AnCora-Verb: A Lexical Resource for the Semantic Annotation of Corpora

Juan Aparicio, Mariona Taulé, Ma. Antònia Martí

CLiC, Centre de Llenguatge i Computació - University of Barcelona
Gran Via de les Corts Catalanes 585, 08007 Barcelona
{juanaparicio, mtaule, amarti}@ub.edu

Abstract

In this paper we present two large-scale verbal lexicons, AnCora-Verb-Ca for Catalan and AnCora-Verb-Es for Spanish, which are the basis for the semantic annotation with arguments and thematic roles of AnCora corpora. In AnCora-Verb lexicons, the mapping between syntactic functions, arguments and thematic roles of each verbal predicate it is established taking into account the verbal semantic class and the diatheses alternations in which the predicate can participate. Each verbal predicate is related to one or more semantic classes basically differentiated according to the four event classes ─accomplishments, achievements, states and activities─, and on the diatheses alternations in which a verb can occur. AnCora-Verb-Es contains a total of 1965 different verbs corresponding to 3671 senses and AnCora-Verb-Ca contains 2151 verbs and 4513 senses. These figures correspond to the total of 500,000 words contained in each corpus, AnCora-Ca and AnCora-Es. The lexicons and the annotated corpora constitute the richest linguistic resources of this kind freely available for Spanish and Catalan. The big amount of linguistic information contained in both resources should be of great interest for computational applications and linguistic studies. Currently, a consulting interface for these lexicons is available at (http://clic.ub.edu/ancora/).

1. Introduction

In this paper we present two large-scale verbal lexicons, AnCora-Verb-Ca for Catalan and AnCora-Verb-Es for Spanish, which are the basis for the semantic annotation with arguments and thematic roles of AnCora corpora. At present, AnCora (Martí et al., 2008; Taulé et al., 2008) is the largest multilevel annotated corpus of Spanish and Catalan consisting of 500,000 words each mostly from newspaper articles. AnCora is annotated with morphological (PoS), syntactic (constituents and functions) and semantic (argument structure and thematic roles, semantic class, named entities and WordNet senses) information.

In AnCora-Verb lexicons, the mapping between syntactic functions, arguments and thematic roles of each verbal predicate it is established taking into account the verbal semantic class and the diatheses alternations in which the predicate can participate. Each verbal predicate is related to one or more semantic classes, depending on its senses. The main goal of this paper is to present the content of these lexicons and their resulting projection in the AnCora corpora (section 2). A quantitative analysis of the data it is also presented (section 3). Finally, main conclusions are drawn in section 4.

2. AnCora-Verb Lexicons

AnCora-Verb lexicons were obtained by deriving, for each sense of each verb, all the syntactic schemata in which a verbal predicate appears in AnCora corpora (Taulé et al., 2008). From this information, the mapping from syntactic functions to thematic roles, and the corresponding argument position, was fully manually encoded in the lexicons. The semantic properties used in the characterization of predicates are based on the proposal of lexical decomposition of Rappaport-Hovav & Levin (1998) from which the concept of Lexical Semantic Structure (LSS) has been taken. For the characterization of the argument structure, we follow PropBank annotation system (Palmer et al., 2005)[1]. In this direction, we follow the lines laid down by Kingsbury et al., (2002) in the construction of VerbNet.

In AnCora-Verb lexicons, each predicate is related to one or more semantic classes (LSS), depending on its senses, basically differentiated according to the four event classes ─accomplishments (A), achievements (B), states (C) and activities (D)─, and on the diatheses alternations in which a verb can occur.

Figure 1 shows the full information associated with the entry *reforzar* 'to reinforce': the lemma (*reforzar*), the different senses associated to their corresponding semantic classes (in this case LSS1.1 and LSS2.2), the mapping between syntactic function and thematic role (for instance, SUJ Arg0##CAU), and the diatheses alternations in which the verb occurs (in this case, ANTICAUSATIVA 'inchoative'). As we can observe, the expression of the causative-inchoative alternation entails an argument crossing: the affected object, appears as direct object in the causative structure (CD Arg1##TEM) and as subject in the inchoative structure (SUJ Arg1##TEM). Furthermore, the expression of this alternation also involves an aspectual change, since the causative reading corresponds with an accomplishment (LSS1.1) and the inchoative reading with an achievement (LSS2.2). Finally, examples are also included.

