent ones. The semantic nature of the second argument determines two subclasses. Just like in the rest of stative verbs, the internal argument (arg1) is associated to the *theme* θ-role.

C41.state-benefactive
[x<PLACE>y]
Arg1=TEM
Arg2=BEN
Diatheses: [+causative(A11)] [+impersonal(=)]
Spanish verbs: *servir* 'to be useful', *suceder* 'to happen', *ocurrir* 'to occur', ...
Catalan verbs: *caldre* 'to need', *tocar* 'to get/to be somebody's turn', *correspondre* 'to belong/correspond', ...

Arg1 is syntactically the subject, while arg2 is the indirect object and maps to a *beneficiary* θ-role.

C42.state-experiencer
[x<PLACE>y]
Arg1=TEM
Arg2=EXP
Diatheses: [+causative(A11)] [+impersonal(=)] [+/-resultative(C21)]
Spanish verbs: *parecer* 'to seem', *importar* 'to be important', *desagradar* 'to dislike', ...
Catalan verbs: *passar* 'to happen', *semblar* 'to seem', *agradar* 'to like', ...

Arg1 is syntactically the subject, while arg2 is the indirect object and maps to a *experiencer* θ-role.

4.4 LSS4 (D): activities ([+dynamic],[-telic])

This general event structure is subdivided into three subclasses: inergative-agentive (D11), inergative-experiencer (D21) and inergative-source (D31). As can be seen, 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, whose thematic role determines verb class. Inergative verbs denote an acting entity (x) that does something (ACT).

4.4.1 LSS D1: inergative-agentive

D11.inergative-agentive
[x ACT<MANNER/INSTRUMENT>]
Arg0=AGT
Diatheses: [+impersonal(=)] [+causative(A11)] [+object extension(A21)] [+resultative(C21)]
Spanish verbs: *actuar* 'to act', *regresar* 'to come back', *volar* 'to fly', ...
Catalan verbs: *treballar* 'to work', *anar* 'to go', *viatjar* 'to travel', ...

Arg0 is syntactically the subject, and its θ-role is *agent*. Verbs belonging to this class may be made transitive by adding a direct object, as the diatheses alternations point out.

4.4.2 LSS D2: inergative-experiencer

D21.inergative-experiencer
[x ACT<MANNER/INSTRUMENT>]
Arg0=EXP
Diatheses: [+causative(A11)] [+impersonal(=)] [+cognate object(A21)] [+resultative(C21)]
Spanish verbs: *dormir* 'to sleep', *vivir* 'to live', *oír* 'to hear', ...
Catalan verbs: *dormir* 'to sleep', *respirar* 'to breathe', *al·lucinar* 'to allucinate', ...

Arg0 is syntactically the subject, and its θ-role is *experiencer*. Verbs belonging to this class may be made transitive by adding a direct object whose semantic content is normally already present in the verb (like in *dream a dream*).

4.4.3 LSS D3: inergative-source

D31.inergative-source
[x ACT<MANNER/INSTRUMENT>]
Arg0=SRC
Diatheses: [+causative(A11)] [+impersonal(=)] [+cognate object(A21)] [+resultative(C21)]
Spanish verbs: *llorar* 'to cry', *toser* 'to cough', *sudar* 'to sweat', ...
Catalan verbs: *callar* 'to shut up', *riure* 'to laugh', *bordar* 'to bark', ...

Arg0 is syntactically the subject, and its θ-role is *source*. Verbs belonging to this class may be made transitive by adding a direct object whose semantic content is somehow already present in the verb (like in *cry tears*).

4.5 Special diatheses: impersonal and causative alternations

To end up this section, a word on these two special diathesis alternations must be made. It has already been said above that causative and impersonal alternations are possible for all LSS (see footnote 3 in section 4). Let us further explain this fact.

All verb classes (and almost all verbs) admit impersonal alternation by means of the addition of the pronoun *se*[8]. What this pronoun normally does is to reduce the valence of the verb in one argument. Thus, a verb having three arguments in its argument structure will have only two if accompanied by *se*. That is to say, the presence of this pronoun blocks the apparition either of the subject or of the direct object (typically). When blocking the apparition of the direct object, passive or anticausative alternations are found, and it only affects transitive verbs (LSS1), changing their class. When blocking the apparition of the subject (be it an internal or an external argument), any LSS may be affected, but there is no change in verb class. Let us present some examples.

- Cuando **se llega**$_{[B11]}$ tarde, no **se pide**$_{[A32]}$ explicaciones ni **se pone**$_{[A21]}$ excusas.
  - [When one is late one does not ask for explanations nor give excuses.]
- En el cine sólo **se llora**$_{[D31]}$ si **se es**$_{[C21]}$ sensible.
  - [One only cries in the cinema if one is sensitive.]

