

Semantic Annotation of CESS-ECE: Lexical Semantic Structures

1. Theoretical fundamentals in predicate analysis

The correspondence between syntactic functions and thematic roles is carried out following the predicate analysis presented in Levin & Rappaport-Hovav [9] and Rappaport-Hovav & Levin [12]. We consider that their proposal is appropriated for our work for a number of reasons. First, because in this 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 the construction of *VerbNet* (Kipper *et al.* [7]), a lexicon with lexical semantic, argument and diathesis information for English predicates. *VerbNet* follows Levin's semantic classification (Levin [8]) and adopts *PropBank* semantic annotation (Palmer *et al.* [11]).

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 [17], Dowty [4]). These general classes can be split in subclasses, as we will see in section 3. Thematic roles are determined by the event class that the predicate belongs to and the type of diathesis alternation that the predicate presents. Thus, not only thematic roles are assigned, but also predicates are characterized both from the aspectual and argument perspective. In fact, the semantic classes determine the mapping between syntactic functions and semantic roles.

This information is currently being stored in the lexicon CESS-LEX for both languages.

2. Lexical Semantic Structures (LSS)

We have developed general Semantic classes that can be subclassified depending on thematic roles and diatheses. In the definition of these main classes we have taken into consideration only Argument0 and Argument1 because they are the basic arguments in defining the predicate structure. This gives rise to a very coarse grained classification that can be further split into subclasses. This subclassification has not been developed as much detailed as the thematic role assignment, since, although it can be very useful, it is not the main goal of this methodology (mapping thematic roles into syntactic functions). We take four main structures that correspond to the four ontological event classes (Vendler 1967, Dowty 1991), *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 LSS in (1) corresponds to the ontological class state, with just one entity involved in the event, and focuses in the state. The LSS (2) corresponds to activities or processes

it usually presents agentive subjects and allows passive constructions¹. The LSS (3) corresponds to accomplishments that refer to resulting states in external cause processes, usually with a causative subject. Finally, the LSS of (4) corresponds to achievements that refer to a resulting state in processes without external cause.

LSS determines the number of arguments that a verbal predicate requires and the thematic role of these arguments.

In our proposal each LSS restricts the set of all possible diatheses.² Each verb sense is associated to one LSS, and the diatheses that each sense allows are the result of focusing different components of the LSS they belong to. In other words, diatheses are surface structures that result from focusing different components of the predicate LSS (CAUSE, ACT and BECOME).

3. Spanish and Catalan Semantic Classes

In this section, we present the basic Lexical Semantic Classes derived from the LSS mentioned above. These classes are the result of combining the LSS with the Argument Structure and the thematic roles that can fulfill each argument. Each verbal class is also characterized for admitting specific diathesis alternations. All this information is captured in the verbal lexicon CESS-LEX where the syntactic-semantic interface is expressed. For each verbal sense a semantic class is established and the mapping between their syntactic functions³ with the corresponding argument structure and thematic roles is declared.

The semantic classes used to characterize verbal predicates are hierarchically arranged in two levels. The first level contains information about the LSS structure, which is closely related to the event structure. The second level contains information about argument structure and thematic roles. Thus, if a verb is related to a semantic class, it will provide access to syntactic and semantic information, and it will be possible to infer its event structure.

In the next section we present the 11 semantic classes that we so far have compiled, which are grouped around the 4 main LSS types. These classes are the result of the analysis of 1462 verbs in the corpora.⁴

On the basis of a draft of the annotation guide, annotator agreement tests have been carried out. In a first step, 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 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

¹ In Catalan and Spanish there are two types of passive constructions: passives with the participle verb form and passives with 'se' (Esta mañana han sido vendidos cinco libros – Esta mañana se han vendido cinco libros 'Five books have been sold this morning').

² We follow in essence the diathesis classification of Vázquez *et. al* (2000).

³ We extracted the verbal syntactic frames from the corpus as it has been described in Taulé *et al.* (2005) and Civit *et al.* (2005).