[1] The arguments selected by the verb are incrementally numbered –Arg0, Arg1, Arg2, Arg3, Arg4– expressing their degree of proximity in relation to its predicate. The adjuncts are labelled as ArgM. The list of thematic roles consists of 20 different thematic labels: AGT (Agent), AGI (Induced Agent), CAU (Cause), EXP (Experiencer), SCR (Source), PAT (Patient), TEM (Theme), ATR (Attribute), BEN (Beneficiary), EXT (Extension), INS (Instrument), LOC (Locative), TMP (Time), MNR (Manner), ORI (Origin), DES (Goal), FIN (Purpose), EIN (Initial State), EFI (Final State) and ADV (Adverbial).

reforzar - 01
LSS1.1 (A1)

SUJ	Arg0##CAU
CD	Arg1##TEM
EX:	"La subida en dos décimas de la tasa de paro reforzó la tendencia al alza"[2]

+ANTICAUSATIVA
LSS2.2 (B2)

SUJ	Arg1##TEM
EX:	"Si dos neuronas se activan, sus conexiones se refuerzan"[3]

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

In order to guarantee the coherence and quality and to ensure the correct mapping between arguments, thematic roles, syntactic functions and LSS, inter-annotator agreement tests were carried out in the building process of the verbal lexicons. After a first proposal of verb classes and their corresponding arguments and theta-roles, a group of seven trained linguists elaborated a subset of 30 verbal entries. The resulting entries were compared, the disagreements discussed and the verb classes modified when necessary. This process was applied over several subsets of 30 verbs until no relevant disagreements arose. Disagreements were mainly due to differences in class assignment (LSS), and therefore also in the thematic role assignment. For example, in Spanish, a verb in a passive ('pasiva refleja') or inchoative ('anticausativa) construction can appear with the pronoun *se*, and it is not always easy to decide which of them the correct interpretation is and, obviously, the consequences are also very different. If we opt for the passive reading, the Arg0 is an Agent, whereas if we choose the inchoative reading the Arg0 is a Causer. The identification of multiwords, for instance the treatment of light verbs, is also especially problematic, basically when it is necessary to decide if a given structure corresponds to a verb and its complements or to an idiom (*tener* + *ganas* vs. *tener_ganas*, 'to need' or 'to want').
Next we present the 13 semantic classes that have been used for the characterization of verbal predicates:

Accomplishments (A)

A1: Transitive-Causative class:
LSS1.1 [x CAUSE [BECOME [y <STATE >]]]
Arg0##CAU
Arg1##TEM
Diatheses: [+Inchoative] [+Resultative]
Spanish verbs: *abrir* 'to open', *causar* 'to cause', *cerrar* 'to close', *romper* 'to break'…
Catalan verbs: afectar 'to affect', convertir 'to turn into', omplir 'to fill'…

A2: Transitive-agentive class:
LSS1.2 [[x DO-SOMETHING] CAUSE [BECOME [y <STATE>]]]
Arg0##AGT
Arg1##PAT
Diatheses: [+Passive]
Spanish verbs: *comer* 'to eat', *escribir* 'to write'…
Catalan verbs: afirmar 'to affirm', *llegir* 'to read'…

A3.1: Ditransitive-agentive locative class:
LSS1.3.1 [[x DO-SOMETHING] CAUSE [BECOME [y <PLACE> z]]]
Arg0##AGT
Arg1##PAT
Arg2##LOC
Diatheses: [+Passive]
Spanish verbs: *colocar* 'to place', *dejar* 'to leave',
Catalan verbs: *moure* 'to move', *posar* 'to put'...