On the other hand, all verbs in Spanish and Catalan accept causative alternation by means of the addition of the light (semiauxiliary) verbs *hacer/fer* or *dejar/deixar* plus the main verb in infinitive form. Even causative verbs (already belonging to a subclass of LSS A1) may participate in this kind of constructions. When undergoing causative alternation, verb class is changed to transitive-causative (A11), with a *cause* argument external to the verb (Arg0-CAU, the syntactic subject), a *theme* argument internal to the verb (Arg1-TEM, the syntactic direct object) and two more optional arguments, a *beneficiary* one (Arg2-BEN, the syntactic indirect object) and an *instrumental* one (Arg3-INS, an adjunct). In these cases, the semantic subject of the main verb (the participant in the event which actually carries out the action expressed by the verb) is the one expressed in the Arg2-BEN argument, when present, or in the Arg1-TEM when not. The causer is sometimes called *inducer agent*. Let us present an example.

- *Unmarked ditransitive sense:*
  - Pedro$_{[Arg0-AGT]}$ **dijo**$_{[A32]}$ algunas cosas feas$_{[Arg1-PAT]}$ a causa de la actitud de Juan$_{[ArgM-CAU]}$.
  - [Pedro said some ugly things due to Juan's attitude.]
- *+Causative alternation:*
  - Con su actitud$_{[Arg3-INS]}$ Juan$_{[Arg0-CAU]}$ **hizo decir**$_{[A11]}$ a Pedro$_{[Arg2-BEN]}$ algunas cosas feas$_{[Arg1-TEM]}$.
  - [With his attitude, Juan made Pedro say some ugly things.]

Apart from this basic causative alternation, a causative-reflexive alternation is also possible for most verb classes, constructed with pronoun *se*, plus a semiauxiliary causative verb (*hacer/fer* or *dejar/deixar*), plus the main verb in infinitive. The most remarkable feature of this causative-reflexive diathesis is that causer and theme coincide in the same argument (the subject-inducer agent). In these cases, we have decided to annotate the subject as Arg0-CAU, considering it an external argument, and not as Arg1-TEM, because we do not want to raise the ambiguity between reflexivity and anticausativity. Therefore, causative-reflexive verb phrases may only have two arguments: Arg0-CAU and Arg2-INS. Let us illustrate this with an example.

- *Unmarked transitive sense:*
  - **Engañó**$_{[A21]}$ al árbitro$_{[Arg1-PAT]}$ con palabras bonitas$_{[ArgM-INS]}$.
  - [(He) deceived the referee with fair words.]

[8]This pronoun has a wide variety of uses both in Spanish and Catalan. For further information on this subject, see the syntactic annotation guide to AnCora corpora: http://clic.ub.edu/ancora/

– *+Causative-reflexive alternation:*

  · El árbitro$_{[\text{Arg0}-\text{CAU}]}$ **se dejó engañar**$_{[\text{A11}]}$ con palabras bonitas$_{[\text{Arg2}-\text{INS}]}$.

  · [lit.: The referee let himself be deceived with fair words.]

5 AnCora-Verb 2.0 and AnCora 2.0: Annotation Criteria

Before starting the discussion on the criteria used in the annotation of AnCora-Verb 2.0 lexicon (section 5.1) and AnCora 2.0 corpora (section 5.2), a word must be said on the format of the data and the general principles driving our philosophy of data storage.

Data in the corpora and lexicon are stored in UTF-8 encoded XML format. XML allows portability and takes advantage of the several tools and libraries available in a variety of platforms and programming languages. Furthermore, XML has a hyerarchical tree structure itself, which maps naturally to the syntactic constituent structure of natural language annotation. Besides, UTF-8 allows the format to be cross-lingual (AnCora data format has been succesfully used in the anotation of corpora in Latin, Arabic and Cyrillic writting).

Our XML structure is guided by three principles. It must be easy to read (intuitive), easy to mantain (coherent), and robust (small changes or errors must not affect overall coherence). For these principles to hold, some design principles have been observed:

Small set of node names. Nodes are generic and specificity is reached through attributes.

Atomic attributes. Each attribute labels one and only one feature of the node. This allows for annotation levels to be independent.

Attributes describe only their node. They do not describe their children/father/sibling nodes. This way, moving, deleting or creating nodes is easy and coherence is guaranteed.

No redundant data.

Easy to add new annotation levels. To do so, only the design of a new attribute or set of attributes with its possible values is needed.

AnCora 2.0 corpora and AnCora-Verb 2.0 lexicons are handled, revised and annotated using AnCoraPipe, a set of java plugins developed for being used in the Eclipse platform. For more information on AnCoraPipe annotation platform and AnCora data format, see [Bertran et al., 2010].

5.1 AnCora-Verb 2.0 lexicon

AnCora-Verb 2.0 is the guide for annotating AnCora 2.0 corpora with verbal semantics (LSS, arguments and θ-roles). It growed and evolved throughout all the annotation process (stretching over two years long) and has been growing ever since, with the addition of new layers of annotation such as deverbal nominal semantics (which cristallized in AnCora-Nom, [Peris et al., 2011, Peris and Taulé, 2011]), or multilingual verb linkings (AnCora-Net project [Taulé et al., 2010]). Nowadays, AnCora-Verb is composed of 2828 files for Spanish and 2248 files for Catalan. Each file contains the information relating to one verb in all its senses and alternations. The name of the file corresponds to the verb in its infinitive form (the enunciation form).

