



Universidad
del País Vasco

Euskal Herriko
Unibertsitatea

Analysing specialized use of verbs: the case of some Catalan trajectory verbs in the sport domain

Author: Mar Rodríguez i Alvarez

Advisors: Izaskun Aldezabal Roteta and Itziar Gonzalez-Dios

hap/lap

Hizkuntzaren Azterketa eta Prozesamendua
Language Analysis and Processing

Final Thesis

June 2022

Departments: Computer Systems and Languages, Computational Architectures and Technologies, Computational Science and Artificial Intelligence, Basque Language and Communication, Communications Engineer.

Abstract

This master thesis presents a syntactic-semantic analysis of 5 Catalan verbs of trajectory class (*aconseguir* (to achieve), *arribar* (to arrive), *avançar* (to move forward), *entrar* (to enter) and *sortir* (to come out)) in the sport domain. The master thesis starts from the hypothesis based on the Communicative Theory of Terminology (CTT) (Cabr , 2001; Adelstein and Cabr , 2003) and the study done by Casademont (2014) about specialized uses of verbs. To perform the analysis of verbs, reliance has basically been placed on the book *Clasificaci3n verbal. Alternancias de di tesis* (V zquez et al., 2000) and on the AnCora’s verbal lexicon (Taul  et al., 2011). The analysis has been done on the data of a sport news corpus, actually created for this study. The results of this study have shown that some verb senses appear most often in the sports domain than others, they have also shown that some specific structures exist for the sport domain and that arguments referring to the trajectory can present ellipsis for some verbs. In addition, in this master thesis, a Natural Language Processing task on event argument extraction (EAE) has been carried out too. Due to lack of resources and that the model used was not trained for Catalan, no reliable results have been achieved.

Laburpena

Master-tesi honetan kokapen-aldaketa motako 5 aditz katalan aztertu dira kirol domeinuan: *aconseguir* (lortu), *arribar* (iritsi), *avançar* (aurrera egin), *entrar* (sartu) eta *sortir* (atera). Azterketa honen abiapuntua Terminologiaren Teoria Komunikatiboa (Cabr , 2001; Adelstein and Cabr , 2003) eta Casademont-ek (2014) aditzen erabilera espezializatuei buruz egindako azterketa dira. Horietan oinarrituta, aditzen analisisa egiteko, *Clasificaci3n verbal. Alternancias de di tesis* (V zquez et al., 2000) liburua eta katalanezko AnCora lexikoa hartu dira oinarri gisa (Taul  et al., 2011). Azterketa egiteko, kirolari buruzko corpus bat bildu da. Azterketaren emaitzek erakutsi dute kokapen-aldaketa aditzen adiera batzuk maizago agertzen direla kirol-arloan beste batzuk baino. Orobat, emaitzek erakutsi dute, batetik, badirela kirol-arloan berariazkoak diren egitura batzuk, eta, bestetik, kokapen-aldaketa adierazten duten argumentuek elipsia izan dezaketela zenbait aditzetan. Horretaz gain, hastapeneko esperimentu bat egin da Hizkuntzaren Prozesamenduko ataza batean, ebentuen argumentuen erauzketan hain zuzen ere. Nolanahi ere, baliabide faltagatik eta erabilitako ereduak ez delako katalanarentzat entrenatu, ez da emaitza fidagarririk lortu.

Contents

1	Introduction	1
2	Literature Review	3
2.1	Previous work on the analysis of language in the sport’s domain	3
2.2	Specialized knowledge	3
2.2.1	Specialized knowledge in verbs	5
3	Basis for our work	8
3.1	Class delimitation of trajectory verbs	8
3.1.1	Verbs of displacement and movement	8
3.1.2	Verbs of displacement and location	9
3.1.3	Verbs of change of possession and communication	9
3.2	Syntactic and semantic characteristics of trajectory verbs	9
3.2.1	Eventive structure and focalization	11
3.2.2	Alternations of diathesis, underspecification and ellipsis	12
3.3	AnCora-Verb 2.0 lexicon and its annotation criteria	14
4	Methodology	16
4.1	Description of the corpus	16
4.2	Verbs’ selection	20
4.3	AnCora-Verb 2.0’s verb senses for <i>aconseguir</i> , <i>arribar</i> , <i>avançar</i> , <i>entrar</i> and <i>sortir</i>	21
4.4	Process of the analysis of the verbs	22
5	Results and Discussion	26
5.1	Results of the structures of the verb <i>aconseguir</i>	26
5.2	Results of the structures of the verb <i>arribar</i>	29
5.3	Results of the structures of the verb <i>avançar</i>	34
5.4	Results of the structures of the verb <i>entrar</i>	37
5.5	Results of the structures of the verb <i>sortir</i>	40
6	Preliminary experiment: Applying the analysis to information extraction	43
7	Conclusion and Future Work	49
A	List of Catalan trajectory verbs	56
B	AnCora’s correspondences between Arguments, Thematic Roles and Syntactic Functions	57

C	Ancora's structure proposals for <i>aconseguir</i>, <i>arribar</i>, <i>avançar</i>, <i>entrar</i> and <i>sortir</i>	58
C.1	Aconseguir	58
C.2	Arribar	59
C.3	Avançar	61
C.4	Entrar	63
C.5	Sortir	65
D	Tables of results	66
D.1	Tables of results of the verb <i>aconseguir</i>	66
D.2	Tables of results of the verb <i>arribar</i>	67
D.3	Tables of results of the verb <i>avançar</i>	69
D.4	Tables of results of the verb <i>entrar</i>	70
D.5	Table of results of the verb <i>sortir</i>	71

List of Figures

1	Appearance of the raw text of the corpus.	18
2	Appearance of the corpus automatically annotated.	18
3	Appearance of the corpus manually annotated.	19
4	Frequency of grammatical categories in the corpus.	20
5	Example of annotation of the corpus.	25
6	Arguments structures with a longer appearance than 4 times of the verb <i>aconseguir</i>	27
7	Syntactic structures with a longer appearance than 4 times of the verb <i>aconseguir</i>	27
8	Thematic structures with a longer appearance than 4 times of the verb <i>aconseguir</i>	28
9	Arguments structures with a longer appearance than 4 times of the sense 1 of the verb <i>arribar</i>	30
10	Syntactic structures with a longer appearance than 4 times of the sense 1 of the verb <i>arribar</i>	31
11	Thematic structures with a longer appearance than 4 times of the sense 1 of the verb <i>arribar</i>	31
12	Arguments structures with a longer appearance than 4 times of the sense 2 of the verb <i>avançar</i>	35
13	Syntactic structures with a longer appearance than 4 times of the sense 2 of the verb <i>avançar</i>	35
14	Thematic structures with a longer appearance than 4 times of the sense 2 of the verb <i>avançar</i>	36
15	Arguments structures with a longer appearance than 4 times of the sense 1 of the verb <i>entrar</i>	38
16	Syntactic structures with a longer appearance than 4 times of the sense 1 of the verb <i>entrar</i>	38
17	Thematic structures with a longer appearance than 4 times of the sense 1 of the verb <i>entrar</i>	39
18	Arguments structures with a longer appearance than 4 times of the sense 1 of the verb <i>sortir</i>	40
19	Syntactic structures with a longer appearance than 4 times of the sense 1 of the verb <i>sortir</i>	41
20	Thematic structures with a longer appearance than 4 times of the sense 1 of the verb <i>sortir</i>	41
21	AnCora's correspondences between Arguments, Thematic Roles and Syntactic Functions.	57
22	Ancora's structure proposal for the active voice of the verb <i>aconseguir</i> . . .	58
23	Ancora's structure proposal for the passive voice of the verb <i>aconseguir</i> . .	58
24	Ancora's structure proposal for the active voice of sense 1 of the verb <i>arribar</i>	59

25	Ancora's structure proposal for the impersonal voice of sense 1 of the verb <i>arribar</i>	59
26	Ancora's structure proposal for the active voice of sense 2 of the verb <i>arribar</i>	59
27	Ancora's structure proposal for the impersonal voice of sense 2 of the verb <i>arribar</i>	59
28	Ancora's structure proposal for the active voice of sense 3 of the verb <i>arribar</i>	59
29	Ancora's structure proposal for the impersonal voice of sense 3 of the verb <i>arribar</i>	59
30	Ancora's structure proposal for the active voice of sense 4 of the verb <i>arribar</i>	60
31	Ancora's structure proposal for the passive voice of sense 4 of the verb <i>arribar</i>	60
32	Ancora's structure proposal for the active voice of sense 5 of the verb <i>arribar</i>	60
33	Ancora's structure proposal for the active voice of sense 6 of the verb <i>arribar</i>	60
34	Ancora's structure proposal for the active voice of sense 7 of the verb <i>arribar</i>	60
35	Ancora's structure proposal for the active voice of sense 1 of the verb <i>avançar</i>	61
36	Ancora's structure proposal for the causative voice of sense 1 of the verb <i>avançar</i>	61
37	Ancora's structure proposal for the active voice of sense 2 of the verb <i>avançar</i>	61
38	Ancora's structure proposal for the benefactive voice of sense 2 of the verb <i>avançar</i>	61
39	Ancora's structure proposal for the impersonal voice of sense 2 of the verb <i>avançar</i>	61
40	Ancora's structure proposal for the passive voice of sense 2 of the verb <i>avançar</i>	61
41	Ancora's structure proposal for the active voice of sense 3 of the verb <i>avançar</i>	62
42	Ancora's structure proposal for the active voice of sense 4 of the verb <i>avançar</i>	62
43	Ancora's structure proposal for the active voice of sense 5 of the verb <i>avançar</i>	62
44	Ancora's structure proposal for the active voice of sense 6 of the verb <i>avançar</i>	62
45	Ancora's structure proposal for the active voice of sense 1 of the verb <i>entrar</i>	63
46	Ancora's structure proposal for the impersonal voice of sense 1 of the verb <i>entrar</i>	63
47	Ancora's structure proposal for the active voice of sense 2 of the verb <i>entrar</i>	63
48	Ancora's structure proposal for the active voice of sense 3 of the verb <i>entrar</i>	63
49	Ancora's structure proposal for the impersonal voice of sense 3 of the verb <i>entrar</i>	63
50	Ancora's structure proposal for the active voice of sense 4 of the verb <i>entrar</i>	63
51	Ancora's structure proposal for the passive voice of sense 4 of the verb <i>entrar</i>	64
52	Ancora's structure proposal for the active voice of sense 5 of the verb <i>entrar</i>	64
53	Ancora's structure proposal for the active voice of sense 6 of the verb <i>entrar</i>	64
54	Ancora's structure proposal for the active voice of sense 7 of the verb <i>entrar</i>	64
55	Ancora's structure proposal for the active voice of sense 1 of the verb <i>sortir</i>	65
56	Ancora's structure proposal for the impersonal voice of sense 1 of the verb <i>sortir</i>	65
57	Ancora's structure proposal for the active voice of sense 2 of the verb <i>sortir</i>	65
58	Ancora's structure proposal for the active voice of sense 3 of the verb <i>sortir</i>	65

59	Ancora's structure proposal for the active voice of sense 4 of the verb <i>sortir</i>	65
60	Ancora's structure proposal for the active voice of sense 5 of the verb <i>sortir</i>	65

List of Tables

1	Dates of news' extraction and publication for each newspaper.	16
2	Total number of news items per newspaper and its location within the corpus marked by ordinal numbers	19
3	Mapping of VerbNet labels for verb senses and the labels used in this study.	23
4	Proposal of the canonical structure of the verb <i>aconseguir</i> in the specialized field of sports news.	29
5	Proposal of the canonical structure of the sense 1 of the verb <i>arribar</i> in the specialized field of sports news.	33
6	Proposal of the canonical structure of the sense 2 of the verb <i>arribar</i> in the specialized field of sports news.	34
7	Proposal of the canonical structure of the sense 2 of the verb <i>avançar</i> in the specialized field of sports news.	37
8	Proposal of the canonical structure of the sense 1 of the verb <i>entrar</i> in the specialized field of sports news.	39
9	Proposal of the canonical structure of the sense 1 of the verb <i>sortir</i> in the specialized field of sports news.	42
10	Entailment probability of verbalizations containing the argument <i>patient</i> for the verb <i>aconseguir</i>	45
11	Entailment probability of verbalizations containing the argument <i>destination</i> for the verb <i>arribar</i>	45
12	Entailment probability of verbalizations containing the argument <i>agent</i> for the verb <i>avançar</i>	46
13	Entailment probability of verbalizations containing the argument <i>location</i> for the verb <i>entrar</i>	46
14	Entailment probability of verbalizations containing the argument <i>origin</i> for the verb <i>sortir</i>	47
15	Quantitative results of the verb <i>aconseguir</i>	66
16	Quantitative results of the verb sense 1 of <i>arribar</i>	67
17	Quantitative results of the verb sense 2 of <i>arribar</i>	68
18	Quantitative results of the verb sense 3 of <i>arribar</i>	68
19	Quantitative results of the verb sense 1 of <i>avançar</i>	69
20	Quantitative results of the verb sense 2 of <i>avançar</i>	69
21	Quantitative results of the verb sense 1 of <i>entrar</i>	70
22	Quantitative results of the verb sense 2 of <i>entrar</i>	70
23	Quantitative results of the verb sense 1 of <i>sortir</i>	71

1 Introduction

This master thesis presents a syntactic-semantic analysis of 5 Catalan verbs of trajectory class (*aconseguir* (to achieve), *arribar* (to arrive), *avançar* (to move forward), *entrar* (to enter) and *sortir* (to come out)) in the sport domain. As Cabré i Castellví (1999); Cabré (2001); Cabré and Adelstein (2001); Adelstein and Cabré (2003) say, a lexical unit is not a word or term *per se*, but as a lexical form where the semantic information is organised into features and modules or feature packages, considered as *macro-senses*. Depending on communicative situation the lexical unit is used, some or other features are activated, giving as a result the specific senses typically found in dictionaries. Within this framework Adelstein et al. (2007) studies the activation of certain semantic features in nouns and Casademont (2014), under the same hypothesis, studies the activation of semantic features in Catalan verbs.