A3.2: Ditransitive-agentive beneficiary class:
LSS1.3.2 [[x DO-SOMETHING] CAUSE [BECOME [y <PLACE> z]]]
Arg0##AGT
Arg1##PAT
Arg2##BEN
Diatheses: [+Passive]
Spanish verbs: *dar* 'to give', *decir* 'to tell',
Catalan verbs: *enviar* 'to send', *vendre* 'to sell'…

Achievements (B)

B1: Unaccusative-motion class
LSS2.1 [BECOME [y <PLACE>]]
Arg1##TEM/PAT
Spanish verbs: *llegar* 'to arrive', *salir* 'to go_out'…
Catalan verbs: *entrar* 'to go_in', *venir* 'to come'…

B2: Unaccusative-state class
LSS2.2 [BECOME [y <STATE>]]
Arg1##TEM/PAT
Arg2##EFI
Spanish verbs: *crecer* 'to grow', *florecer* 'to bloom'…
Catalan verbs: *enfonsar-se* 'to collapse'…

States (C)

C1: Existence-state class
LSS3.1 [x <STATE>y]
Arg1##TEM
Arg2##LOC
Spanish verbs: *estar* 'to be', *existir* 'to exist'…
Catalan verbs: *haver-hi* 'there_is/are'…

C2: Attributive-state class
LSS3.2 [x <STATE>y]
Arg1##TEM
Arg2##ATR
Spanish verbs: *ser* 'to be', *tener* 'to have'…
Catalan verbs: *estar* 'to be', *tenir* 'to have'…

[2] 'The rise in two tenths of the unemployment rate reinforced the bullish tendency'.
[3] 'If two neurons are activated, their connections are reinforced'.

C3: Scalar-state class
LSS3.3 [x <STATE>y]
Arg1##TEM
Arg2##EXT
Spanish verbs: *medir* ‘to measure’, *pesar* ‘to weigh’…
Catalan verbs: costar ‘to cost’, durar ‘to last’…

C4: Beneficiary-state class
LSS3.4 [x <STATE>y]
Arg1##TEM
Arg2##BEN/EXP
Spanish verbs: *gustar* ‘to like’, *parecer* ‘to seem’…
Catalan verbs: *agradar* ‘to like’, *preocupar* ‘to worry’…

Activities (D)

D1: Agentive-inergative class
LSS4.1 [x ACT <MANNER/INSTRUMENT >]
Arg0##AGT
Spanish verbs: *caminar* ‘to walk’, *nadar* ‘to swim’…
Catalan verbs: *jugar* ‘to play’, *navegar* ‘to sail’…

D2: Experiencer-inergative class
LSS4.2 [x ACT <MANNER/INSTRUMENT >]
Arg0##EXP
Spanish verbs: *dormir* ‘to sleep’, *soñar* ‘to dream’...
Catalan verbs: *respirar* ‘to breath’…

D3: Source-inergative class
LSS4.3 [x ACT <MANNER/INSTRUMENT >]
Arg0##SRC
Spanish verbs: *roncar* ‘to snore’, *sudar* ‘to sweat’…
Catalan verbs: *cridar* ‘to shout’, *plorar* ‘to cry’…

2.1. Automatic Annotation

AnCora-Verb lexicons were used for the semiautomatic tagging of the AnCora corpora with arguments, thematic roles and semantic classes. A set of manually written rules automatically mapped part of the information declared in these lexicons onto the syntactic structure (Martí et al., 2007). We defined three different types of rules taking into account the kind of information they were based on:

a) Rules based on a specific function or morphosyntactic property. For example, if the predicate has associated the verbal morpheme ‘PASS’ (passive voice), then its subject has the argument position Arg1 and the thematic role patient (SUJ-Arg1-PAT).
b) Rules based on the semantic properties of the predicates. For instance, when predicates are monosemic, the mapping between syntactic function and argument and thematic role as well as the assignment of the semantic class is directly realized. In the case of polysemic verbs, the mapping can be partial because it is only automatically assigned the unambiguous information.
c) Rules based on the type of adverb or prepositional multiword appearing in a specific constituent. For instance, if the prepositional multiword *a_causa_de* (‘because_of’) or the adverb *aún* (‘still’, ‘yet’) in Spanish, appears in an adverbial complement (function = CC), then it is automatically assigned the argument and thematic role ArgM-CAU (an adjunct argument with the thematic role cause) as well as ArgM-TMP (an adjunct argument with the thematic role temporal) respectively.