⁴ The total number of verbs in the corpus is 1462, which corresponds to 11708 occurrences.

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’). The analysis of the remaining 1000 verbs has been done by the annotators independently.

3.1. LSS1: [x CAUSE [BECOME [y <STATE/THING/PLACE>]]]

LSS1 usually corresponds to the event structure of the accomplishments⁵ and shares the resultative alternation. In this LSS, we distinguish three main classes, the transitive-causative class, the causative agent class and hacer-causative class.

Transitive-Causative class:

The transitive-causative class is characterized by the fact that the verbs belonging to this class accept, as a specific characteristic, the anticausative alternation⁶. These verbs can be characterized as verbs of change of state, where the object is always affected. The subject of these verbs in the causative alternation maps into the Argument-0 with the thematic role *Cause*, and the direct object maps to Argument-1 and *Patient* role. In this case, we are dealing with a direct cause.

LSS1.1 = A1

[x CAUSE [BECOME [y <STATE >]]]

SUJ Arg0##CAU

OD Arg1##PAT

Diatheses: [+Anti-causative] [+Resultative] [+/- Passive]

Example: ‘Juan_{Arg0-CAU} abre la ventana_{Arg1-PAT}’

Spanish verbs: *abrir, aclarar, agotar, alegrarse, alisar, asustarse, babosear, balancear, cerrar, construir, descomponer, emocionar(se), empeorar, enamorar, encandilar, encantar, enfadar, freír, hervir, hundir, inspirar, intrigar, irritar, limpiar, mejorar, molestar, nivelar, orientarse, oscurecer, purificar, reducir, romper, sacralizar, tintar...*⁷

Catalan verbs: *aclarir, bullir, construir, emocionar, enfonsar, esgotar, espantar, fregir, millorar, obrir, purificar, tancar, trencar, ...*

Causative Agent class:

The causative agent class includes basically those verbs implying a change of location, where an acting agent, the syntactic subject, causes the direct object to become in another location or position. Thus, in these cases, we are dealing with indirect causes in which the Argument-0 is represented as an *Agent* and the Argument-1 as a *Patient*. Most of these verbs allow the passive alternation and not the anticausative one. We consider that all these facts support the treatment of the subject as an *Agent*.

LSS1.2 = A2

[x CAUSE [BECOME [y <PLACE >]]] or [x CAUSE [BECOME [<THING> IN y]]]

SUJ Arg0##AGT

OD Arg1##PAT

⁶ Anti-causative alternation is also known as ergative or inchoative alternation.

⁷ We understand that it is one of the possible senses of these verbs. Obviously, we can find that the same verb belongs to different semantic classes because of its polysemy.

Diatheses: [- Anti-causative] [+/-Resultative] [+Passive]

Example: ‘El médico Arg0##AGT hospitalizó al paciente Arg1##PAT’
‘Juan Arg0##AGT ensilla el caballo Arg1##PAT’

Spanish verbs: *almacenar, bajar (an object), colocar, encarcelar, empaquetar, empapelar, enharinar, ensillar, ensobrar, hospitalizar, meter (las sillas), poner, subir (an object)*...

Catalan verbs: *baixar (an object), col·locar, emmagatzemar, emmantegar, empaquetar, empresonar, ensobrar, hospitalitzar, posar, pujar (an object), ...*

Hacer-causative class

In this class are included complex predicates composed by the verb ‘hacer’ followed by an infinitive. The presence of the ‘hacer’ predicate determines the causative interpretation of the infinitive verb. These complex predicates can be characterized as verbs of change of state, where the object is always affected. In contrast with the transitive-causative class, these verbs cannot accept the anticausative alternation neither the passive. This is the reason that the direct object maps to Argument-1 with the thematic role Theme. The subject of these verbs maps into an Argument-0 with the thematic role Cause.