As in Casademont (2014), in this study verbs for Catalan have been analysed. Catalan is a Romance language of the Occitano-Romance branch, sometimes considered a minoritised language because of its unequal power relationship with major languages with which it shares territory. The current amount of usual Catalan speakers in the Catalan-speaking territories (Valencian Country, Franja de Ponent, Catalonia, Northern Catalonia, Andorra and Alghero) is only the 32.4% of the territories' population (4,517,706 speakers) (Melero et al., 2022). Melero et al. (2022) describe the current situation of the Languages Technologies for Catalan and in their report it can be seen that despite the socio-linguistic situation of Catalan, several language resources and tools for this language can be found. Still, the number is not high enough.

Unlike Casademont (2014), who covers more than one domain (*Genomics* and *Economics*) with the aim of extracting common generalisations across domains, this study relies on a single domain: *Sport*. This is due to the fact we want to observe specific features of this particular domain. Sport and the sport press move a lot of money. In addition, artificial intelligence (AI) techniques are increasingly being used. Thus, linguistic analysis in a sports domain which help to create Natural Language Processing (NLP) tools can be very necessary. For these purposes, we have created a Catalan sport news corpus.

For the present study, those lexical units or lemmas which, in principle, prototypically evoke a general notion equivalent to a semantic class, such as verbs of change of location (or verbs of trajectory) have been chosen. For that, we have selected some trajectory verbs, according to Vázquez et al. (2000), and we have proposed a manual analysis of the verbs and their arguments, following Vázquez et al. (2000) and Taulé et al. (2011).

Finally, to see if these analysis can help to perform tasks on Information Extraction (IE), exactly on event (Event Argument Extraction), we have performed a preliminary experiment on a set of templates to capture the vent argument roles, based on Sainz et al. (2022). Information Extraction refers to automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from structured sources Sarawagi (2008). Thus, Event Argument Extraction (EAE), aims to identify the entities serving as event arguments and classify the roles they play in an event Wang et al. (2019).

With all this in mind, our specific aim in this work is to analyse how some verbs of trajectory behave in a specialized corpus of sport. And our sub-objectives are

- to compare the results of the analysis of the chosen verbs with the description found for them in Ancora's lexicon (Taulé et al., 2011) and to see if there is any remarkable tendency in the field of sport with respect to general "trajectory" use (to see if there is something missing or with which we do not agree very much).
- to analyse the phenomenon of ellipsis (of the verb arguments in the specialised corpus)
- to see if the analysed structures can help to perform tasks on event argument extraction (EAE).

The contents of this study are organized as follows: Section 2 contains the previous work on the sport's domain, and on the specialized knowledge in verbs; Section 3 contains basic works on the description of the trajectory verbs class, and the description of AnCora's lexicon and AnCora's criteria of annotation; Section 4 contains the methodology followed to create the corpus, to select the verbs for the analysis and the methodology followed to annotate the corpus and analyse the verbs; Section 5 contains the results and the discussion of the analysis; Section 6 contains the preliminary experiment on EAE; and, finally, Section 7 contains the conclusions of the study and proposals for future work.

2 Literature Review

In this section previous work related to the analysis of language in the domain of sport and previous work related to specialized knowledge and specialised knowledge in verbs are exposed.

2.1 Previous work on the analysis of language in the sport's domain

When looking at previous work related to the analysis of language in the domain of sport, most commonly encountered are works related to the genre. Omrčen (2017), for example, analysed the usage of gender-fair language in competitive sport and in the scientific journals related to sport and exercise topics; and Segrave et al. (2006) described differences in the ways in which many different people speak differently about men's and women's sport and male and female athletes.

On the other hand, when looking at works on Natural Language Processing (NLP) in the same domain, works related to sentiment analysis (Ljajić et al., 2015) and sports news generation are more frequent. Two works related to sports news generation are Belemkoabga et al. (2021), that proposed a global summarizing method for live sport commentaries, and Li et al. (2019), that used Generative Adversarial Networks (GAN) architecture to create a sports news generator that produces sport news immediately after each match is over.

We want to point out that we have not found previous studies related to language structures in a sports context. Moreover, we have not found data-sets in Catalan for this domain either.

2.2 Specialized knowledge

As it has been said in Section 1, this master thesis starts from the hypothesis based on the Communicative Theory of Terminology (CTT) (Cabré i Castellví, 1999; Cabré, 2001; Cabré and Adelstein, 2001; Adelstein and Cabré, 2003) where is claimed that:

“The lexical unit is not per se a word or term, but a lexical form. According to the communicative situation in which it is used, activates a specialised value or not. Semantic information, according to Cabré, is organised into features and modules or feature packages. They are activated in different ways according to each situation, leading to one of the possible senses of the Lexical Unit (LU). The CTT explanation that lexical semantic information is organised in activatable features and modules of features allows to account, not only for the distinction between specialised and non-specialised values, but also for all the semantic variation related to a lexical form.”(Adelstein et al., 2007, p. 91)

These features or modules are considered *macro-senses*. They are made up of different types of semantic elements that maintain different types of relationships with each other. Consequently, all senses relating to a form (the typical dictionary list of senses) are particular activations of certain elements of the *macro-sense*. Thus, the activation of a specialised

value would consist in an activation of certain semantic elements of the LU. The elements of the semantic information of the lexical entry have the nature of semantic predicate and argument. Their possible combinations constitute each of the senses of the LU, regardless of the fragmentation degree of the description or analysis.

Adelstein et al. (2007) postulate that

“(i) *the multiplicity of diverse senses of a LU can be reduced to two or three basic notions*” (we understand these notions to be the previously mentioned *macro-senses* or feature packages).

“(ii) *the semantic information of a LU has a basic semantic structure: It consists of a basic predicate and one or more arguments; thus, each sense of the LU would constitute a different way of specifying the basic predicate and its arguments.*” (p. 101)

Adelstein (2001) based his work on nouns, specialised lexical units (terms) per se. She carried out a meticulous study of the specialised semantic information of the lexical units *mother*, *head*, *cup* and *family*. From these lexical units, empirical evidence was presented for the lexical unit hypothesis formulated by Cabré (1999). For example:

- for *mother* defines two basic notions which are: (1) mother as progenitor, and (2) mother as origin
- or for *head*, a single basic generalising notion: *part of* (meronymic relation between two elements), of which their form and nature are relevant: x is a part a (which has a form c and a location d in the whole) of a whole b (which is of a certain nature e and has a form f).

All the senses listed in the dictionaries would be concretisations of this notion, focusing on one or the other aspect according to the context. Thus, in his work he presented the most relevant observations about the semantic information of the lexical entry, the conditions and properties of activation of lexical senses, and, finally, the properties of specialised senses.

Under the same hypothesis, but focusing on the verbal lexical unit, Casademont (2014) also carried out an exhaustive work on lexical units or lemmas of Catalan in the domain of Genomics and Economics. As Casademont (2014) stressed, “*verbs by their relational conditions, would then seem to be less subject than nouns to present terminological values on their own, and therefore the possible specialized behaviours of a verb would seem more difficult to grasp unless its interaction with the surrounding text is considered*” (p. 93). In her study, she leaved aside the verbalization of noun terms, such as *crystallize*.

As Casademont (2014) specified, “*when aiming at analysing the possible terminological use of verbs, it will be necessary to observe characteristics of the verb itself, properties of its arguments and elements linking the verb and its argument*” (p.94). This characterisation is based on lexicalist approaches such as Levin (1993 and 2000) or Pustejovsky (1995). In these approaches, the verbal class is understood as a grouping of lexical units evoking the same general predicate or general cognitive value, that we relate to the *macro-sense*, mentioned above for nouns.

Indeed, this notion of verbal class is not trivial, and we have chosen and worked precisely with sources which in our opinion are more based on these theoretical notions. Thus, as we will see in Section 3, for the verbal class characterisation we have relied on:

- *Clasificación verbal. Alternancias de diátesis* (Vázquez et al., 2000) and
- AnCora's verbal lexicon (Taulé et al., 2011), which is largely designed on the basis of Vázquez et al. (2000), and which contains argumental information, thematic roles and semantic features, among other characteristics.

2.2.1 Specialized knowledge in verbs

Specialized knowledge in verbs has been studied by L'Homme (1998), Lerat (2002), Lorente Casafont (2001, 2002a, 2002b, 2007), Pimentel (2012). However, as far as we are aware, in Catalan just the work of Casademont (2014) is carried out on the specialized knowledge of verbs in a certain domain.

As it has been introduced above, Casademont (2014) proposed an analysis of phraseological verbs¹ in specialized discourse in order to establish syntactic and semantic characteristics likely to indicate the transmission of specialized knowledge. Casademont (2014), for this study, considered elements from the relation between the verb and its arguments, such as argument structure and thematic roles, and aspects from the arguments themselves, such as semantic features and termhood.

The domains studied in their research were Genomics and Economics. Casademont (2014) chose lemmas that could appear in very different domains to ensure that the specialized meaning was detected based on the cooperative use of the context in which the verb appears.

The occurrences of the corpus were analysed according to 10 different levels of description:

- **Adicity:** The number of arguments a function takes. It can be applied to syntax by showing if some verbs always follow the same argument structure, if the use of one structure is domain-dependent, and if sentences are the result of canonical projections or of marked forms (Levin, 1993, 2000, 2004).
- **Subcategorization of the constituents:** The assignment of a lexical item to a subclass of its part of speech, especially with respect to the syntactic elements with which it can combine (Matthews, 2014).

¹Phraseological verbs express an action, a change or a cause of change. They have a dyadic or monadic structure (two or one arguments) and their accompanying noun group is a relevant term of the domain. According to fixation, they can be phraseological units or collocations. They convey specialized knowledge when combined with other units (Lorente, 2007).

- **Thematic Role:** The part that a unit plays in a larger structure (Matthews, 2014). In order to observe this, Casademont (2014) took into account Dowty's, (1991) proposal. This proposal follows the idea of a continuous non-compartmented classification where thematic roles are ordered according to their prototypicality. This classification is framed by two basic poles: *proto-agent* and *proto-patient*.
- **Semantic Selection:** Bloomfield's term for the feature of a construction by which certain individual words or morphemes enter into it (Matthews, 2014). Casademont (2014) used semantic features of lexicon units based on Pustejovsky (1995).
- **Termhood (T1, T2, T3):** It refers to the status of index terms as units of representative of document contents (Kageura and Umino, 1996).
- **Semantic Class:** Any class of word established by similarities in semantics (Matthews, 2014). The semantic classes in Casademont (2014) are based on SIMPLE Lenci et al. (2000). On the other hand, in our work, we take Vázquez et al. (2000) and Taulé et al. (2011) as reference for establishing the semantic classes, because they seem to us to be more appropriate.
- **Syntactic-semantic Pattern:** It can express an identificative state, a stative place, a judgment, a perception, an action with a goal, a communicative action, a natural transition, a cause of change of value, a creation, etc. (Casademont, 2014). The syntactic-semantic patterns used in her study to select the lemmas express an event and have dyadic or monadic structure (two or one arguments).
- **Productivity according to domain:** It is not specified in the work of Casademont (2014). However, it might refer to when the phenomena *change* and *cause of change* occur in each of the domains studied in her work.
- **Sense:** It is the meaning of a lexical unit (Matthews, 2014). Casademont (2014) marks it with capital letters. Each letter is assigned to a definition of the lemma taken from reference dictionaries.

With this study, Casademont (2014) was able to see that the transmission of specialized knowledge of a verb could slightly vary according to its type. For verbs expressing events, in monadic occurrences, the verb and the argument corresponding to the subject act as a minimal construction. In the case of dyadic occurrences, the relevant group would be the verb and the argument acting as the first complement of the verb, and in the triadic occurrences another internal argument is also added to the relevant minimal construction. Thus, the transmission of specialized knowledge in verbs expressing an event is done through the verb and the argument being its subcategorized constituent. On the other hand, for verbs expressing stative situations, the marks of transmission take place between the verb and the two complements being linked by it. The construction of these three elements would be then the minimal construction to convey specialized knowledge. Thus, the transmission

of specialized knowledge in verbs expressing a stative condition takes place through the verb and all the arguments.