We applied these rules following a decreasing heuristic according to the degree of generality, that is, we applied first the more general rules of type a), secondly the type c) rules and, finally, the type b) rules. In the automatic annotation process we obtained either full annotations – containing information about the arguments and the thematic roles– or partial annotations with only arguments or thematic roles. This procedure permits to automatically annotate 60%[4] of the expected arguments and thematic roles with a fairly low error (below 2%) (Martí et al., 2007). Given the high quality of the results obtained we claim that this methodology is very suited for the semi-automatic approach to corpus annotation and able to save a significant amount of manual effort. Afterwards we manually completed the thematic role annotation in order to guarantee the accuracy required to support the final resource. The Catalan corpus, AnCora-Ca, is already completed for the 500,000 words, while the semantic manual checking covers, up to now, the 100,000 words of the Spanish corpus, AnCora-Es. The Spanish corpus will be completed at the end of this year.

3. Quantitative Analysis of Data

The Spanish lexicon, AnCora-Verb-Es, contains a total of 1965 different verbs (corresponding to 3671 senses) and the Catalan lexicon, AnCora-Verb-Ca, contains 2151 verbs (corresponding to 4513 senses).

In table 1, the distribution of these verbs’ senses in semantic classes it is shown for both languages. The average of senses per lemmata is 1,86 for Spanish and 2,09 for Catalan.

Table 1 shows that the semantic class with the highest number of different verbs, in both languages, is by far the transitive-agentive class (A2) followed by the unaccusative-state class (B2) and the causative-transitive class (A1). It has to be noticed that in B2 class the passive or inchoative constructions coming from other classes (A1, A2 and A3) as result of a diatheses alternation are also included. For instance, the passive alternation of the verbal predicate *verificar* ‘to verify’ (from A2 semantic class) is annotated as B2 (See figure 2). The expression of most alternations entails an aspectual change, which necessarily implies a change of semantic class.

[4] From which the 94% corresponds to full annotations and 6% to partial annotations.