For the file names in the lexicon, special characters have been avoided in order to prevent format or encoding conflicts. Thus, *ñ* is written down as *n* (or *nn* when ambiguity is possible), *l·l* is written down as *l-l*, *ç* is written down as *c*, and accents have been left out. Reflexive pronoun *se* and other clitics are separated from the infinitive they accompany by -, while _ is used to write down multi-words. For instance:

· *banar.lex.xml* stands for *bañar*

- *sonnar.lex.xml* stands for *soñar* (due to the possible ambiguity with *sonar*)
- *al-lucinar.lex.xml* stands for *al·lucinar*
- *amenacar.lex.xml* stands for *amenaçar*
- *abstenir-se.lex.xml* (with reflexive pronoun *se*)
- *treure_foc_pels_queixals.lex.xml* (multi-word)
- *anar-se-n_a_fer_punyetes.lex.xml* (multi-word with several clitics)

File names in the lexicon may be simple words (infinitives) or multi-words (an infinitive followed by a complement or set of complements), such as the Catalan verbal expression *treure_foc_pels_queixals* ('to be extremely angry'; literally: 'to spit fire out of one's teeth'). The similitude between these multi-words and lexicalized arguments (labelled as argL) is great. It has been difficult to decide when to annotate one way or the other. The determining factor is the degree of lexicalization of the expression (the more lexicalized, the more it is likely to become a multi-word). That means that normally lexicalized arguments (argL) allow for some degree of variation, while multi-words do not. Multi-words are lemmatized and treated as a single word (complements included in the expression have no function, they are not considered arguments, they do not receive thematic role and the LSS is assigned to the whole multi-word).

Verb senses within each file are incrementally numbered, with no meaning whatsoever associated to that index. Each sense in each file is represented by a *sense* node father to a number of *frame* nodes, which represent the diatheses in which that particular sense might occur. All senses must have at least one frame, termed and labelled *default*, representing its unmarked LSS. Additional frames, if present, represent diathesis alternations.

All files must have at least one *sense* node, but there is no upper boundary to the number of senses which can be associated with a given lemma. Also, there is no other upper boundary to the number of frames in a sense than its semantics. Just to illustrate, the verb with the most senses, *fer* ('to do' in Catalan), has 101 senses which add up to 123 frames.

The internal structure of each frame is divided into two groups of nodes. First, *argument* nodes reflect constituent structure. Each one is labelled with argument slot, syntactic function and thematic role information concerning a single constituent, either argumental or not. *Argument* nodes can have daughter *constituent* nodes when deemed necessary, which encode the kind of constituent (phrase) that instantiates that particular argument. Second, *example* nodes illustrate the frame with usage samples taken from AnCora 2.0, ranging from a single instance to dozens. It must be noted that, unless stated otherwise, all frames in AnCora-Verb 2.0 have been documented to exist in AnCora 2.0. The only exception to that rule are frames derived from the creation of AnCora-Nom lexicon, which are nevertheless labelled *undocumented*.

Thus, nodes available in AnCora-Verb 2.0 are reduced to 5, all of them hyerarquically organized and dependant of a *lexentry* root node. Each one of these nodes serves as an anchor for different bits of information, encoded through attributes. Each attribute refers to only one feature of the node it accompanies and affects no other nodes. Table 2 illustrates the attributes that accompany each node and the information they encode.

Node name	**Attribute** and *description*
lexentry (root node)	**lemma** *verb in its enunciation form* **lng** *entry language (2 digits: ca/es/en/...)* **type** *syntactic cathegory of the entry (verb)*[9]
sense	**id** *sense index (senses are incrementally numbered)*
frame	**type** *diathesis described; the first frame of any sense must have "default" as value* **lss** *lss the frame belongs to* **undocumented** *"yes" means the frame has not been attested in AnCora 2.0*
argument	**argument** *argument slot of the argument described* **thematicrole** *θ-role associated to this argument* **function** *syntactic function of the argument in the current diathesis*
constituent	**type** *type of phrase that instantiates the argument* **preposition** *preposition that introduces the argument (only for PPs)*
examples	Anchor node for grouping individual examples
example	Text node giving an example

Table 2: Nodes and attributes in AnCora lexicons. Indentation reflects hyerarchy.

[9]AnCora has currently two lexicons: a verbal one, AnCora-Verb 2.0, and a nominal one, AnCora-Nom, where deverbal nominalizations are defined. For further information, see [Peris et al., 2011, Peris and Taulé, 2011].