Casademont (2014) could validate that there are elements in the syntactic projection of verbs and their arguments that show the activation of specialized value. This is because the syntactic and/or semantic characteristics in the occurrences show the activation of terminological value in some realizations compared to others of the same verb. However, she observed that not all the elements considered entail the activation of terminological value in all cases. Moreover, they do not always have the same influence on the transmission of specialized knowledge. The most important element would be the semantic selection and sometimes even the lexical selection of a unit or a group of units.

Finally, their results confirmed that a phraseological verb transmits specialized knowledge by the cooperation of the units appearing in its context. Thus, it is not an item just appearing randomly in discourse, but highly interacting with the units expressed in its context.

As it has been said in 1, in our study, we start from Casademont's (2014) conclusions. Thus, as it will be showed in Subsection 4.4, the elements selected for the analysis of our study correspond to those elements showing relevant results on the transmission of specialized knowledge in Casademont (2014).

3 Basis for our work

As pointed out in Section 1, the verbs analysed in this study belong to the class of trajectory verbs. For the delimitation of the class we base mostly on Vázquez et al. (2000), since we agree with the point of view they adopt for the analysis of verb classes. Moreover, Vázquez et al.'s (2000) point of view is in line with the main semantic characterization seen in Adelstein et al. (2007) and Casademont (2014).

The following subsection (3.1) is a collection of the aspects and discussions of Vázquez et al.'s (2000) work that, in our opinion, have most relevance for our study and our final decisions.

On the other hand, for the representation of the verbs, we have used the lexicon described in Ancora-Verb 2.0 (Taulé et al., 2011). In Subsection 3.3 the description of this lexicon can be found.

3.1 Class delimitation of trajectory verbs

Trajectory verbs are defined as verbs that express the movement of an object. The semantics of these verbs denotes a process that results in the change of location of an entity from a point of origin to a destination. This change of location is defined in a path that corresponds to the trajectory component. The diatheses² that accept these predicates allow the different subcomponents associated with it to be omitted or expressed (Vázquez et al., 2000).

Trajectory includes both physical and abstract displacements of entities. The latter refer to changes of possession (*to buy, to sell*) or communicative exchanges (*to say, to answer*) (Vázquez et al., 2000).

3.1.1 Verbs of displacement and movement

The concepts of displacement and movement can be distinguished from each other depending on whether there is an orientation or not respectively. Some predicates such as *to run* are difficult to be classified in one class or the other, since lexically they are a type of predicate that does not include the limit of the route.

Jackendoff (1990) and Vázquez et al. (2000) differentiate predicates that have a path argument such as *to go* or *to run*, which are represented by *to go*; from those that do not, such as *to dance*, which is associated with *to move*. With respect to the latter, the displacement that can be carried out is secondary. The focus of these predicates is on movement.

²According to Vázquez et al. (2000), the diathesis is “*the syntagmatic expression of different semantic oppositions*” (p. 90).

3.1.2 Verbs of displacement and location

A differentiation between location and movement/displacement should be notice since stationary location predicates “do not describe a motion, but rather the location of an object relative to some other object” (Jackendoff, 1976 (p. 95)).

Related to this question, verb such us Spanish *poner* (to put) need to be taken into account. Some authors (J.-P. Boons (1976), Lamiroy (1991) and Cifuentes Honrubia (1999)) consider that the predicate *poner* (to put) does not express a direction but a change of possession. However, it has to be noticed that what characterises this type of verbs is its aspectual content, since it focuses on the end point of the trajectory. Thus, these verbs should also be considered within the class of verbs of trajectory.

3.1.3 Verbs of change of possession and communication

Vázquez et al. (2000) include the verbs of change of possession and communication in the class of trajectory because they share the same basic characteristics, both syntactic and semantic, as the verbs of displacement. Moreover, the relationship between source and recipient in a transfer event (*to give, to buy*) is equivalent to the relationship between source and target in a displacement (*to go, to come*): in both cases there is an initial position, an object and a final position.

3.2 Syntactic and semantic characteristics of trajectory verbs

As for the semantic decomposition of trajectory verbs, Vázquez et al. (2000) identify three meaning components: the entity, the initiator and the trajectory. The entity corresponds to the element *figure*, described by Talmy (1985) and to the element *trajector*, described by Langacker (1987). Vázquez et al. (2000) say that, prototypically, the entity is represented by a noun phrase (NP). In the following sentence (1) the entity of a sentence containing a trajectory verb can be seen:

- (1) El cartero lleva *las cartas* a sus destinatarios.

The postman delivers *the letters* to their addressees. (Vázquez et al. (2000) p. 183)

The initiator is defined as the cause of the displacement and it is typically agentive in character, but it can also be passive, as in *receiving*. The initiating component can be performed autonomously or it can coincide with some other component, such as, for example, the displaced entity. The latter situation occurs when the entity moves autonomously, i.e. with prototypical displacement verbs. One example of sentence is 2:

- (2) *El conserje* va a la sala de estudios.

Caretaker goes to the study room. (Vázquez et al. (2000) p. 184)

Another possibility is that the initiator coincides in the subject position with some point on the trajectory, either the *destination* or the *origin* of the route. The trajectory

component is the path, real or abstract, experienced by a moving entity. Trajectory can be understood as a realisation of the space component, such as in autonomous (*to run, to go*) and non-autonomous (Spanish *trasladar* (to transfer), *carretear* (to haul)) displacement verbs. The trajectory component in these verbs is represented by locative complements (3).

- (3) Ha ido *a Valencia* este fin de semana.
He/she went *to Valencia* this weekend. (Vázquez et al. (2000) p. 184)

On the other hand, the predicates of change of possession or transfer (*to give, to buy*) and communication predicates (*to say, to speak*) require animate entities as the beginning and end of the trajectory. Thus, the trajectory is not understood as a subspecification of the space component, but it is interpreted in a more abstract way as in Example 4.

- (4) Sofía ha dado el regalo a Juan.
Sofía has given the gift to Juan. (Vázquez et al. (2000) p. 185).

Vázquez et al. (2000) propose that both the *origin* (To) and the *end* (Td) of the path along which the entity moves are part of the trajectory component. Within this, the route subcomponent (Tr) and the direction subcomponent (Tdir) are also distinguished. They consider that the expression of one of these subcomponents is enough to consider that the path component has a syntactic realisation. However, they point out that there are certain incompatibilities in the realisation of these subcomponents. For example, the expression of the *destination* cannot go together, in the same sentence, with the *direction*. Similarly, some verbs do not allow the expression of the *origin* and the *direction* in the same sentence. The two types of incompatibility mentioned are typical of predicates such as *to go*. Finally, some verbs also cannot express the *destination* together with the *origin*.

Vázquez et al. (2000) note that in Spanish when the *origin* is made with a prepositional phrase (PP), the preposition used is *de* and sometimes *desde*. The prototypical prepositions for the expression of the *destination* of the trajectory are *a*, *en* or *hasta*. Regarding the *route* subcomponent, it is expressed by a SP typically introduced by the preposition *por* (5).

- (5) Miguel ha ido *de Madrid* (To) *a Barcelona* (Td) *por Valencia* (Tr).
Miguel has gone *from Madrid* (To) *to Barcelona* (Td) *via Valencia* (Tr). (Vázquez et al. (2000) p. 185)

Direction can be expressed syntactically through a PP introduced by the preposition *hacia*, the prepositional locution *en dirección a*, or it can also be embedded in the verb (*subir, avanzar, o caer*) (Dervillez-Bastuji (1982) and Boons (1987)). Trajectory can also be expressed through a NP that indicates the totality of the displacement, either by specifying the type of territory travelled or by its spatial measures (6) (Vázquez et al., 2000).

- (6) Los atletas corrieron *40 km*.
Athletes run *40 km*. (Vázquez et al. (2000) p. 187)

Another feature of the trajectory verbs is that they often have the *manner* component incorporated in them. These verbs are hyponyms of general verbs of displacement, such as Spanish *desplazarse* (to travel/to move) or *recorrer* (to travel/to go through), which also do not incorporate trajectory boundaries (Vázquez et al., 2000).

3.2.1 Eventive structure and focalization

On the basis of a two-argument structure (one of the arguments being the *agent*), the subcategorized argument tends to increase in importance (focalization) and progressively takes the place of the *agent*. At the same time, the intervention in the discourse of the agentive external argument of the canonical structure progressively decreases, the occurrences mainly moving from a cause of change to a change. The way in which language uses its resources to express focalizations is also made in some cases from change towards cause of change. In most cases, the aim of focalization is to focus the internal argument of the canonical structure and place it in a prominent position in the marked structure. It must be noted that focalization also works as an impersonalization resource, aiming at creating an objective discourse, typical of specialized texts (Casademont, 2014).

According to Rappaport and Levin (1992), not all verbs of movement share the same eventive structure. Thus, verbs such as *to arrive* differ from predicates such as *to run* in that they contain the direction component and are telic³. In contrast, *to run* incorporates the *manner* component but not the direction, moreover it is atelic. However, *to run* can behave like *to arrive* when it is complemented by a phrase expressing the *destination*. Vázquez et al. (2000) state that within the class of trajectory, the same aspectual subclassification can be established for verbs of non-autonomous movement. Thus, verbs such as Spanish *arrastrar* (to drag) present the same essential characteristics as those represented by *to run*. They only differ in the fact that in the latter the displacement is carried out autonomously and in the first one an external causer is required. Likewise, among the transitive verbs involving displacement, there are verbs (*to give*) which are telic, since they also have a built-in subcomponent of the trajectory (Serzisko, 1988). Boons (1987) and Laur (1993) say that the type of boundary lexicalised by trajectory verbs needs to be considered too. Thus, it can be differentiate between predicates that focus the point of origin (*to leave, to move away, the destination (to land, to come)*) or the route (understood as the midpoint of the journey). Among the verbs that focus the *origin* (7) or the *end* (8), different degrees of focus can be observed, depending on whether the emphasis is placed only on the boundary or also on the movement itself. Thus, while Spanish *partir* (to depart) (*origin*) and *llegar* (to arrive) (*end*) focus very precisely on the incorporated boundary, in other verbs, such as Spanish *alejarse* (to move away) or *to go*, the focus on the boundary (*origin* and *end*, respectively) is minor (Vázquez et al., 2000).

(7) Partieron *desde el aeropuerto*.

They departed *from the airport*.

³Aspect distinguishing events that have a limit (Matthews, 2014).

(8) Han llegado *a la ciudad*.

They have arrived *at the city*.

Vázquez et al. (2000) consider that what characterises trajectory verbs as a whole is that they can express complex events, formed by two sub-events: one that expresses the process and another that expresses a point in the trajectory, i.e. a state. However, this complexity is given by different circumstances. In the case of trajectory verbs, the characteristic of telicity can be obtained from two different procedures: either it is inherent to them by nature, i.e. they are lexically formed by a process and a sub-event of telic type (*to arrive*, Spanish *dirigirse* (to head), *to give*); or they are simple events that can be bounded by a phrase associated with the path, thus becoming complex events (*to swim* vs. *to swim to the buoy*).

Among the verbs that are telic in nature, there are verbs that focus the end point (*to arrive*, *to put*, *to enter*) or both the boundary and the process (*to go*, *to come*). Others, however, focus the *origin*, and as in the previous case, the focus may be on the limit (*to leave*), which in this case precedes the process, or both sub-events may be focused on (Spanish *alejarse*, *huir* (to move away, to flee)) (Vázquez et al., 2000).

For verbs of change of possession, of transfer and verbs of communication (*to give*, *to send*, *to say*), it is postulated that, from a lexical point of view, both the process and the boundaries are equally focused. Thus, syntactic construction of the elements corresponding to process and the boundaries will determine which element is more relevant (Vázquez et al., 2000).

3.2.2 Alternations of diathesis, underspecification and ellipsis

All verbs of trajectory can have underspecification, “*the deliberate omission of information from linguistic descriptions to capture several alternative realisations of a linguistic phenomenon in one single representation*” (Egg, 2010). Thus, in structures governed by verbs of trajectory, complements expressing one or more of the subcomponents of the trajectory can be omitted Vázquez et al. (2000).

The subcomponents *origin*, *destination* and *address* When the verb can syntagmatically express different subcomponents of the path, there are restrictions on which of them can be underspecified. Process verbs such as Spanish *carretear* (to haul) can drop the *origin* without restriction. However, they only allow the omission of the *destination* if the omission of the *origin* occurs simultaneously (9 and 10).

(9) Los operarios han carreteado la mercancía.

Operators have hauled the goods. (Vázquez et al. (2000) p. 193)

(10) *Los operarios han carreteado la mercancía *del camión*.