LSS1.3 = A3

[x CAUSE [BECOME [y <STATE >]]]

SUJ Arg0##CAU

OD Arg1##TEM

Diatheses: [- Anti-causative] [-Resultative] [-Passive]

Example: ‘Este tratamiento Arg0##CAU hace crecer al niño Arg1##TEM’

Spanish verbs: *hacer crecer, hacer llorar,*

Catalan verbs: *fer créixer,*

3.2. LSS2: [BECOME [y <STATE >/<PLACE>]]

Verbs belonging to the LSS2 correspond to the event structure of achievements, and they are basically unaccusative verbs. Currently, we have distinguished two classes: the unaccusative class and the unaccusative with final state. In the unaccusative class we have included the verbs of inherent directed motion and verbs of appearance and disappearance. In the unaccusative with final state class we have included those verbs indicating the final state. Verbs belonging to the LSS2 neither participate in the passive, the anticausative nor the resultative alternation. The subject maps into the Argument-1 with the thematic role *Theme*.

Unaccusative class:

This class includes intransitive verbs whose subject behaves as an internal argument. In some languages, such as Catalan⁸, this subject is characterized by the fact that allows the cliticization with the pronoun ‘en’; for example: ‘Han arribat quinze turistes’ vs. ‘N’han arribat quinze’ (‘Fifteen tourists have arrived’). The subject of these verbs usually appears in the postverbal position. This last characteristic is also found in

⁸ As well as in Italian and French.

Spanish. Most of these verbs are included in the Levin's *inherently directed motion class*, which is a subgroup of verbs that can express the Origin (Argument-1) and the Goal (Argument-2), as in 'Ana viene de París_{CREG-Arg1-ORI}' ('Ana comes from Paris'); 'Ana sale de casa_{CREG-Arg2-DES}' ('Ana leaves home').

LSS2.1 = B1

[BECOME [y <PLACE>]] or [BECOME [y <STATE>]]

SUJ Arg1##TEM

Diatheses: [-Passive]

Example: Los niños_{Arg1##TEM} llegaron tarde
 'The kids arrived late'
 Los ladrones_{Arg1##TEM} desaparecieron sin dejar rastro
 'The thieves vanished without a trace'

Spanish verbs: *acabar_en*, *aparecer*, *caer*, *crecer*, *desaparecer*, *desembocar* (*concluir_en*), *emerger*, *entrar*, *florecer*, *llegar*, *morir*, *salir*, *venir*,...

Catalan verbs: *aparèixer*, *arribar*, *caure*, *créixer*, *desaparèixer*, *entrar*, *morir*, *sortir*, *venir*, ...

Unnacusative final state class:

This class includes intransitive verbs whose subject behaves as an internal argument, mapping the Argument-1 with the thematic role *Theme*. This class is characterized by the fact that they have an Argument-2 indicating the *Final State*.

LSS2.2 = B2

[BECOME [y <STATE>]]

SUJ Arg1##TEM

CC Arg2##EFI

Diatheses: [-Passive]

Example: 'Los niños_{Arg1##TEM} caen enfermos'

Spanish verbs: *caer enfermo*, *entrar en coma*, *entrar en silencio*

Catalan verbs: *entrar en coma*

3.3. LSS3: [x/y <STATE >]

The verbal classes related to LSS3 denote states, and typically they can not be controlled by an *Agent*. We have distinguished four classes depending on the argument structure and the type of subject allowed by the verbal predicates. We basically differentiate between *state unaccusative*, *state unergative*, *state transitive* and *state measure* classes.

State Unaccusative class:

All the members of this class have intransitive uses and they are specifically treated as unaccusative. We represent their subject as an Argument-1 mapping the thematic role *Theme*. We also include aspectual verbs in this class, that is to say, verbs that basically describe the initiation and termination of an activity.