For the sake of clarity, an actual file from the lexicon is given in Figure 1 and explained next. It corresponds to the Catalan verb *resumir* 'to summarize'.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<lexentry lemma="resumir" lng="ca" type="verb">
  <sense id="1">
    <frame default="yes" lss="A21.transitive-agentive-patient" type="default">
      <argument argument="arg0" function="suj" thematicrole="agt"/>
      <argument argument="arg1" function="cd" thematicrole="pat"/>
      <argument argument="argM" function="cc" thematicrole="mnr">
        <constituent preposition="en" type="sp"/>
      </argument>
      <argument argument="argM" function="cc" thematicrole="loc">
        <constituent preposition="a" type="sp"/>
      </argument>
      <examples>
        <example>
Fernández_Díaz el resumeix en una frase
        </example>
        <example>
que *0* ha resumit a l' Informe que el passat 14_de_març *0* va entregar al president del Parlament , Joan_Rigol
        </example>
      </examples>
    </frame>
    <frame lss="B22.unaccusative-passive-transitive" type="passive">
      <argument argument="arg1" function="suj" thematicrole="pat"/>
      <argument argument="arg0" function="cag" thematicrole="agt"/>
      <examples>
        <example>
Els deu punts es resumeixen en un : Estimar -se com_a família sense presentar mai factura ' , afegeix *0*
        </example>
      </examples>
    </frame>
  </sense>
  <sense id="2">
    <frame default="yes" lss="C21.state-attributive" type="default">
      <argument argument="arg1" function="suj" thematicrole="tem"/>
      <argument argument="arg2" function="cd" thematicrole="atr"/>
      <examples>
        <example>
Aquesta frase pronunciada davant les càmeres de televisió per l'alcalde de Saint-Etienne-en-Dévouly, Jean-Marie_Bernard,
amb llàgrimes als ulls i la veu escanyada, resumeix el sentiment dels habitants del poblet
        </example>
      </examples>
    </frame>
  </sense>
</lexentry>
```

Figure 1: Lexical entry for the Catalan verb *resumir* ('to summarize').

The first line of code is common to most XML files and encodes information about XML version and file encoding, among others. The *lexentry* node is the root node, ancestor to the rest of nodes in the file. Its attributes bear information concerning the whole lexical entry (its lemma, language –*lng*– and syntactic cathegory –*type*). The file is composed of two *senses*, each one with a numeric index specified in the attribute *id*. The first sense has two *frames* associated: the *default* one, conveying a transitive meaning, and its passive diathesis, which entails a change in class (*lss*). The second sense just has one frame, with a stative meaning. Each *frame* has a number of *arguments*, with attributes specifying their argument slot, syntactic function, and θ-role, and a number of *examples*, grouped under a common node. When deemed necessary, constituent information is declared in a *constituent* node daughter to *argument*, as in the case of the two adjuncts of the first frame of the file (labelled *ArgM*). Notice that these *constituent* nodes are optional, as are *frame* nodes other than the default one. Notice also that the number of *examples* is undetermined.

The construction of the lexicon has been a dynamic process, carried out along the annotation of AnCora 2.0 corpora. For every verb instance in a corpus, the corresponding lexicon file was consulted. If the sense and frame was already present in the file, the information contained was mapped to the corpus, thus encoding verbal LSS and argument type and number of the complements, along with their thematic role. If considered informative enough, the sentence containing the particular verb instance was also included as an example (all verb senses and alternations in the lexicon which are present in the corpora must include at least one example). If the sense or diathesis was not present in the file, it was included, and the annotation was carried out as

normal. If the file did not exist, it was created at the moment, after a brief discussion among the annotators on the LSS of the verb when necessary.

5.2 AnCora 2.0 corpora

Next, we summarize the guides for the annotation of semantic elements (LSS, arguments and thematic roles) in the corpora. We will specify which nodes were to be annotated, which attributes and values they should bear, and which annotation exceptions are to be observed. We will begin by describing the annotation of verbs and verb phrases and, next, we will move on to the annotation of syntactic complements of verbs.

5.2.1 Verbs and verb phrases

In the annotation of the corpora, LSS is encapsulated in the *lss* attribute associated to verbal elements, whose possible values are any of the final classes described in section 4. These verbal elements must be terminal nodes[10], including most verbs and some adjectives. Here is complete list of such elements with examples. In the examples, the whole verb phrase is underlined, but only the word in bold letters gets the tag:

1) inflected verbs in their synthetic tense forms (*no* ***acabarem*** *fins que no* ***arribi*** *el material* – 'we will not finish until the material arrives').

2) non personal forms in periphrases (the inflected auxiliary verb does not get LSS):

   a. periphrastic tense forms (*no vàrem* ***acabar*** *fins que el material no va* ***arribar*** – 'we did not finish until de material arrived').

   b. passive constructions with *ser* (*el treball fou* ***acabat*** *quan el material va ser* ***enviat*** – 'the work was finished when the material was sent').

   c. aspectual periphrases[11] (*acabàvem d'****acabar*** *quan el material va començar a* ***arribar*** – 'we had just finished when the material started arriving').

   d. causative constructions with *hacer/fer* or *dejar/deixar* (*els vaig fer* ***acabar*** *el treball després de deixar* ***arribar*** *el material* – 'I made them finish the work after letting the material arrive').

3) infinitives when appearing as a main verb (in non-personal completive clauses) (***acabar*** *el treball va ser complicat sense el material* – 'finishing the work was hard without the material').

4) gerunds when appearing as a main verb (in non-personal adverbial clauses) (***enviant*** *el material, ens van permetre acabar el treball* – 'by sending the material, they let us finish the work').

5) participles when appearing as a main verb:

   a. in absolute clauses (past participle is then annotated as verb-participle)(***acabat*** *el treball, ja no calia esperar el material* – 'once the work was finished, there was no need to wait for the material').

   b. in non-personal adjective clauses with explicit complements (past participle is then annotated as adjective-participle)(*el material,* ***enviat*** *a darrera hora, va arribar just a temps* – 'the material, sent in the last minute, arrived just in time').

On the other hand, there is still a set of verbal elements which do not receive LSS annotation:

1) auxiliary verbs in periphrases (see examples 2a to 2d above).

2) participles in non-personal adjective clauses without explicit complements (*el material enviat va arribar just a temps* – 'the material sent arrived just in time').

[10] Terminal nodes are the lowest nodes in XML hyerarchy. A node is termed terminal if it has no descendants. In AnCora 2.0 corpora, terminal nodes correspond to Parts of Speech (words, speaking in general terms) and punctuation marks.

[11] A complete list of periphrases can be found in the syntactic annotation guide for AnCora corpora ([Soriano et al., 2008])

3) verbal elements participating in verbless sentences or clauses (*el material va ser enviat pel distribuidor i rebut a la fàbrica* – 'the material was sent by the dealer and received at the factory').

An actual example of the annotation of a verb phrase in given in figure 2. It corresponds to the phrase *va ser suspesa* ('was cancelled-FEM'). Notice that only the third terminal node has the *lss* attribute, accordingly to the description just given. The first terminal node corresponds to the auxiliary verb *anar*, used for constructing periphrastical tenses, and the second one corresponds to the semiauxiliary *ser*, used for constructing the passive. Notice also the value of the *lss* attribute, depicting the class this particular instance of the verb belongs to (B22.unaccusative-passive-transitive). All other attributes accompanying *v* terminal nodes label a morphological feature each, and are not relevant for this discussion[12].