*Operators have hauled the goods *from the truck*. (Vázquez et al. (2000) p. 193)

Boons (1987) observes that, often, in order to express the *origin*, it is necessary to express the end as well, but not the other way round (concept of *dependent origin*). In contrast to process verbs, there are other trajectory predicates that cannot leave this component completely unspecified. These are telic verbs that lexically focus the *destination* or *origin* of the movement. For verbs that focus the end point of the trajectory, such as Spanish *venir* (to come), the lack of expression of this subcomponent does not correspond to an underspecification, but to an omission by ellipsis. In the ellipsis, the omission of the elements is due to the element had appeared previously, thus, the context directly supplies it (Matthews, 2014). Some verbs which focus the *destination* can express the *origin* in the same sentence and there are others which do not admit its appearance in any case (Vázquez et al., 2000).

As with verbs that focus the *destination*, verbs that focus the *origin* do not admit the underspecification of the incorporated component. Thus, in the event that the latter is not expressed, the procedure used is again ellipsis (11).

- (11) Claudia ha venido (*a casa*) caminando.
Claudia has come (*home*) walking. (Vázquez et al. (2000) p. 194)

In constructions in which the predicates focus the *origin*, the expression of the *destination* component cannot appear in the same sentence as the *origin* (Spanish *fugarse*, *irse*, *marcharse* and *huir*) (to escape, to go, to leave and to flee) (Vázquez et al., 2000).

The route subcomponent The route may refer to a precise point between the *origin* and the *destination* or it may express an unbounded path, i.e. without a specific *origin* and *destination*. The expression of the midpoint of the trajectory is generally accepted for most verbs of this class, regardless of telicity. As it has been said at the beginning of this section, one of the ways of expressing the route as a midpoint in Spanish is by complements introduced by the preposition *por*.

Verbs that accept the route understood as the space where the displacement is carried out focus the process. Therefore, the realisation of this constituent is not compatible with the focus of the *destination*. When a verb of this type has an argument introduced by the preposition *por*, it is a case of a constituent that expresses the route as a mid point (12).

- (12) Ha ido a Madrid (*por Zaragoza*).
He/she has gone to Madrid (*via Zaragoza*). (Vázquez et al. (2000) p. 198)

Thus, the phrase expressing a route as an indefinite space is not compatible with the expression of other constituents expressing the limits of the trajectory. The trajectory component can also be expressed in the form of a NP through which the entire path is accounted for (13).

- (13) Han atravesado *el bosque*.
They have gone through *the forest*. (Vázquez et al. (2000) p. 199)

This distance can be designated by measurements or by an entity that corresponds to the complete journey. These types of complements are restricted to verbs expressing a physical movement (Vázquez et al., 2000).

3.3 AnCora-Verb 2.0 lexicon and its annotation criteria

As it has been said at the beginning of the section, the lexicon Ancora-Verb 2.0 (Taulé et al., 2011) has been used for the representation of the verbs in this study. AnCora-Verb 2.0 is the guide used for annotating AnCora 2.0 corpora (a Spanish and Catalan corpora containing 105,101 predicates, 55,990 in Spanish and 49,111 in Catalan, mainly made up of journalistic texts) with verbal semantics (Lexical Semantic Structures, arguments and thematic roles). Ancora-Verb 2.0 contains 2828 files for Spanish and 2248 files for Catalan. In this study, just the Catalan files have been taken into account. Each file contains the information relating to one verb in all its senses and alternations and the name of each file corresponds to the verb in its infinitive form Taulé et al. (2011).

Both the lexicon and the corpus were annotated by 5 people and they were handled, revised and annotated using AnCoraPipe (Bertran et al., 2010), a tool for corpora annotation. AnCora-Verb 2.0 and AnCora 2.0 can be downloaded from AnCora's Web page⁴ just with prior authorisation.

In order to assign each verbal predicate of the lexicon to a specific semantic class and to tag every syntactic function with arguments and thematic roles, Taulé et al. (2011) followed Levin et al.'s (1995) definition of lexical decomposition, Rappaport Hovav and Levin's (1998) concept of Lexical Semantic Structure⁵ (LSS) and Kipper et al.'s (2002) and Kingsbury et al.'s (2002) guides in the construction of VerbNet.

According to Taulé et al. (2011), LSS restricts the set of all possible diatheses for each verb. In order to define the sense of each verb of the corpus, they follow the idea that each verb sense is associated to one LSS. Thus, according to Taulé et al. (2011) each sense of a verb is determined by a different LSS. When determining the LSS for each verb Taulé et al. (2011) also expose the semantic class of the verb (*causative, agentive, of motion, state, existential, attributive, scalar, benefactive, experiencer* and *source*). In addition, for each of their verb senses, Taulé et al. (2011) add the corresponding verb sense according to PropBank (Babko-Malaya, 2005), VerbNet (Kipper et al., 2006), FrameNet (Johnson et al., 2016) and WordNet (Miller, 1995).

Despite AnCora-Verb 2.0 containing the semantic class for every verb in the lexicon, in this study we disagree with the idea of every LSS being a different verb sense. It is true that we agree with the idea of a lexical unit activating a sense according to the the context in which appears Cabré (1999), yet we do not think that each sense should correspond strictly to an specific structure according to the number of arguments (and their syntactic function) that a verbal predicate requires. It is because of this reason, that verb senses

⁴<http://clic.ub.edu/corpus/es/ancora-descarregues>

⁵LSS 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 Taulé et al. (2011)

used in this study do not match exactly with AnCora-Verb 2.0's verb senses. The verb sense selection will be explain in more detail in Subsection 4.4.

The syntactic functions of the AnCora-Verb 2.0 lexicon are determined by the LSS and their mapping to thematic roles is given by the semantic class (Taulé et al., 2011). The criteria followed by Taulé et al. (2011) to annotate the arguments of the corpus is based on PropBank's proposal (Palmer et al., 2005). Taulé et al. (2011) distinguish between seven possible arguments: *arg0*, *arg1*, *arg2*, *arg3*, *arg4*, *argM* and *argL*. The first five tags are numbered according to the degree of proximity in relation to the verb; ArgM corresponds to adjuncts and argL corresponds to lexicalized complements of light verbs. When annotating thematic roles, the criteria followed by them is the semantic relation that each argument maintains with the event denoted by the verb. According to this idea, Taulé et al. (2011) proposed 20 different thematic roles, that can be found in Appendix B.

4 Methodology

In this section the description of how the study has been carried out is presented. The section is divided into 3 subsections. The first subsection describes, both qualitatively and quantitatively, the corpus used in this study. In this subsection, the steps followed to create the corpus are also explained. In the second subsection, which verbs and how they have been selected is explained. And the third subsection contains the aspects of verb structures that have been analysed and the steps that have been followed to do the analysis.

4.1 Description of the corpus

The corpus used in this study has been created from sports news gathering from Catalan newspapers. The newspapers selected for the creation of the corpus are *La Veu de l’Anoia*, *Regió7*, *El Punt Avui*, *L’Esportiu*, *VilaWeb*, *Diari de Girona*, *Tarragona Digital* and *Segre*. The news selected for the corpus were published between October of 2020 and February 2022 and were extracted from the 25th of January of 2022 until the 8th of February of 2022 (in Table 1 the collection and publication dates for each newspaper can be seen). In order to extract the news, two Python libraries have been used: *Feedparser* (McKee, 2021) (a Python module for downloading and parsing syndicated feeds) and *BeautifulSoup* (Richardson, 2007) (a Python library for pulling data out of HTML and XML files). The raw text has been saved in a *.txt* format and then, it has been analysed either automatically or manually depending on the purpose.

Newspaper	Dates of news’ collection	Dates of news’ publication
La Veu de l’Anoia	2022/01/25 - 2022/01/26	2021/04/14 - 2022/01/25
Regió7	2022/02/03 - 2022/02/08	2022/02/02 - 2022/02/08
El Punt Avui	2022/02/08	2022/02/08
L’Esportiu	2022/02/03 - 2022/02/09	2021/03/21 - 2022/02/03
VilaWeb	2022/02/03 - 2022/02/04	2020/10/27 - 2022/02/03
Diari de Girona	2022/02/05 - 2022/02/08	2022/02/04 - 2022/02/08
Tarragona Digital	2022/02/05 - 2022/02/08	2021/08/28 - 2022/02/05
Segre	2022/02/08	2022/02/05 - 2022/02/08

Table 1: Dates of news’ extraction and publication for each newspaper.

On the one hand, the R library *UDPipe* (Wijffels, 2022) has been used. This package allows doing out-of-the-box annotation of tokenization, Parts of Speech tagging (POS), Lemmatization and Dependency Parsing. Moreover, it also allows training new annotator models directly from R. In this study, the *UDPipe*’s pre-trained model *catalan-ancora* has been used for getting the automatic annotation of the corpus (in a *.csv* format). The resulting corpus after having been automatically annotated with *UDPipe* contains the following information for each token:

- Index number according to its order of appearance in the corpus

- Document in which appears
- Paragraph in which appears
- Index number of the sentence in which appears according to its order of appearance in the corpus
- Sentence in which appears
- The position of the first character of the token in the corpus
- The position of the last character of the token in the corpus
- A row identifier in which the token appears
- Token's index in the corpus
- The token itself
- Token's lemma
- PoS tag
- The token's index of the head of the token (indicating which other token in the sentence it is related to)
- Semantic-syntactic information
- Dependency relations

On the other hand, manual annotations for the selected verbs have been done (in a *.ods* format). How the annotations are made will be explained in Subsection 4.4. For the manual annotations, 4 columns have been added:

- A column referring to arguments structure
- A column referring to syntactic functions
- A column referring to thematic roles structure
- A column referring to verb senses

The information related to these 4 columns has not been annotated for each token in the corpus. It has only been annotated for each predicate of each of the occurrences of the verbs studied (*aconseguir*, *arribar*, *avançar*, *entrar* and *sortir*). In order to avoid overlapping information, for each of the verbs studied, a different sheet of the document *aconseguir* has been used.

Only one person has annotated the corpus. Thus, the Inter-Annotator Agreement (IAA) has not been possible to calculate.

3835 new_445
 3836 El Nàstic té set ocasions clares de gol al camp de la Balompèdica Linense i no és capaç de veure porteria. L'Andorra, fent un partit semblant al Nou Estadi davant els tarragonins, s'ha endut els tres punts amb un marcador contundent (0-3). Dos equips, dues definicions diferents a l'àrea contrària i dos resultats totalment oposats que deixen clar perquè els andorrans són a la part alta i els de Raül Agné no.
 3837 El FC Andorra ha saltat a la gespa del Nou Estadi i ha demostrat ser superiors del primer minut. Els tricolors han gaudit de les primeres arribades a l'àrea i de la primera clara ocasió que han transformat des dels onze metres. Protestat penal sobre Manu Nieto que Marc Fernández transformava en el minut 4 en el 0 a 1.
 3838 El duel comença malament per a un Nàstic que veuria com arriba a pisjar quan Romera es lesionava en el minut 18. Els visitants dominaven la possessió, trenaven accions combinades amb extrema facilitat i els grana sempre arribaven tard i malament. Semblava que l'Andorra no tingués rival, semblava que el partit era del Nàstic contra un equip de superior categoria.
 3839 Fruït de la insistència i el domini, ha arribat el 0 a 2 en el minut 39 després d'una clamorosa errada defensiva de Nil al lateral esquerre de la defensa. Manu Nieto li roba L'esfèric, es planta sol davant Manu i el supera amb una elegant retallada. El partit semblava ja quasi impossible però s'ha convertit en impossible del tot quan en el minut 33 arribava el 0 a 3 amb un cop de cap de Nieto.
 3840 El Nàstic, que només havia marcat un gol en els darrers cinc partits, en necessitava quatre per guanyar. Els andorrans seguien treballant i mostrant-se superiors, gaudint d'alguna ocasió més per ampliar la golejada. I quan els de Raül Agné han gaudit d'alguna ocasió, ha acabat als novells.
 3841 Com en el minut 44 quan Pedro Martín, amb el porter batut, no engaltava bé el cop de cap i desaprofitava una clara ocasió local del primer temps. Desafortunadament per als interessos locals, aquesta acció seria un miratge i la segona part del matx ha estat un tràmit. El Nàstic quasi no ha generat perill i l'Andorra no ha patit per conservar el resultat.
 3842 Amb aquesta derrota, els de la Budaiera queden catorzens amb 25 punts i el descens aproximant-se a només dos punts. Pel que fa a la promoció d'ascens, aquesta queda marcada per l'Algeciras amb 30 punts. Dimecres a les 17.00 hores, una final: visita al Betis, cuer de la lliga.
 3843 FITXA TÈCNICA
 3844 Gimnàstic
 3845 Manu, Ortel, Aythami, Quintanilla, Nil (Bontilla 52'), Ribelles (Valverde 70'), Del Campo, Elias, Robert, Pedro (Fernández 46') i Romera (Carbia 19')
 3846 Andorra
 3847 Ratti, Vilanova, Aguado, Hevel, Fernández (Altt 63'), Molina (Gll 78'), Olego, Nieto (Enri 78'), Vllà, Bover (Morer 74') i Pastor (Vergés 63')
 3848 Gols
 3849 0-1 / Fernández (4')
 3850 0-2 / Nieto (30')
 3851 0-3 / Nieto (33')
 3852 Targetes grogues
 3853 Per al local Aythami (36') i per als visitants González (71') i Aguado (83')
 3854 Àrbitre
 3855 Iván Muñoz, del comitè territorial valencià
 3856 Incidències
 3857 Matx corresponent a la jornada 20 del grup II de Primera RFEF amb la presència de 2.195 espectadors
 3858 new_446
 3859 «Que ningú ens doni per morts, aquí no ens rendirem mai perquè ningú és millor que nosaltres». Amb aquesta frase ha presentat Raül Agné, tècnic del Nàstic, el partit de dimecres a les 19.00 hores al Nou Estadi davant l'Andorra. Els grana enceten la segona volta en plena crisi golejadora i buscant el primer triomf de l'any per reenganxar-se a la zona alta.
 3860 La prova per al conjunt tarragoní serà exigent, ja que es presenta al Nou Estadi el tercer classificat, un Andorra que acumula 32 punts després de nou victòries i cinc empats. Els tricolors encara no han guanyat en el 2022 després de perdre contra el Castella i empatar davant el Costa Brava. Tot i els dos darrers encontres, cal recordar que van finalitzar el 2021 amb quatre triomfs seguits.
 3861 Agné, tècnic grana, reconeix que davant l'Andorra «tenim una nova oportunitat i esperem que el procés que estem tenint el pugem retindicar amb resultats». A més, és conscient que «visc dels resultats, he de passar i creure en el procés i aquest està sent el que vol tot entrenador». Javi Ribelles tornarà a estar disponible després de complir el partit de sanció a la Lliga de la Concepció.
 3862 Iván Muñoz, acompanyant a les bandes per Carlos Martínez i Rubén Ortega, seran els col·legiats del matx.
 3863 new_447
 3864 La companya #NotesEnJoc va aconseguir els objectius pels quals va néixer. Una iniciativa creada a Flix per reivindicar que una jugadora de 13 anys, futbolista de la JE Flix, continuï jugant amb el seu equip la temporada vinent, a la categoria cadet.
 3865 Una intenció que, segons la Federació Catalana de Futbol (FCF), podria complir. La normativa va canviar a l'assemblea celebrada a Blanes l'any 2020, quan es va ampliar la possibilitat del futbol mixt fins a la categoria cadet. Un fet que possibilita que la jugadora flixana vesteixi la samarreta del club.
 3866 Tot i això, els impulsors de la iniciativa no es volen quedar aquí. I vol lluitar que les notes adolescents puguin jugar a futbol fins a la categoria juvenil, és a dir, que l'edat límit siguin els 18 anys i no els 16, com és ara.
 3867 El motiu principal és que les notes, sobretot aquelles que viuen en zones rurals llunys de la possibilitat de competir en un equip femení, puguin gaudir de l'esport i el futbol. En aquest sentit, quan fan els 18 anys, normalment, segueixen els seus estudis a la ciutat i poden trobar més facilitats per canviar a un equip femení.
 3868 Per tot plegat, la companya #NotesEnJoc continua activa per reclamar equips mixtos fins a la categoria juvenil. Des que es va engegar la iniciativa, s'han acumulat més de 1.800 firmes de suport.