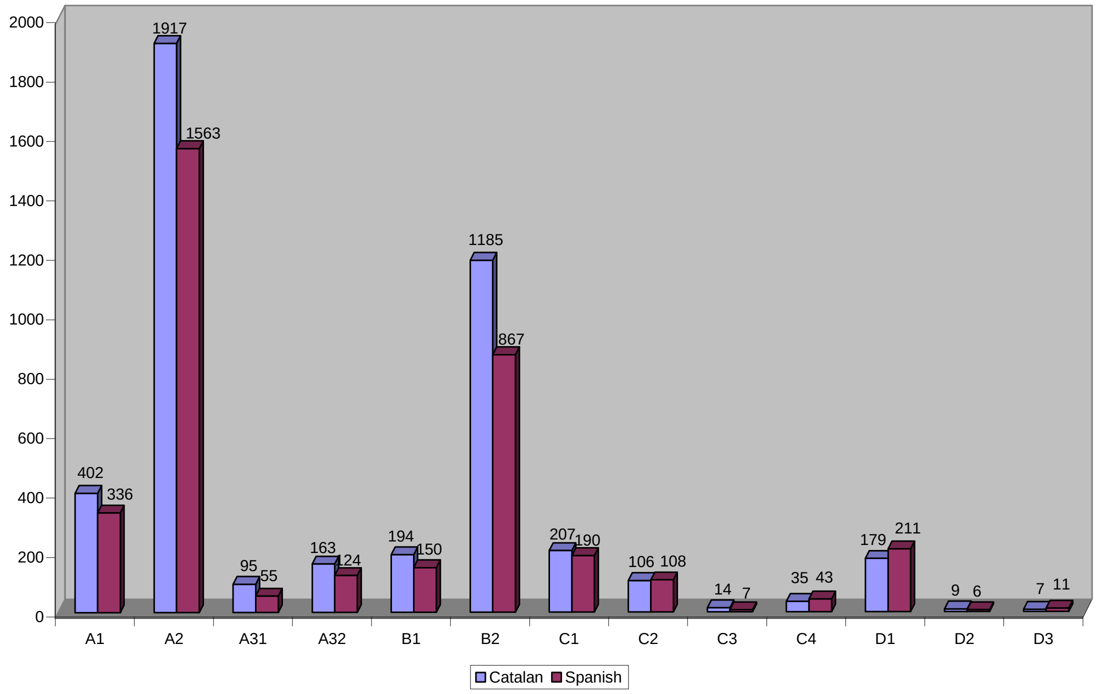

Table 1: Verbs associated to each semantic class

verificar - 01
LSS1.2 (A2)
SUJ Arg0##AGT
CD Arg1##PAT
EX: "(…)verificar la responsabilitat dels actuals directius"[5]

+PASSIVA
LSS2.2 (B2)
SUJ Arg1##PAT
EX: "(…) que es verifiqui l'honradesa dels càrrecs publics"[6]

Figure 2: Lexical entry of *verificar* 'to verify' in AnCora-Verb-Ca

Next we present the figures corresponding to the projection of AnCora-Verb-Es and AnCora-Verb-Ca lexicons in AnCora-Es and AnCora-Ca corpora respectively (See table 2 and table 3).

For the quantitative analysis of the data we have taken into account the 500,000 words fully annotated for Catalan and a subset of 100,000 words for Spanish. These figures correspond to the total amount of semantic annotated data manually checked.

The Spanish subset comprises a total amount of 11,061 verbal tokens, corresponding to 2613 senses (Table 2). The Catalan subset comprises a total of 48,319 verbal tokens corresponding to 4102 senses (Table 3). In this context, we understand for sense the number of different lemmata associated to each verbal class.

Spanish Semantic Classes			
LSS	**Tokens**	**lemmata (senses)**	**%**
A1	485	192	4.38
A2	3833	886	34.65
A31	210	46	1.90
A32	755	82	6.82
B1	681	184	6.16
B2	1406	756	12.71
C1	736	160	6.65
C2	2299	110	20.78
C3	4	1	0.04
C4	195	33	1.76
D1	428	149	3.87
D2	15	5	0.14
D3	14	9	0.13
Total	11,061	2613	100

Table 2: Figures corresponding to a corpus-sample of 100,000 words fully annotated

Even though the number of words annotated in Catalan is upper than that of the Spanish, if we compare the data to level of percentages the results are parallel. In both languages the verbal classes with the highest number of occurrences are: the transitive-agentive class (A2) with 3833 in Spanish and 8165 in Catalan; the attributive-state class (C2) with 2299 in Spanish and 8117 in Catalan; and the unaccusative-state class (B2) with 1406 in Spanish and 7631 in Catalan. The sum of which

[5] '(…) verifying the responsibility of the current managers'

[6] '(…)that the honesty of the publics charges is verified'

means the 68.14% and the 70.18% of the total verbal occurrences annotated in AnCora-Es and AnCora-Ca respectivelly. If we take into account the number of different lemmata associated to each of these three classes, we can observe that A2 presents the highest number of different types (886 for Spanish and 1311 for Catalan), followed by B2[7] (756 and 1496 for Spanish and Catalan respectively); while C2 class has the lower number (only 110 for Spanish and 115 for Catalan). It means that A2 and B2 classes are more sparsely distributed than C2.