```
<grup.verb>
  <v lem="anar" mood="indicative" num="s" person="3" pos="vaip3s0" postype="auxiliary" tense="present" wd="va"/>
  <v lem="ser" mood="infinitive" pos="vsn0000" postype="semiauxiliary" wd="ser"/>
  <v gen="f" lem="suspendre" lss="B22.unaccusative-passive-transitive" mood="participle" num="s" pos="vmp00sf"
   postype="main" wd="suspesa"/>
</grup.verb>
```

Figure 2: Annotation of a verb phrase in AnCora 2.0.

There are 105,101 predicates with the *lss* attribute in AnCora 2.0: 55,990 in Spanish (52,298 verbs and 3,692 adjective-participles) and 49,111 in Catalan (45,817 verbs and 3,294 adjective-participles).

5.2.2 Syntactic complements of verbs

The syntax-semantics interface is broad, and the interactions between meaning and structure have been a matter of discussion in the linguistics field for long. To acknowledge that fact while still being able to annotate AnCora corpora consistently and robustly, we have stablished a hyerarchy between syntactic function (encoded in the attribute *func*), argument slot (encoded in the attribute *arg*) and thematic role (encoded in the attribute *tem*). For an element to be labelled with a thematic role, it must have its argument slot declared, and only elements labelled with syntactic function may have its argumental slot specified. This means that each one of these attributes is bounded to the preceding one following the order func>arg>tem. Nevertheless, none of these attributes is compulsory for any given node type. Rather, they depend on context, yelding a small set of possibilities: we can have nodes with *func, arg* and *tem* attributes, nodes with *func* and *arg*, nodes with just *func*, and nodes with none of the three arguments. Therefore, it is important to determine which nodes may have each one of these tags.

Function attribute (*func*)

More information on the annotation of syntactic functions can be found in [Soriano et al., 2008]. Here, we will just make a brief note to declare which constituents cannot bear function annotation.

All nodes daughter (immediatley dependent) of a sentence node (*sentence*) or a clause node (*S*) must be annotated with their syntactic function via the *func* attribute except in the following cases:

- the verb phrase node (*grup.verb* for finite forms, *participi* for non-finite past participle forms, *gerundi* for non-finite gerund forms, or *infinitiu* for non-finite infinitive forms)
- interjections (*interjeccio*)
- inserted elements (*inc*)
- verbal pronominal morphemes (*morfema.pronominal*)
- conjunctions (*coord* for coordinating ones or *conj* for subordinating ones)
- adjoined constituents
- any node daughter of a verbless sentence or clause (with the *verbless* attribute valued *yes*)
- punctuation marks (*f*)

[12]For more information on the morphological annotation of AnCora 2.0, refer to http://clic.ub.edu/corpus/ancora-documentacio/

As a rule of thumb, it could be said that constituents (phrases and clauses) other than the verb phrase in non-verbless sentences and clauses must receive a *func* attribute. The possible values for this attribute are given in the first column of Table 4 in appendix A.

Argument attribute (*arg*)

This attribute specifies the type of argument each constituent fulfills, and its closeness to the predicate. It has seven possible values (see section 3 and Table 4 in appendix A):

arg0 is associated to the external causer argument of the verb; this value is given to the subject of transitive and inergative verbs and to the agent complement of passives.

arg1 is associated to the first internal argument of the verb; this value is typically given to the direct object of transitive verbs and to the subject of unaccusatives and statives.

arg2 is associated to the second internal argument of the verb; this value is typically given to indirect or prepositional objects of ditransitive verbs and to attributes or prepositional objects of statives and unaccusatives.

arg3 is normally associated to the starting point of change of place or change of state predicates.

arg4 is normally associated to the ending point of change of place or change of state predicates.

argM is associated to non-argumental constituents (adjuncts) in all verb classes.

argL is associated to lexicalized complements of light verbs which admit variation to some extent.

Though incrementally numbered, the values reflecting argumental complements are not incrementally assigned by any means, but rather associated to particular types of constituents from both a semantic and a syntactic point of view. This entails that a given predicate might have an *arg0* and an *arg2* without it needing to also have the intermediate *arg1*.

All nodes with a *func* attribute must be annotated with a corresponding *arg* attribute except in the following cases:

- passive markers (*func="pass"*)
- impersonal markers (*func="impers"*)
- non-argumental verb modifiers (*func="mod"*)
- textual elements (*func="et"*)
- orational adjuncts (*func="ao"*)
- vocatives (*func="voc"*)

Thematic role attribute (*tem*)

This attribute specifies the θ-role each argument carries out. There are 20 possible values, given in table 1 in section 3 and repeated below as a handy reminder. The actual values are given in italics between brackets. Also, consult table 4 in appendix A.

adverbial (*adv*), agent (*agt*), attribute (*atr*), beneficiary (*ben*), cause (*cau*), cotheme (*cot*), destination (*des*), final state (*efi*), initial state (*ein*), experiencer (*exp*), extension (*ext*), purpose (*fin*), instrument (*ins*), location (*loc*), manner (*mnr*), origin (*ori*), patient (*pat*), source (*src*), theme (*tem*), time (*tmp*)

Table 3: Reminder of thematic roles and labels in AnCora 2.0

All nodes with an *arg* attribute must receive a corresponding *tem* attribute, whose value is determined in the lexicon entry for the verb they complement. The only exception to this are lexicalized complements of light verbs (with the *arg* attribute valued *argL*).

The attributes discussed in this section (*func, arg* and *tem*) have to be interpreted in relation to the head verb of the sentence or clause they occur in. It is that verb's *lss* which determines their values, captured and described in the lexicon. Verbs cannot assign this kind of information across sentence or clause boundaries.

Finally, let us present an instance taken from the corpora, to illustrate the annotation format. Semantic tags dealt with in this guide are highlighted in color: red for syntactic function (*func*), green for argument slot (*arg*), blue for θ-role (*tem*) and purple for verb lexical semantical structure (*lss*). The sentence, taken from the Catalan corpus, is *Ha dit diverses vegades que la seva intenció era fer rock and roll com Jerry Lee Lewis* 'He has said several times that his intention was to play rock and roll just like Jerry Lee Lewis'.