Figure 1: Appearance of the raw text of the corpus.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	doc_id	paragraph_id	sentence_id	start	end	term_id	token_id	token	lemma	upos	xpos	feats	head_token	l/dep	rel	deps	misc	
2	1	1	1	1	7	1	1	new_001	new	NUM	NUM	NumForm=Digit NumType=0	0	root	NA	NA	SpacesAfter=0	
3	2	doc2	1	1	1	8	1	IGUALADA	Igualada	PROPN	PROPN	NA	3	subj	NA	NA		
4	3	doc2	1	1	10	14	2	RIGAT	RIGAT	PROPN	PROPN	NA	1	flat	NA	NA		
5	4	doc2	1	1	16	16	3	3	1	NUM	NUM	NumForm=Digit NumType=0	0	root	NA	SpacesAfter=0		
6	5	doc3	1	1	1	5	1	1BARÇA	Barça	ADV	ADV	NA	2	advmod	NA	NA		
7	6	doc3	1	1	7	8	2	2	10	NUM	NUM	NumForm=Digit NumType=0	0	root	NA	SpacesAfter=0		
8	7	doc4	1	1	1	4	1	1Dues	dos	NUM	NUM	Gender=Fem Number=Plur	2	nummod	NA	NA		
9	8	doc4	1	1	6	10	2	2cares	cara	NOUN	NOUN	Gender=Fem Number=Plur	0	root	NA	NA		
10	9	doc4	1	1	12	12	3	3i	i	CCONJ	CCONJ	NA	5	cc	NA	NA		
11	10	doc4	1	1	14	16	4	4ben	ben	ADV	ADV	NA	5	advmod	NA	NA		
12	11	doc4	1	1	18	30	5	5diferenciades	diferenciat	ADJ	ADJ	Gender=Fem Number=Plur	2	conj	NA	SpaceAfter=No		
13	12	doc4	1	1	31	31	6	6.	PUNCT	PUNCT	PunctType=Peri	2	punct	NA	NA			
14	13	doc4	1	2	33	39	7	1Gairebé	gairebé	ADV	ADV	NA	0	root	NA	NA		
15	14	doc4	1	2	41	43	8	2com	com	SCONJ	SCONJ	NA	4	mark	NA	NA		
16	15	doc4	1	2	45	46	9	3la	el	DET	DET	Definite=Def Gender=Fem N	4	det	NA	NA		
17	16	doc4	1	2	48	50	10	4nit	nit	NOUN	NOUN	Gender=Fem Number=Sing	1	obj	NA	NA		
18	17	doc4	1	2	52	52	11	5i	i	CCONJ	CCONJ	NA	7	cc	NA	NA		
19	18	doc4	1	2	54	55	12	6el	el	DET	DET	Definite=Def Gender=Masc	7	det	NA	NA		
20	19	doc4	1	2	57	59	13	7dia	dia	NOUN	NOUN	Gender=Masc Number=Sing	4	conj	NA	SpaceAfter=No		
21	20	doc4	1	2	60	60	14	8.	PUNCT	PUNCT	PunctType=Peri	1	punct	NA	NA			
22	21	doc4	1	3	62	62	15	1A	ADP	ADP	AdpType=Prep	4	case	NA	NA			
23	22	doc4	1	3	64	65	16	2la	el	DET	DET	Definite=Def Gender=Fem N	4	det	NA	NA		
24	23	doc4	1	3	67	73	17	3primera	primer	ADJ	ADJ	Gender=Fem Number=Sing	4	amod	NA	NA		
25	24	doc4	1	3	75	78	18	4part	part	NOUN	NOUN	Gender=Fem Number=Sing	9	obj	NA	NA		
26	25	doc4	1	3	80	89	19	5Iguualada	Iguualada	PROPN	PROPN	NA	9	subj	NA	NA		
27	26	doc4	1	3	91	92	20	6HC	HC	PROPN	PROPN	NA	5	flat	NA	NA		
28	27	doc4	1	3	94	98	21	7Rigat	Rigat	PROPN	PROPN	NA	5	flat	NA	NA		
29	28	doc4	1	3	100	101	22	8va	anar	AUX	AUX	Mood=Ind Number=Sing Per	9	aux	NA	NA		
30	29	doc4	1	3	103	110	23	9competir	competir	VERB	VERB	VerbForm=Inf	0	root	NA	NA		
31	30	doc4	1	3	112	113	24	10de	de	ADP	ADP	AdpType=Prep	11	case	NA	NA		
32	31	doc4	1	3	115	116	25	11tu	tu	PRON	NOUN	Case=Nom Number=Sing Pr	9	iobj	NA	NA		

Figure 2: Appearance of the corpus automatically annotated.

In Figure 1, the appearance of the raw text can be seen; in Figure 2, the appearance of the automatic annotations can be seen; and, finally, in Figure 3, the appearance of the manual annotations can be seen.

In relation to the quantitative description of the corpus, the corpus has a total of 755 news and 288651 tokens. In Table 2, the amount of news items per newspaper and its distribution within the corpus can be seen. Regarding the grammatical categories of the corpus' tokens, it can be seen that the one appearing the most is the category *Nouns*. It has an appearance of 48538 items. The second grammatical category appearing the most is the category *Prepositions*. It has an appearance of 41098 items. Even the third grammatical

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	
215428	215427	doc477	1	1	Els del	76	77	18	17	al	al	ADP	ADP	AdpType	16	fixed	NA	NA						
215429	215428	doc477	1	1	Els del	79	85	19	18	Pavelló	Pav	PROP	PROP	NA	15	nmod	NA	NA						
215430	215429	doc477	1	1	Els del	87	90	20	19	Joan	Joa	PROP	PROP	NA	18	flat	NA	NA						
215431	215430	doc477	1	1	Els del	92	97	21	20	Ortoll	Orto	PROP	PROP	NA	18	flat	NA	SpaceAfter=No						
215432	215431	doc477	1	1	Els del	98	98	22	21	.	.	PUNO	PUNCT	PunctTyp	7	punct	NA	NA						
215433	215432	doc477	1	2	Jan Es	100	102	23	1	Jan	Jan	PROP	PROP	NA	3	nsubj	arg0	sub	agt					
215434	215433	doc477	1	2	Jan Es	104	109	24	2	Escala	esc	PROP	PROP	NA	1	flat	arg0	sub	agt					
215435	215434	doc477	1	2	Jan Es	112	119	25	3	avançav	ava	VERB	VERB	Mood=In	0	root	NA	NA					Sense 2	
215436	215435	doc477	1	2	Jan Es	121	127	26	4	l'equip	l'eq	NOUN	NOUN	Gender=>	3	obj	arg1	cd	pat					
215437	215436	doc477	1	2	Jan Es	129	130	27	5	en	en	ADP	ADP	AdpType	7	case	argM	cc	tmp					
215438	215437	doc477	1	2	Jan Es	132	133	28	6	el	el	DET	DET	Definite=>	7	det	argM	cc	tmp					
215439	215438	doc477	1	2	Jan Es	135	139	29	7	minut	min	NOUN	NOUN	Gender=>	3	obl	argM	cc	tmp					
215440	215439	doc477	1	2	Jan Es	141	141	30	8	6	6	NUM	NUM	NumFonn	3	obl	argM	cc	tmp					
215441	215440	doc477	1	2	Jan Es	143	149	31	9	després	des	ADV	ADV	NA	8	advmod	argM	cc	tmp					
215442	215441	doc477	1	2	Jan Es	151	154	32	10	d'un	d'un	DET	DET	Gender=>	12	det	argM	cc	tmp					
215443	215442	doc477	1	2	Jan Es	156	161	33	11	potent	pote	ADJ	ADJ	Number=>	12	amod	argM	cc	tmp					
215444	215443	doc477	1	2	Jan Es	163	165	34	12	xut	xut	NOUN	NOUN	Gender=>	9	obl	argM	cc	tmp					
215445	215444	doc477	1	2	Jan Es	167	169	35	13	des	des	ADP	ADP	AdpType	16	case	argM	cc	tmp					
215446	215445	doc477	1	2	Jan Es	171	172	36	14	de	de	ADP	ADP	AdpType	13	fixed	argM	cc	tmp					
215447	215446	doc477	1	2	Jan Es	174	175	37	15	la	el	DET	DET	Definite=>	16	det	argM	cc	tmp					
215448	215447	doc477	1	2	Jan Es	177	181	38	16	banda	ban	NOUN	NOUN	Gender=>	12	nmod	argM	cc	tmp					
215449	215448	doc477	1	2	Jan Es	183	190	39	17	esquerr	esq	ADJ	ADJ	Gender=>	16	amod	argM	cc	tmp					
215450	215449	doc477	1	2	Jan Es	191	191	40	18	.	.	PUNO	PUNCT	PunctTyp	3	punct	NA	NA						
215451	215450	doc477	1	3	Sergi	193	197	41	1	Sergi	Ser	PROP	PROP	NA	8	nsubj	NA	NA						
215452	215451	doc477	1	3	Sergi	199	203	42	2	Miras	mir	PROP	PROP	NA	1	flat	SpaceAfter=No							
215453	215452	doc477	1	3	Sergi	204	204	43	3	,	,	PUNO	PUNCT	PunctTyp	4	punct	NA	NA						

Figure 3: Appearance of the corpus manually annotated.

category appearing the most would be *Determiners* (with an appearance of 33729 items), it has to be noticed that if *Verbs* and *Auxiliary verbs* are added together, they become the third most frequent grammatical category in the corpus. Both categories together have an appearance of 40935 items.

Newspaper	Amount of news	Location of the news within the corpus
La Veu de l'Anoia	198	1 st - 198 th
Regió7	18	199 th - 206 th and 581 st - 590 th
El Punt Avui	18	207 th - 214 th and 571 st - 580 th
L'Esportiu	176	215 th - 233 rd , 596 th and 600 th - 755 th
VilaWeb	194	234 th - 427 th 428 th - 437 th , 591 st - 595 th and 597 th - 599 th
Diari de Girona	18	
Tarragona Digital	118	438 th - 555 th
Segre	15	556 th - 570 th

Table 2: Total number of news items per newspaper and its location within the corpus marked by ordinal numbers

In Figure 4 the frequency of appearance of the token according to its grammatical category can be seen. It also can be seen that the appearance of lexical and functional categories is quite evenly matched.