LSS3.1 = C1

[y <STATE >]

SUJ Arg1##TEM

Diatheses: [-Passive] [-Cognate Object]

Example: 'El año Arg1##TEM acaba el 31 de diciembre ArgM##TMP'
'El 31 de diciembre ArgM##TMP acaba el año Arg1##TEM'

Spanish verbs: *acabar, comenzar, empezar, existir, habitar, terminar, vivir (en Barcelona),...*

Catalan verbs: *acabar, començar, existir, haver-hi,...*

State Unergative class:

This class comprises unergative verbs denoting a state. Though intransitive, they are different from LSS2.1 in their thematic role assignment, since subjects of *State Unergative* verbs are Argument-0 mapping the role *Experiencer*.

LSS3.2 = C2

[x <STATE > y]

SUJ Arg0##EXP

OD Arg1##TEM

Diatheses: [-Passive], [+Cognate Object]⁹

Example: 'Juan Arg0##EXP sueña'

Spanish verbs: *babear, brillar¹⁰, burbujear, centellar, crecer (niño), chorrear, destellar, dormir, llorar, nacer, oír (p.e.: oír bien/mal), parpadear, respirar, roncar, soñar, sudar, temblar, ver, vivir,...*

Catalan verbs: *brillar, dormir, roncar, somiar, suar, tremolar, viure...*

State Transitive class:

This class is mainly integrated by copulative verbs. Passive alternation does not occur in this class of verbs. The Argument-1 maps the role *Theme* and the Argument-2 maps the role *Attribute*.

LSS3.3 = C3

[x <STATE > y]

SUJ Arg1##TEM

CD /ATR/CPRED Arg2##ATR

Diatheses: [-Passive]

Example: 'Juan Arg1##TEM tiene dos hipotecas Arg2##ATR'
'La película Arg1##TEM es interesante Arg2##ATR'

Spanish verbs: *anteceder, estar, inspirar, oler, parecer, poseer, ser, significar, tener ...*

Catalan verbs: *estar, posseir, semblar, ser, tenir...*

⁹ Posar-ne un exemple: 'Respirava l'aire fresc'.

¹⁰ Aquesta manera de tractar els papers temàtics, independentment de les restriccions selectives de l'entitat, ens permet tractar amb el mateix paper temàtic tant el subjecte d'un verb com somriure 'el nen somriu' com el subjecte d'un verb com brillar 'l'estel brilla' (l'ús metafòric, 'aquella nit l'actor va brillar'). El martillo/viento/niño es tracten tots com Cause.

State Measure class:

This class includes those verbs that describe the value of some attribute of an entity along a scale (*Measure verbs* in the Levin's classification (Levin B., 1993)). We represent the Argument-1 mapping the thematic role *Theme* and the Argument-2 mapping the thematic role *Extension*.

LSS3.4 = C4

[x <STATE > y]

SUJ Arg1##TEM

CD Arg2##EXT

Diatheses: [-Passive]

Example: 'Juan_{Arg1##TEM} mide dos metros_{Arg2##EXT}'

Spanish verbs: *costar, medir, pesar, valer...*

Catalan verbs: *costar, medir, pesar, valer...*

3.4. LSS4: [x ACT <MANNER/INSTRUMENT> y]

Most verbal classes related to LSS4 denote activities and, consequently, the verbs involved share an agentive subject. That is, Argument-0 always maps into thematic role *Agent*, while the Argument-1, if there is any, always fits with *Patient*. If there is a *Patient*, the passive alternation is necessarily possible. By the moment, we have distinguished five different semantic classes, mainly depending on the predicate's arity: *agentive inergative class*, *agentive transitive class*, *agentive ditransitive class*, *locative ditransitive class* and *benefactive ditransitive*.

Agentive Unergative class:

All the verbs in this class have intransitive uses and most of them typically describe manner of motion, involving or not displacement. Most verbs involving movement (i.e. *nadar, correr*, etc.) can display the extension object alternation, that is, they can be used in a transitive way expressing an extension or a measure phrase.