```
<sentence id="8">
  <sn arg="arg0" coreftype="ident" elliptic="yes" entity="entity1" entityref="spec" func="suj"
   tem="agt"/>
  <grup.verb>
    <v lem="haver" mood="indicative" num="s" person="3" pos="vaip3s0" postype="auxiliary"
     tense="present" wd="Ha"/>
    <v gen="m" lem="dir" lss="A32.ditransitive-patient-benefactive" mood="participle" num="s"
     pos="vmp00sm" postype="main" wd="dit"/>
  </grup.verb>
  <sn arg="argM" entityref="nne" func="cc" tem="tmp">
    <grup.nom gen="f" num="p">
      <s.a gen="f" num="p">
        <grup.a gen="f" num="p">
          <a gen="f" lem="divers" num="p" pos="aq0fp0" postype="qualificative" wd="diverses"/>
        </grup.a>
      </s.a>
      <n gen="f" lem="vegada" num="p" pos="ncfp000" postype="common" sense="16:05449233"
       wd="vegades"/>
    </grup.nom>
  </sn>
  <S arg="arg1" clausetype="completive" func="cd" tem="pat">
    <conj conjunctiontype="subordinating">
      <c lem="que" pos="cs" postype="subordinating" wd="que"/>
    </conj>
    <sn arg="arg1" entityref="nne" func="suj" tem="tem">
      <spec gen="f" num="s">
        <d coreftype="ident" entity="entity1" entityref="spec" gen="f" lem="el_seu" num="s"
         person="3" pos="dp3fs0" postype="possessive" wd="la_seva"/>
      </spec>
      <grup.nom gen="f" num="s">
        <n gen="f" lem="intenció" num="s" pos="ncfs000" postype="common" sense="16:04588594"
         wd="intenció"/>
      </grup.nom>
    </sn>
    <grup.verb>
      <v lem="ser" lss="C21.state-attributive" mood="indicative" num="s" person="3"
       pos="vsii3s0" postype="semiauxiliary" tense="imperfect" wd="era"/>
    </grup.verb>
    <S arg="arg2" clausetype="completive" func="atr" impersonal="yes" tem="atr">
      <infinitiu>
        <v lem="fer" lss="A21.transitive-agentive-patient" mood="infinitive" pos="vmn0000"
         postype="main" wd="fer"/>
      </infinitiu>
      <sn arg="arg1" entityref="nne" func="cd" homophoricDD="yes" tem="pat">
        <grup.nom gen="m" num="s">
          <n gen="m" lem="rock-and-roll" num="s" pos="ncms000" postype="common" sense="16:cs1"
           wd="rock-and-roll"/>
        </grup.nom>
      </sn>
      <sp arg="argM" func="cc" tem="adv">
```