Due to the restrictive licenses of the newspapers selected, none of the versions of the corpus can be published. They do not allow reproduction, distribution or public communication of their content. Even these restrictions do not allow open research, there is no other way to carry out the analysis for this study. Moreover, it needs to be pointed out

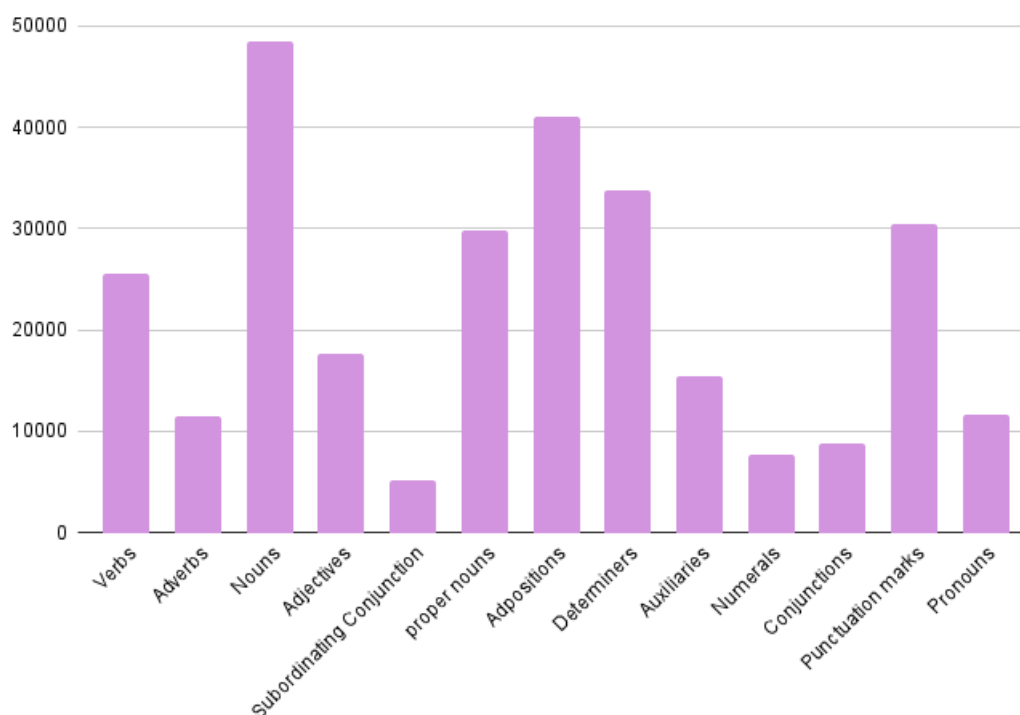


Figure 4: Frequency of grammatical categories in the corpus.

that this kind of restrictions do not help increasing the reduced number of studies and tools for Catalan related to language in the domain of sports.

4.2 Verbs' selection

In this study 3 parameters have been taken into account in order to select the verbs for the analysis.

The first aspect considered has been knowing which verbs of the corpus belong to the studied class. In order to carry out this first step, Vázquez et al.'s (2000) list of Catalan trajectory verbs (which can be found in Appendix A has been considered).

The second aspect has been selecting those verbs which are easily found in a sports context. Related to this second aspect, and after knowing which verbs belong to the trajectory class, their frequency of appearance in the corpus has been considered too. Thus, 5 of the class' verbs with a high appearance in the corpus have been *aconseguir* (with an appearance of 318 occurrences), *arribar* (with an appearance of 435 occurrences), *avançar* (with an appearance of 111 occurrences), *entrar* (with an appearance of 133 occurrences) and *sortir* (with an appearance of 159 occurrences).

Finally, the third, and last aspect that has been considered for the verbs selection has been the representation of different canonical structures of the predicates. In this step, different phenomena have been taken into account: transitivity, telicity, ellipsis, under-

specification and focalization. Analysing verbs with different natural structures allows to see if verbs of the same class behave different in a specific field according to their structure.

Regarding transitivity, it can be said that in general *aconseguir* is transitive and *sortir* is intransitive. However, *arribar*, *avançar* and *entrar* can be either transitive or intransitive depending on their sense selection.

Regarding telicity, it can be said that, in general, *aconseguir*, *arribar*, *entrar* and *sortir* are telic, because have an specific limit; but *avançar* is atelic because it does not have an specific limit (Vázquez et al., 2000).

Ellipsis and underspecification are so closely related that it can be said that when there is an omission of information in a predicate where the verb does not allow underspecification, it is a case of ellipsis. As it has been said before, verbs accepting underspecification are atelic verbs. Thus, *avançar* allows underspecification. On the other hand, when there is an omission of information in some part of the predicate of the other four verbs (*aconseguir*, *arribar*, *entrar* and *sortir*) it is a case of ellipsis.

According to focalization, it can be said that, in general, *arribar* and *entrar* focalise the *destination* or the *location*; it can also be said that *sortir* focalises the *origin*. However, neither *aconseguir* nor *avançar* focalise anything (Vázquez et al., 2000).

4.3 AnCora-Verb 2.0's verb senses for *aconseguir*, *arribar*, *avançar*, *entrar* and *sortir*

In Section 3.3 it has been said that AnCora-Verb 2.0 (Taulé et al., 2011) has been followed in this study to do the representation of the selected verbs. In this subsection the verb senses proposed by Taulé et al. (2011) for each of the 5 verbs selected are exposed; however, AnCora-Verb 2.0 mapping of the syntactic structures, the arguments and the thematic structures of all verb senses are located in the Appendix C.

Aconseguir The verb *aconseguir* just has 1 sense, which corresponds to VerbNet's *obtain-13.5.2*. However, it has two different structures, one for the active voice and another for the passive.

Arribar The verb *arribar* has 7 different senses. The 1st one corresponds to Verbnet's *escape-51.1*, the 2nd, the 3rd and the 5th ones correspond to VerbNet's *get-13.5.1*, the 4th sense corresponds to VerbNet's *send-11.1*, and finally neither 6th nor 7th senses correspond to any of Verbnet's classes. 1st, 2nd and 3rd senses have two different structures: one referring to active voice and another to impersonal voice. 4th sense also has two different structures: one for the active voice and another one for the passive. Finally, 5th, 6th, and 7th senses just have an structure for the active voice.

Avançar The verb *avançar* has 6 different senses. 1st, 4th and 6th senses correspond to Verbnet's class *other_cos-45.4*, 2nd sense corresponds to VerbNet's *roll-51.3.1*, and finally neither 3rd nor 5th senses correspond to any of Verbnet's classes. 1st sense has 2 different

structures: one referring to active voice and another to causative voice. 2nd sense has 4 different structures: one for the active voice, another for the impersonal voice, another for the benefactive voice and another one for the passive voice. Finally, 3rd, 4th, 5th and 6th senses just have an structure for the active voice.

Entrar The verb *entrar* has 7 different senses. 1st and 6th senses correspond to Verbnet's *escape-51.1*, 4th, and 7th senses correspond to VerbNet's *characterize-29.2* and finally neither 2nd, 3rd nor 5th senses correspond to any of Verbnet's classes. 1st and 3rd senses have two different structures: one referring to active voice and another to impersonal voice. 4th sense also has two different structures: one for the active voice and another one for the passive. Finally, 2nd, 5th, 6th, and 7th senses just have an structure for the active voice.

Sortir The verb *sortir* has 5 different senses. 1st sense has two different structures: one referring to active voice and another to impersonal voice. The rest of senses just have the active structure. The active structure of 1st sense and 4th sense correspond to Verbnet's *appear-48.1.1*, the impersonal structure of 1st sense corresponds to Verbnet's *escape-51.1*. Finally, neither 2nd, 3rd nor 5th senses correspond to any of Verbnet's classes.

4.4 Process of the analysis of the verbs

Different aspects related to the structure of the verbs *aconseguir*, *arribar*, *avançar*, *entrar* and *sortir* have been analysed. Regarding the work of Casademont (2014), a mapping between the aspects she analysed and the ones analysed in this study can be done. As it has been said in Section 2.2.1, Casademont (2014) analysed the corpus according to 10 different levels of description. In this study, just 4 of them have been considered (those which showed relevant results on the transmission of specialized knowledge): *Adicity*, *Subcategorization of the constituents*, *Thematic Role* and *Sense*.

The labels used for the levels of description have been slightly modified (except for the label *Sense* which has been maintained): *Adicity* has been named *Arguments structure*, *Subcategorization of the constituents* has been named *Syntactic structure* and *Thematic Role* have been named *Thematic structure*.

The first step followed to carry out the analysis has been differentiating how many senses each verb has and which they are. In this first step, AnCora's verb senses have been taken into account. However, the selected senses for this study do not match exactly with AnCora's proposal. As it has been said in Section 3.3, Taulé et al. (2011) follow the idea that every verb sense is associated to one LSS. Thus according to them, (and based on the senses described in Section 4.3) *aconseguir* would have 1 sense, *arribar* would have 7 different senses, *avançar* would have 6 different senses, *entrar* would have 7 different senses and *sortir* would have 5 different senses.

Not all AnCora senses are mapped with a VerbNet class. In this study, only the AnCora senses which had a mapping with VerbNet's classes have been selected for the analysis. Thus, those structures that did not have a mapping sense in VerbNet have been added to

existing senses or have been discarded. The verb senses considered for this study are the following:

- Aconseguir: 1 sense (corresponding to VerbNet’s class *obtain-13.5.2*)
- Arribar: 3 senses (corresponding to VerbNet’s classes *escape-51.1*, *get-13.5.1* and *send-11.1*)
- Avançar: 2 senses (corresponding to VerbNet’s classes *other_cos-45.4* and *roll-51.3.1*)
- Entrar: 2 senses (corresponding to VerbNet’s classes *escape-51.1* and *characterize-29.2*)
- Sortir: 2 senses (corresponding to VerbNet’s classes *appear-48.1.1* and *escape-51.1*)

In this study, the names of VerbNet’s classes have not been used. Cardinal numbers have been used, instead. Table 3 shows the labels used in this study for every verb sense and their corresponding VerbNet labels.

Verb	VerbNet’s label	Label used in this study
Aconseguir	obtain-13.5.2	sense 1 (aconseguir)
Arribar	escape-51.1	sense 1 (arribar)
Arribar	get-13.5.1	sense 2 (arribar)
Arribar	send-11.1	sense 3 (arribar)
Avançar	other_cos-45.4	sense 1 (avançar)
Avançar	roll-51.3.1	sense 2 (avançar)
Entrar	escape-51.1	sense 1 (entrar)
Entrar	characterize-29.2	sense 2 (entrar)
Sortir	appear-48.1.1	sense 1 (sortir)
Sortir	escape-51.1	sense 2 (sortir)

Table 3: Mapping of VerbNet labels for verb senses and the labels used in this study.

The second step has been annotating all the occurrences of the selected verbs according to its arguments, syntactic and thematic structures. For the annotation of these structures, the labels used in the AnCora corpus have been used too in this study (all the labels can be found in Appendix B). However, not all of them have appeared in our annotations. The following are the labels that appear in the annotations of the corpus of this study:

- The labels used for the arguments structures annotation are *arg0*, *arg1*, *arg2*, *arg3*, *arg4*, *argM* and *argL*.
- The labels used for the syntactic structures annotation are *subj*, *cd*, *ci*, *creg*, *cag*, *cpred*, and *cc*.

- The labels used for the thematic structures annotation are *agt*, *pat*, *tem*, *atr*, *loc*, *ori*, *des*, *ben*, *mnr*, *cau*, *fin*, *tmp* and *adv*.

The annotations were done in the corpus with the *.ods* format due to displaying purposes. When annotating, colours for the different type of arguments were used. As cells' colouring can not be done in the *.csv* format, the annotations were done in the *.ods* format. The colours added in the arguments structure column help with the visualisation of the structure. The corresponding colours used for each of the arguments are the following:

- arg0 - red
- arg1 - orange
- arg2 - yellow
- arg3 - purple
- arg4 - light green
- argM - dark green
- argL - brown

In Figure 5 an example of an annotation of the corpus can be seen.

When doing the analysis, some structures have been omitted. Thus, they have not been analysed. These structures have been the following:

- Repeated sentences, news and paragraphs
- Sentences in which the selected verb for the analysis was nominalized or worked as an adverb
- Periphrastic structures containing the verbs⁶ *aconseguir*, *arribar*, *avançar*, *entrar* or *sortir*
- Factive structures (with the exception of factive structures containing the verb *arribar*, which the verb phrase of one of its LSS is *fer arribar* (to make arrive))

After having omitted these structures in the analysis, the number of occurrences analysed for each verb is as follows:

- *Aconseguir* - 312 occurrences analysed
- *Arribar* - 389 occurrences analysed
- *Avançar* - 109 occurrences analysed

⁶Although they are not always so easy to distinguish from more general syntactic-semantic structures.