Catalan Semantic Classes			
LSS	**Tokens**	**lemmata (senses)**	**%**
A1	1826	311	3.78
A2	8165	1311	37.59
A31	1001	77	2.07
A32	4129	140	8.55
B1	2368	273	4.90
B2	7631	1496	15.79
C1	3208	182	6.64
C2	8117	115	16.80
C3	139	18	0.29
C4	561	30	1.16
D1	1143	138	2.37
D2	20	5	0.04
D3	11	6	0.02
Total	48,319	4,102	100

Table 3: Figures corresponding to 500,000 words fully annotated

Verbs belonging to A32, C1, B1, A1 and D1 semantic classes represent a little bit more than the 25% of the total verbal predicates appeared in the corpora, the 26.24% for Spanish and 27.88% for Catalan. Whereas the rest of verbal classes -A31, C4, C3, D2 and D3- represent the 3.97% and the 3.58% of the total verbal occurrences in AnCora-Es and AnCora-Ca corpora respectively.

In order to get more information about how verbal predicates are distributed in each semantic class, we have obtained the frequency of the 10 more frequent lemmata for each class and its corresponding percentage with respect to the total amount of the class (See Table 4 for Spanish and Table 5 for Catalan). Notice that despite the difference in corpus size, the percentages overlap to a great extent. However, this overlapping does not take place in all verbs.

Table 4 and 5 show that, for example, in the attributive-state class (C2) and the beneficiary-state class (C4), the 10 more frequent lemmata represent the 83,9% and the 75.89% of the total verbal tokens of these classes for Spanish, and for Catalan the 88,94% (C2) and the 92.8% (C4). The same subset in the scalar-state class (C3) in Catalan covers also the 92.8% of the total class tokens. Therefore, the state classes have few verbal types but they present a very high occurrence in both corpora. In fact, the verb *ser* ('to be') is the one with the highest frequency in both languages. On the opposite side we find the unaccusative-state class (B2), in which the 10 more frequent lemmata only represent the 8.8% for Spanish and the 14.34% for Catalan.

LSS	**Spanish verbs**[8]	**%**
A1	provocar:18; convertir:15; poner:14; hacer:13; abrir:13; mejorar:13; reducir:12; afectar:12; quemar:11; aumentar:10	27.0
A2	hacer:122; ver: 98; saber:93; querer:76; hablar:42; creer:41; lograr:39; mantener:38; intentar:38; pensar:37	16.2
A31	poner:31; señalar:23; alcanzar:19; sacar:13; colocar:9; llevar:7; introducir:6; contemplar:6; situar: 6; tirar:6	60.0
A32	decir:205; dar:50; permitir:37; pedir:28; explicar:28; presenter:24; asegurar:23; afirmar:23; indicar:22; ofrecer:19	60.7
B1	llegar:69; caer:30; salir:29; entrar:26; ocurrir:23; aparecer:19; nacer:17; acudir:14; producir:13; registrar:13	37.1
B2	convertir:30; considerar:12; hacer:12; conducir:12; utilizar:10; conocer:10; llamar:10; ver:10; aumentar:9; abrir:9	8.8
C1	haber:120; estar:57; existir:38; vivir:33; tratar:23; morir:22; acabar:20; encontrar:17; referir:15; contar:13	48.6
C2	ser:1367; tener:209; estar:188; quedar:40; suponer:27; resultar:24; mostrar:20; llevar:19;ver:18; sentir:17	83.9
C4	parecer:64; servir:17; suceder:16; quedar: 14; gustar:10; pasar:9; importar:8; ir:6; llegar:5; faltar:5	75.8
D1	trabajar:32; ir:32; hacer:16; volver:15;pasar:13; echar:11; dar:10; actuar:10;regresar:10; huir:10;	37.1

Table 4: The 10 more frequent verbs for each semantic class in Spanish

It is important to highlight that nine of the thirteen Catalan semantic classes -A31, A32, C1, C2, C3, C4, D1, D2 and D3- cover, with the 10 more frequent lemmata, more than the 50% of the total amount of verbal occurrences in each class. In the case of the Spanish subset, the number of classes is eight -A31, A32, C1, C2, C3, C4, D2 and D3-. Only four classes in Catalan –A1, A2, B1 and B2- and five in Spanish –A1, A2, B1, B2 and D1- are below 50%, probably because they are also the classes with more different verbal lemmata, and more sparsely distributed too.