```
        <prep>
          <c lem="com" pos="cs" postype="subordinating" wd="com"/>
        </prep>
        <sn entityref="ne" ne="person">
          <grup.nom gen="m" num="s">
            <n lem="Jerry_Lee_Lewis" ne="person" pos="np0000p" postype="proper" sense="16:cs1"
             wd="Jerry_Lee_Lewis"/>
          </grup.nom>
        </sn>
      </sp>
    </S>
  </S>
  <f lem="." pos="fp" punct="period" wd="."/>
</sentence>
```

References

[Aparicio et al., 2008] Aparicio, J., Taulé, M., and Martí, M. (2008). AnCora-Verb: A Lexical Resource for the Semantic Annotation of Corpora. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*.

[Bertran et al., 2010] Bertran, M., Borrega, O., Martí, M., and Taulé, M. (2010). *AnCoraPipe: A new tool for corpora annotation (Working paper 1: TEXT-MESS 2.0)*. Universitat de Barcelona.

[Civit and Martí, 2005] Civit, M. and Martí, M. (2005). GramCat and GramEsp: two Grammars for Chunking. *Intelligent Information Processing and Word Mining*.

[Dowty, 1991] Dowty, D. (1991). Thematic Proto-Roles and Argument Selection. *Language*, 67.

[Kingsbury et al., 2002] Kingsbury, P., Palmer, M., and Marcus, M. (2002). Adding Semantic Annotation to Penn TreeBank. In *Proceedings of the 2002 Conference on Human Language Technology*.

[Kipper et al., 2002] Kipper, K., Palmer, M., and Rambow, O. (2002). Extending PropBank with VerbNet Semantic Predicates. In *Workshop on Applied Interlinguas, held in conjunction with AMTA-2002*.

[Levin and Rappaport-Hovav, 1995] Levin, B. and Rappaport-Hovav, M. (1995). *Unaccusativity. At the Syntax-Lexical Semantics Interface*. MIT Press.

[Palmer et al., 2005] Palmer, M., Kingsbury, P., and Gildea, D. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 21(1).

[Peris and Taulé, 2011] Peris, A. and Taulé, M. (2011). AnCora-Nom: A Spanish lexicon of deverbal nominalizations. *Procesamiento del Lenguaje Natural*, 46:11–18.

[Peris et al., 2011] Peris, A., Taulé, M., and Rodríguez, H. (2011). Semantic Annotation of Deverbal Nominalizations in the Spanish corpus AnCora. In *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories (TLT9)*, pages 187–198. University of Tartu.

[Rappaport-Hovav and Levin, 1998] Rappaport-Hovav, M. and Levin, B. (1998). Building Verb Meanings. In Butt, M. and Geuder, W., editors, *The Projection of Arguments: Lexical and Compositional Factors*. CSLI.

[Soriano et al., 2008] Soriano, B., Borrega, O., Taulé, M., and Martí, M. (2008). *Syntactic Annotation Guidelines to AnCora Corpora*. Universitat de Barcelona.

[Taulé et al., 2005] Taulé, M., Aparicio, J., Castellví, J., and Martí, M. (2005). Mapping Syntactic Functions into Semantic Roles. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT2005)*, pages 185–196. Universitat de Barcelona.

[Taulé et al., 2010] Taulé, M., Martí, M., and Borrega, O. (2010). AnCora-Net: Mapping the Spanish AnCora-Verb lexicon to VerbNet. In *Proceedings of the Interdisciplinary Workshop on Verbs. The Identification and Representation of Verb Features*.

[Taulé et al., 2008] Taulé, M., Martí, M., and Recasens, M. (2008). AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*.

[Vendler, 1967] Vendler, Z. (1967). *Linguistics in Philosophy*. Cornell University Press.

[Vàzquez et al., 2000] Vàzquez, G., Fernández, A., and Martí, M. (2000). *Clasificación Verbal. Alternancias de diátesis*. Edicions de la Universitat de Lleida.

A Correspondences between Arguments, θ−Roles and Functions

Attribute <func> **value** / *function*	**Attribute <arg>** **value**	**Attribute <tem>** **value** / *thematic role*	**Example**
suj / *subject*	**arg0**	**agt** / *agent*	**Juan** lee una novela
		cau / *cause*	**El viento** abrió la puerta
		exp / *experiencer*	**Juan** sueña
		src / *source*	**Juan** sudaba
	arg1	**pat** / *patient*	**Clara** es amada por todos