L'exroig-i-negre	L	P	P	N	4	n	N	NA	arg0	su	agt	
Jan	J	P	P	N	1	fl	N	NA	arg0	su	agt	
Escala	e	P	P	N	1	fl	N	Sp	arg0	su	agt	
avançava	a	V	V	N	0	r	N	NA				Sense 2
el	e	D	D	D	6	d	N	NA	arg1	cd	pat	
Calafell	C	P	P	N	4	o	N	NA	arg1	cd	pat	
en	e	A	A	A	9	c	N	NA	argM	cc	tmp	
el	e	D	D	D	9	d	N	NA	argM	cc	tmp	
minut	m	N	N	Q	4	o	N	NA	argM	cc	tmp	
4	4	N	N	N	#	n	N	Sp	argM	cc	tmp	
després	d	A	A	A	#	a	N	NA	argM	cc	tmp	
de	d	A	A	A	#	n	N	NA	argM	cc	tmp	
rematar	r	V	A	V	#	a	N	NA	argM	cc	tmp	
al	a	A	A	A	#	c	N	NA	argM	cc	tmp	
segon	s	A	A	Q	#	a	N	NA	argM	cc	tmp	
pal	p	N	N	Q	#	o	N	NA	argM	cc	tmp	
una	u	D	D	Q	#	d	N	NA	argM	cc	tmp	
assistència	a	N	N	Q	#	n	N	NA	argM	cc	tmp	
d'Arnau	d	V	V	N	4	a	N	NA	argM	cc	tmp	
Xaus	X	P	P	N	#	o	N	Sp	argM	cc	tmp	

Figure 5: Example of annotation of the corpus.

- *Entrar* - 124 occurrences analysed
- *Sortir* - 148 occurrences analysed

Therefore, the number of occurrences analysed differ from the total number of occurrences appearing in the corpus (listed in Subsection 4.2)

5 Results and Discussion

In this section the quantitative and qualitative results obtained in this study will be presented.

Statistics results of the different structures and proposals for canonical structures of every verb sense in the specialized field of sports news will be presented. Moreover, this section also contains the differences between the results obtained in this study and AnCora's proposals for structures.

When analysing the results, a threshold has been established. Structures appearing more than 4 times in the corpus are the ones considered in the statistics. Despite those occurrences appearing less than 4 times do not appear in the graphics, they have been analysed too. In Appendix D tables containing the results of all the occurrences for every verb sense can be found.

5.1 Results of the structures of the verb *aconseguir*

Regarding the arguments structures proper to the verb *aconseguir*, the results have shown that the most common one is *arg0 + arg1 + argM* (with an appearance of 110 times) (14). It is followed by the arguments structures *arg1 + argM* (with an appearance of 81 times) and *arg1* (with an appearance of 77 times). The less common arguments structure of this verb has been the structure *arg0 + arg1*. However, even if it has been the less frequent structure, the number of appearances in the corpus is quite relevant (it has an appearance of 44 times). In Figure 6 a graph of the frequencies of appearance of the arguments structures in the corpus can be seen.

- (14) *Un Roger Pérez excel·lent* (arg0, suj) va aconseguir *portar el seu equip a l'empat* (arg1 cd) *a falta de 30 segons* (argM cc).

An excellent Roger Pérez (arg0 suj) managed to *lead his team to a draw* (arg1 cd) *with 30 seconds to go* (argM cc).

With regard to the possible syntactic structures of the verb, the results have shown that the most common syntactic structure for the verb *aconseguir* has been *suj + cd + cc* (with a 109 times frequency) (14). *cd + cc* and *cd* have been the second and the third most common syntactic structures for this verb (with an appearance frequency of 79 and 75 times, respectively). Another relevant syntactic structure found in the corpus has been the structure *suj + cd* (with an appearance of 43 times). Other syntactic structures that have also appeared in the analysis have been *suj*, *suj + cc*, *suj + cag* and *cd + ci*. However, Figure 7 only shows the frequencies of appearance of the arguments structures in the corpus of structures appearing more than 4 times.

Finally, the most common thematic structures of the verb *aconseguir* are the ones showed in Figure 8. This figure shows those thematic structures proper to the verb *aconseguir* that have had a longer appearance than 4 times in the corpus. As it can be seen in the figure, the most common thematic structure for this verb has been *pat* (with an

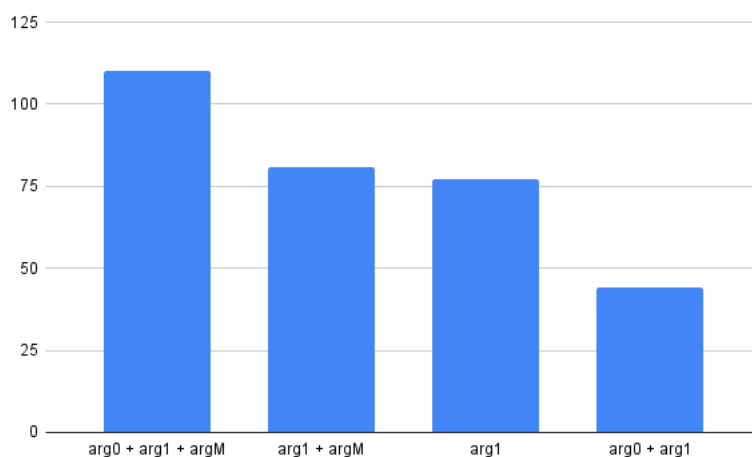


Figure 6: Arguments structures with a longer appearance than 4 times of the verb *aconseguir*

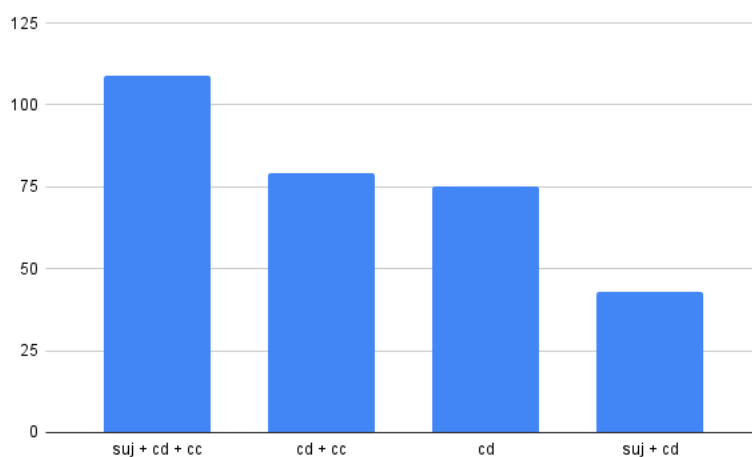


Figure 7: Syntactic structures with a longer appearance than 4 times of the verb *aconseguir*

appearance of 77 times) (15), it is followed by the structures *atg + pat*, *agt + pat + tmp*, and *pat + tmp* (with an appearance of 45, 32 and 29 times, respectively). Other structures which have had a quite relevant appearance in the corpus have been *agt + pat + mnr*, *agt + pat + tmp + mnr*, *pat + mnr* and *agt + pat + loc* (with an appearance of 17, 16, 15 and 14 times, respectively). Finally, 4 other structures that have exceeded 4 times the number of appearances have been *agt + pat + tmp + loc*, *pat + loc*, *pat + tmp + mnr* and *pat + fin*.

- (15) Van aconseguir *tres gols més* (pat).
They achieved *three more goals* (pat).

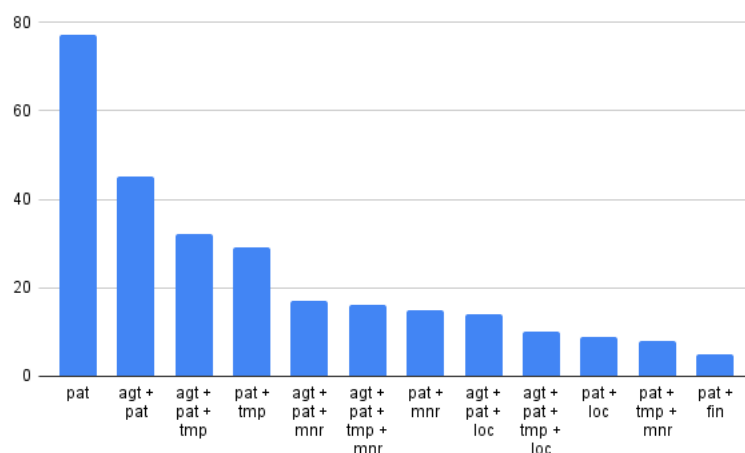


Figure 8: Thematic structures with a longer appearance than 4 times of the verb *aconseguir*

These results have shown that in many cases this verb appears without adjuncts. However, the most frequent with which it appears are *time*, *manner* and *location*. It should also be noted that even the most common arguments structure and syntactic structure contain an *arg0* and a *subj*, respectively, plenty of times the verb *aconseguir* also appears without the argument *agent*. Thus, in many sentences just the *patient* is required (16).

- (16) Van aconseguir *debutar amb el primer equip* (pat) *aquest passat dissabte* (tmp).
They got *their first team debuts* (pat) *last Saturday* (tmp).

There is an adjunct that, even if it does not appear in the most frequent argument structures, should be highlighted from the results. This adjunct is the adjunct *co-agent* (it appeared 11 times in the corpus). This adjunct is not an AnCora's proposal, but our proposal. This adjunct makes reference to an entity (normally a person or a group of people) which does not do the action but without them the action would not be possible. Either is not the *agent* nor the *patient*, but is involved in the performance of the action. We have considered it to be an adjunct and not an argument because it can be easily omitted without changing the meaning of the sentence. This element is introduced by the prepositions *contra* (against) and *a*, or the prepositional phrase *davant de* (in front of).

An example of a sentence of the corpus in which this thematic role appears would be (17):

- (17) L'equip que entrena Marco Rossi ha aconseguït d'empatar *contra la selecció francesa, actual campiona del món i subcampiona d'Europa*.
The team coached by Marco Rossi has managed to draw *against the French team, current world champions and runners-up in Europe*.

We have considered this type of adjunct to be specific to the sports news context.

With these results 3 other differences with respect to AnCora's structure proposal have been found. The first one is that according to AnCora, in those *manner* adjuncts which are introduced by a preposition, the prepositions are *amb* (with) and *en* (in). However, the results have shown that other prepositions or prepositional phrases can also introduce these type of adjunct. These are *de* (of) (18), *per davant* (ahead), *per* (for) and *sense* (without).

- (18) *De manera èpica.*
In an epic way.

The second difference is that in Ancora's structure proposal the locative adjunct is not said to be a proper adjunct to the verb *aconseguir*. However, in this corpus it has appeared 47 times. Thus, we have considered it to be typical of a sports news context (19).

- (19) *Que va aconseguir en la final de Tokio.*
That he/she achieved *in the Tokyo final*.

Finally, the third difference is that in Ancora's structure proposal there is an attribute adjunct which is syntactic function is *cpred*. However, no attribute adjunct has been found in this study for the verb *aconseguir*. Thus, may be this kind of adjunct does not belong to the specialized context studied.

After seeing all the statistical results, a canonical structure for the verb *aconseguir* in a sports news context is proposed. This canonical structure can be seen in Table 4.

Arguments structure	Syntactic structure	Thematic structure
arg0	subj	agt
arg1	cd	pat
argM	cc	tmp
argM	cc	mnr
argM	cc	loc
argM	cc	fin
argM	cc	co-agt
argM	cc	adv

Table 4: Proposal of the canonical structure of the verb *aconseguir* in the specialized field of sports news.

5.2 Results of the structures of the verb *arribar*

According to Table 3, presented in Section 4.4, *arribar* has three different senses. However, the results have shown that sense 1 is the one appearing by far most often in the corpus (366 occurrences). Sense 2 is the second appearing the most and sense 3 is the third. Even so, these two senses just have an 18 and 5 times appearance, respectively. Thus, it could

be said that sense 1 (corresponding to VerbNet's *escape-51.1* sense) is the most frequent in the specialized field of sports news.

With respect to the most common arguments structures of *arribar*'s sense 1, 7 different structures have been taken into account because they appear more than 4 times in the corpus. The structure that appears most often is *arg1 + argM* (it has an appearance of 121 times) (20). It is followed by the structures *arg1 + arg4 + argM* (with an appearance of 60 times), *arg4* (with an appearance of 55 times), *arg 4 + argM* (with an appearance of 44 times), *arg1* (with an appearance of 38 times) and *arg1 + arg4* (with an appearance of 21 times). Finally, the 7th structure is not as frequent as the previous ones, but it can also be taken into account: *argM* (it has an appearance of 6 times). In Figure 9 these results can be seen in a graphical form.

- (20) *Els principals avaladors de Laporta* (arg1 suj tem) han arribat *a última hora* (argM cc tmp).

Laporta's main guarantors (arg1 suj tem) have arrived *at the last minute* (argM cc tmp).