[7] Note that in this class are also included the passive constructions.

[8] We have not considered the C3, D2 and D3 semantic classes because they have less than 6 different lemmata per class.

LSS	Catalan verbs[9]	%
A1	provocar:139; afectar:103, obligar:79; millorar:61; convertir:60; augmentar:59; reduir:58; facilitar:56; incrementar:52; causar:40	38.7
A2	fer:910; voler:315; considerar:292; assegurar: 248; aconseguir:231; afegir:195; participar:194; rebre:192; recordar:189; assenyalar:183	36.1
A31	incloure:148; portar:94; publicar:53; acollir:52; recollir: 48; posar:43; plantejar:40; tenir_en_compte:37; incorporar:37; traslladar:36	58.7
A32	dir:585; explicar:358; demanar:338; presentar:261; permetre:216; donar:194; oferir:175; informar:146; anunciar:139; reclamar:104	60.9
B1	arribar:283; sortir:167; presentar :125; entrar:83; tenir_lloc:65; venir:61; caure:59; tornar:57; néixer:49; situar:48	42.1
B2	fer:328; produir:167; celebrar:112; relacionar:80; preveure:74; passar:73; convertir:72; pujar:68; baixar:64; situar:57	14.3
C1	haver:768; comptar:155; començar:125; tractar:119; ser.103; acabar:103; disposar:102; morir:99; servir:94; estar:93	54.8
C2	ser:4393; tenir:1296; estar:838; quedar:172; suposar:159; patir:102; viure:70; mostrar:68; formar_part:65; resultar:65	88.9
C3	costar:38; durar:37; portar:12; tardar:12; valer:10; fer:6; complir:5; prendre:4; guanyar:3; estar:2	92.8
C4	caldre:200; passar: 136; semblar:69; agradar:43; tocar:18; correspondre:17; interessar:12; faltar:11; costar:9; fer:6	92.8
D1	treballar: 162; anar:162; actuar:108; fer:76; passar:66; assistir: 50; viatjar:33; desplaçar:31; fugir:25; circular:23	64.3

Table 5: The 10 more frequent verbs for each semantic class in Catalan

4. Conclusions and Further Work

We have presented the lexicons AnCora-Verb-Ca and AnCora-Verb-Es, focusing on the content of the entries and the quantitative analysis of the data projected in AnCora corpora. The lexicons and the annotated corpora constitute the richest linguistic resources of this kind freely available for Spanish and Catalan. The big amount of linguistic information contained in both resources should be of great interest for computational applications and linguistic studies.

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

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

5. Acknowledgements

This research has been supported by the projects CESS-ECE (Corpus Etiquetados sintáctica y semánticamente para el Español, Catalán y Euskera) (HUM 2004-21127-E) PRAXEM (HUM2006-27378-E), Lang2World (TIN2006-15265-C06-06) from the Spanish Ministry of Education and Science, and the funding given by the government of the Generalitat de Catalunya. We would also like to thank all the annotators: Oriol Borrega, Núria Bufí, Joan Castellví, Maria Jesús Díaz, Silvia García, Difda Monterde, Aina Peris, Lourdes Puiggròs, Marta Recasens, and Bàrbara Soriano Bautista.

6. References

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

Martí, M.A., Taulé, M., Bertran M., Màrquez L. (2007). Anotación semiautomática con papeles temáticos de los corpus CESS-ECE. In *Procesamiento del Lenguaje Natural*, num. 38. Spain: SEPLN, pp. 67-77.

Martí, M.A., Taulé, M., Bertran M., Màrquez L. (2008). AnCora: Multilingual and Multilevel Annotated Corpora (pending to be published): http://clic.ub.edu/ancora/.

M. Palmer, Kingsbury P., Gildea D. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 21 (1). USA: MIT Press.

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

Taulé, M., Martí, M.A , Recasens, M. (2008). AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of Language, Resources and Evaluation*, LREC 2008. Marrakech, Morocco.

[9] We have not considered the D2 and D3 semantic classes because they have less than 6 different lemmata per class.