LSS4.1 = D1

[x ACT <MANNER/INSTRUMENT> y]

SUJ Arg0##AGT

CD ArgL##

Diatheses: [-Passive], [+/-Extension Object]

Example: 'Juan_{Arg0##AGT} corre '

'Juan_{Arg0##AGT} caminó tres kilómetros_{Arg1##EXT}'

Spanish verbs: *caminar, contonearse, correr, escapar, establecerse, ir, menear(se), nadar, perseguir...*

Catalan verbs: *anar, caminar, córrer, cridar, nadar, ...*

Agentive Transitive class:

This class comprises verbs typically transitive that present Argument-0 with the thematic role *Agent* and Argument-1 with *Patient*. It is the largest class in Catalan and Spanish languages.

LSS4.2 = D2

[x ACT <MANNER/INSTRUMENT> y]

SUJ Arg0##AGT

CD Arg1##PAT

CREG Arg1##

CREG Arg2#en# (*)

CPRED Arg2##ATR (**)

Diatheses: [+Passive]

Example: 'Juan Arg0##AGT lee una novela histórica'

Spanish verbs: *aconsejar, amar, barrer, beber, cantar, cazar, cepillar, comer, desear, escuchar, forzar, fregar, leer, mirar, odiar, oler, orientar, peinar, silbar... oír (oía canciones de amor), ver (veo la película), (como escuchar y mirar)*

(*) *Citar, convencer, dotar, incitar, instar, invertir, traducir*

(**) *Declarar, entender, considerar, llevar, mantener*

Catalan verbs: *beure, cantar, desitjar, escombrar, estimar, fregar, llegir, odiar, pentinar, raspallar, xiular...*

Agentive Ditransitive class:

The verbs of this semantic class are characterized by presenting a double object, one expressing the *Patient* (Argument-1) and another referring to the *Beneficiary* (Argument-2). For example, verbs expressing change of possession and communication verbs can fit this class. That is, when any kind of transfer of possession, information or ideas is carried out.

LSS4.3 = D3

[x ACT <MANNER/INSTRUMENT> y]

SUJ Arg0##AGT

SUJ ArgL##

CD Arg1##PAT

OI Arg2##BEN

CREG Arg1##

Diatheses: [+Passive], [-Subject Locative]

Example: 'Juan Arg0##AGT da un caramelo Arg1##PAT a la niña Arg2##BEN'

Spanish verbs: *cantar, contar, dar, decir, entregar, enviar, explicar,...*

Catalan verbs: *cantar, contar, dir, donar, enviar, explicar, lliurar...*

Locative Ditransitive class:

This class is characterized by admitting the subject locative alternation, that is, Argument-2 with thematic role *Locative* can occur in a subject position, for example: 'El autor aborda la discriminación de género en su ensayo_{CC-Arg2-LOC}' vs. 'El ensayo_{SUJ-Arg2-LOC} aborda la discriminación de género' ('The author tackles gender discrimination in his essay' vs 'The essay tackles gender discrimination').

LSS4.4 = D4

[x ACT <MANNER/INSTRUMENT> y]

SUJ Arg0##AGT

CD Arg1##PAT

CC Arg2##LOC

Diatheses: [+Passive], [+Subject Locative]

Example: 'El autor_{Arg0##AGT} aborda esa temática_{Arg1##PAT} en la novela_{Arg2##LOC}'
'La novela_{Arg2##LOC} aborda esa temática_{Arg1##PAT}'

Spanish verbs: *abordar, acoger, registra, tratar...*

Catalan verbs: *abordar, recollir, registrar, tractar...*

Benefactive ditransitive class

This class comprises verbs with two arguments: Argument-0 with the thematic role *Agent* and Argument-2 with *Benefactive*.

LSS4.5 = D5

[x ACT <MANNER/INSTRUMENT>]

SUJ Arg0##AGT

CD ArgL##

OI Arg2##BEN

Diatheses: [-Passive], [-Subject Locative]

Example: 'Me_{Arg2##BEN} gustan las patatas_{Arg0##AGT}'

Spanish verbs: *gustar, dar_un_beso, dar_ganas, ...*

Catalan verbs: *agradar...*