		tem / *theme*	**Juan** llegó tarde
	arg2	**ins** / *instrument*	**Una lona** cubre el coche de Juan
		loc / *locative*	**El diario** abordó la noticia
	argL	*none*	**El cadáver** fue levantado a la 1 de la tarde
cd / *direct object*	**arg1**	**pat** / *patient*	Juan lee **una novela**
		tem / *theme*	El viento abrió **la puerta**
	arg2	**atr** / *attribute*	Juan tiene **un coche blanco**
		ext / *extension*	El paro subió **15.891 personas**
	argL	*none*	Puso **punto final** a la discusión
creg / *prepositional object*	**arg1**	**tem** / *theme*	Clara se rió **del chiste**
		loc / *locative*	Juan intervino **ante la comisión**
	arg2	**tem** / *theme*	Clara sustituyó el vino **por agua**
		loc / *locative*	Juan apoyó la bicicleta **en un árbol**
		ins / *instrument*	Juan va equipado **con un casco**
		atr / *attribute*	Clara joza **de buena salud**
		cot / *cotheme*	Juan conecta el ordenador **a la impresora**
		efi / *final state*	Reconvirtió la habitación **en un estudio**
		ein / *initial state*	Antonio se recupera **de un accidente**
		ext / *extension*	La crisis situó el paro **en 93.278 personas**
	arg3	**ori** / *origin*	El aceite proviene **de las olivas**
	arg4	**des** / *destination*	Juan llevó el coche **al garaje**
	argL	*none*	El niño estalló **en sollozos**
ci / *indirect object*	**arg2**	**ben** / *beneficiary*	Clara **se** lo sugirió **(a Juan)**
		exp / *experiencer*	**(A Clara) le** gusta pasear
	arg3	**ben** / *beneficiary*	**(A Juan) le** salen muy bien las tortillas
		exp / *experiencer*	El chiste **le** pareció divertidísimo
cag / *agent complement*	**arg0**	**agt** / *agent*	Clara es amada **por todos**
cpred / *predicative complement*	**arg2**	**atr** / *attribute*	El niño se llama **Daniel**
	arg3	**atr** / *attribute*	Juan pasó la tarde **sin pensar en Clara**
	argM	**atr** / *attribute*	Suspiró **embelesado**
	argL	*none*	Clara puso **de relieve** su falta de tacto
atr / *attribute*	**arg2**	**atr** / *attribute*	Clara es **abogado**
cc / *adjunct*	**arg1**	**tem** / *theme*	Juan continúa **con sus estudios**
	arg2	**atr** / *attribute*	Clara se pirra **por las patatas fritas**
		cot / *cotheme*	Juan no está emparentado **con Clara**
		efi / *final state*	Clara está inmunizada **contra disgustos**
		ext / *extension*	El juicio se prolongó **hasta el día 15**
		ins / *instrumental*	Juan adornó el discurso **con metáforas**
		loc / *locative*	Clara y Juan residen **en Barcelona**
		tem / *theme*	Clara recibió un regalo **de Juan**
	arg3	**ein** / *initial state*	El semáforo pasó **de verde** a rojo
		ins / *instrumental*	Clara se cepilla **con un cepillo azul**
		loc / *locative*	Nos alertaron **en un comunicado**
		ori / *origin*	Juan ha regresado **de las vacaciones**
	arg4	**des** / *destination*	Clara vino **a casa** ayer
		efi / *final state*	El semáforo pasó de verde **a rojo**
	argM	**adv** / *non-specific adjunct*	Juan vive **con su hermano**
		atr / *attribute*	El día amaneció **cubierto de niebla**
		cau / *cause*	Rompió la foto **por los malos recuerdos**
		ext / *extension*	Clara asumió **cinco años más** el cargo
		fin / *goal*	Lo hizo **para poder dormir a gusto**
		ins / *instrumental*	Ganó la carrera **con una bicicleta nueva**
		loc / *locative*	A Clara le gusta leer **en el jardín**
		mnr / *manner*	Juan lo hace todo **a su manera**
		tmp / *temporal*	Prefiere comer **a las 2**
ao / *sentence adjunct*	*none*	*none*	**Según ella**, Juan se lo merece
et / *textual element*	*none*	*none*	**Y** ¿qué dice él?
mod / *verbal modifier*	*none*	*none*	**No** quiso venir
impers / *impersonality marker*	*none*	*none*	**Se** trata de llegar a tiempo
pass / *passive marker*	*none*	*none*	Los importes **se** calculan en dólares

Table 4: <func>, <arg> and <tem> labels in AnCora 2.0 corpora