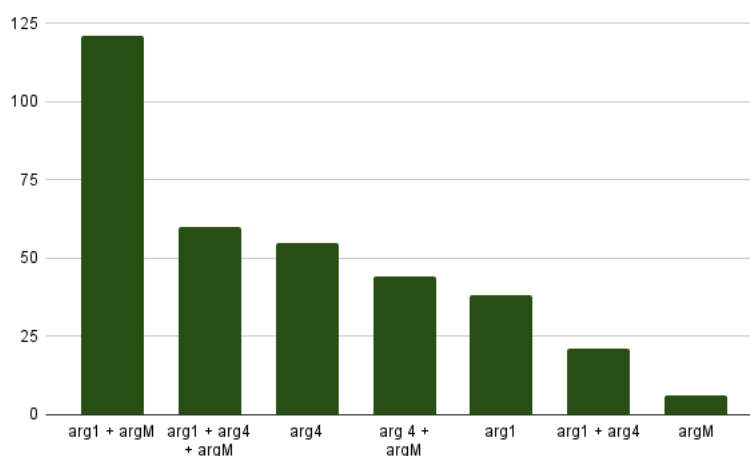


Figure 9: Arguments structures with a longer appearance than 4 times of the sense 1 of the verb *arribar*

Concerning possible syntactic structures for sense 1 of *arribar*, the results have shown that the most frequent one is *suj + cc* (9). It appears 196 times in the corpus. The second structure appearing the most is *cc*. It also has a high frequency number, it appears 105 times in the corpus. Three other syntactic structures can be considered relevant for the study because they also appear more than 4 times in the corpus. These structures are *suj* (with an appearance of 38 times), *suj + cpred + cc* (with an appearance of 10 times) and *suj + ci + cc* (with an appearance of 6 times). These results can be graphically seen in Figure 10.

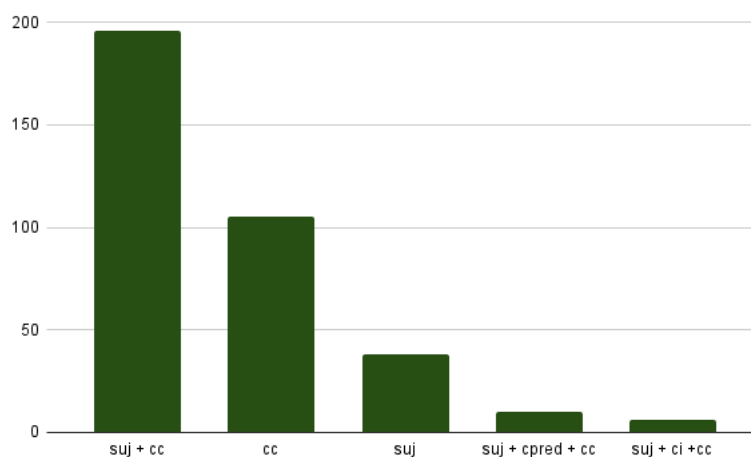


Figure 10: Syntactic structures with a longer appearance than 4 times of the sense 1 of the verb *arribar*

Lastly, the results concerning the thematic structures for sense 1 of *arribar* (11) have shown that the most common structure is *tem + tmp* (it appears 71 times in the corpus) 20. Other structures appearing more than 4 times are *des* (with an appearance of 55 times), *tem* (with an appearance of 38 times), *des + mnr* (with an appearance of 31 times), *tem + des + mnr* (with an appearance of 25 times), *tem + des + tmp* (with an appearance of 23 times), *tem + mnr + tmp* (with an appearance of 21 times), *tem + des* (with an appearance of 21 times), *tem + mnr* (with an appearance of 20 times) and *des + tmp* (with an appearance of 8 times).

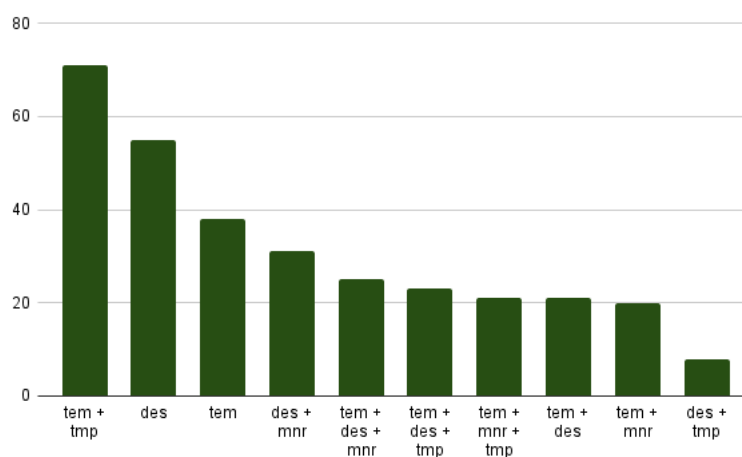


Figure 11: Thematic structures with a longer appearance than 4 times of the sense 1 of the verb *arribar*

The results obtained from the structures of this sense of the verb *arribar* have shown

that adjuncts are quite present in the sentence. Many occurrences contain not only arguments, but also adjuncts. The most common adjuncts accompanying the verb and the arguments are *tmp* (it appears 141 times) and *mnr* (it appears 111 times). However there are also many structures containing only arguments. The two most common arguments going alone with the verb are *des* (21) and *tem* (22). These two arguments appear alone with the verb 55 and 38 times, respectively. Even if *des* appears more times alone with verb than the argument *tem*, this second argument is the one appearing the most in general (it has an appearance of 257 occurrences in total).

(21) Arriben a la presidència del club (des).
They arrive at club's presidency (des).

(22) Un dia (tem) arribarà.
One day (tem) will come.

Regarding focalization, the results have proved that this verb clearly focalises the *destination* and not the *origin*. When looking at the statistics, it can be seen that the total amount of occurrences presenting the structure *arg4 cc des* is 183, while the total presenting the structure *arg3 cc ori* is 12. However it has to be noticed that in many cases neither the *origin* nor the *destination* appear. Thus, they can be easily omitted in a sports news context. Since, *arribar* focalises the destination, it can be said that when the argument *arg4 cc des* does not appear, it is a case of ellipsis. In 23, the argument *arg4 cc des* is omitted because by the context, it can be deduced that the *destination* makes reference to the new team where the player is going to play.

(23) El nou jugador grana arriba cedit fins al 30 de juny per reforçar la punta de l'atac.
The new grana player arrives on loan until 30 June to reinforce the front line.

In the analysis of this verb sense, just one difference with AnCora's analysis has been found. This difference is shared with the results obtained by the analysis of the verb *aconseguir*. This difference makes reference to the fact that according to AnCora, the arguments of *manner argM cc mnr* that go with this verb sense and are introduced by a preposition or a prepositional phrase, are introduced by the preposition *amb* (with) or the prepositional phrase *mitjançant per* (by means of). However, in this corpus other prepositions have been found too. The other prepositions introducing an (*argM cc mnr*) are *en* (in), *de* (of) (24), *per* (by), *contra* (against) and *sense* (without).

(24) Va arribar fins al màstil de mans d'un passadís humà format per nens i adults.
It reached the mast from the hands of a human corridor made up of children and adults.

As regards the results obtained from *arribar*'s sense 2, just two arguments structures can be considered because they have a frequency higher than 4 times. These structures are *arg0 + arg1 + argM* (25) and *arg1 + argM* (with frequencies of 8 and 7, respectively). A

similar thing happens with syntactic structures: two structures appear more than 4 times in the corpus. These structures are *subj + creg + cc* and *creg + cc* (with frequencies of 8 and 7, respectively). Nevertheless, only one thematic structure appears more than 4 times in the corpus: *agt + tem + adv* (it has an appearance of 6 times).

- (25) *El Club d'Esports Vendrell* (arg0 subj agt) *ha arribat a un acord de patrocini* (arg1 creg tem) *amb l'operadora de telecomunicacions Parlem Telecom* (argM cc adv).
Esports Vedrell Club (arg0 subj agt) *has reached a sponsorship agreement* (arg1 creg tem) *with the telecommunications operator Parlem Telecom* (argM cc adv).

Even if these results are not very representative, it can be said that it seems that in this type of structures the argument corresponding to the *creg tem* and an adjunct corresponding to a *cc adv* are always required. On the other hand, the subject can be omitted.

Finally, with regard to the structures of sense 3, none of them (neither arguments, syntactic nor thematic structures) have a frequency higher than 4 times. Thus, it can also be said that sense 3 is not specific to the field of sports news.

After seeing all the quantitative results, a canonical structure for the sense 1 of the verb *arribar* in a sports news context is proposed (Table 5). Even if very few occurrences have corresponded to sense 2, it has been possible to create a proposal of a canonical structure for this sense too (Table 6).

On the other hand, a canonical structure for sense 3 cannot be proposed because there have not been enough results pertaining to this sense.

Arguments structure	Syntactic structure	Thematic structure
arg1	subj	tem
arg2	ci	ben
arg3	cc	ori
arg4	cc	des
argM	cpred	atr
argM	cc	tmp
argM	cc	mnr
argM	cc	fin
argM	cc	adv

Table 5: Proposal of the canonical structure of the sense 1 of the verb *arribar* in the specialized field of sports news.

When looking at Ancora's canonical structures it can be said that our proposal of structure for sense 1 of this verb is identical to theirs.

Regarding our proposal for sense 2, it is very similar to AnCora's canonical structure of sense 3, yet the adjunct *mnr* is found in their proposal for sense 2. However, both AnCora's sense (2 and 3) correspond to the same VerbNet's class (*get-13.5.1*), the one used in this study to determine the sense 2 of the verb *arribar*. Thus, in fact, there have

Arguments structure	Syntactic structure	Thematic structure
arg0	su _j	agt
arg1	creg	tem
argM	cc	adv
argM	cc	tmp
argM	cc	mnr

Table 6: Proposal of the canonical structure of the sense 2 of the verb *arribar* in the specialized field of sports news.

been no novelties in the structure of *arribar*'s sense 2 either. However, we would like to point out that it is difficult to determine weather sentences containing this structure (25) present a specific structure or are a case of an expression containing specific lexical units.

5.3 Results of the structures of the verb *avançar*

The verb *avançar* has two general senses (Table 3 in Section 4.4). The results have shown that most of the occurrences found in the corpus belong to sense 2 (97 occurrences out of 109). On the other hand, the number of occurrences belonging to sense 1 of the verb *avançar* is quite lower (12 occurrences).

The results referring to the arguments structures of *avançar*'s sense 1 have shown that just one arguments structure appear more than 4 times in the corpus. This structure is *arg1 + argM* (it appears 5 times in the corpus).

When looking at the results referring to the syntactic structures of the same verb sense, something similar happens: just the structure *su_j + cc* appears more than 4 times in the corpus (it has a frequency of 5). Moreover, none of the thematic structures of sense 1 of the verb *avançar* exceeds the number 4 of appearances. It is because of this reason, that in this study, with the data we have at this moment, a canonical structure for sense 1 of this verb in a sports news context cannot be proposed.

On the other hand, the results from sense 2 of the verb *avançar* can be analysed. The most common arguments structure for this sense is *arg0 + arg1 + argM* (it has an appearance of 37 times) (26). 5 other arguments structures that appear more than 4 times in the corpus are *arg0 + arg1 + arg2 + argM*, *arg1 + argM*, *arg0 + arg1*, *arg1* and *arg1 + arg2* (they appear 10, 10, 10, 9 and 7 times, respectively). In Figure 12, a graphic of how many times the most frequent arguments structures appear in the corpus can be seen.

Su_j + cd + cc is by far the most common syntactic structure of sense 2 of *avançar* appearing in the corpus (it appears 50 times) (26). 3 more syntactic structures can be considered relevant because they have a higher appearance than 4 in the corpus. This structures are *cd + cc* (with an appearance of 21 times), *su_j + cd* (with an appearance of 10 times) and *cd* (with an appearance of 9 times). A graphic of the results referring to the most common syntactic structures for sense 2 of the verb *avançar* can be seen in Figure 13

In terms of thematic structures, there are 7 structures which appear more than 4 times

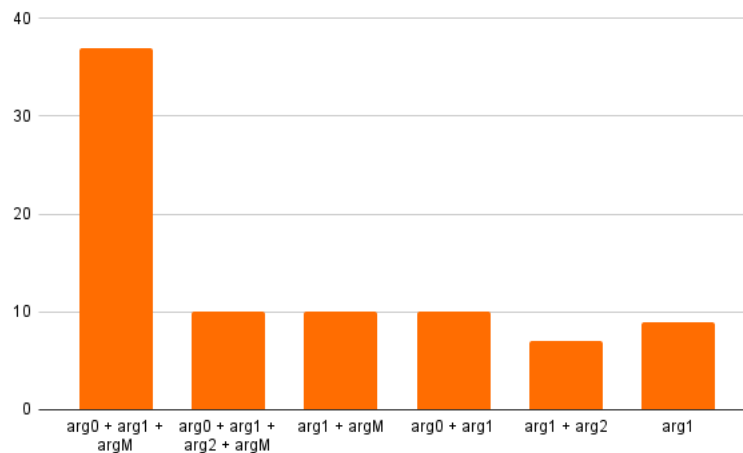


Figure 12: Arguments structures with a longer appearance than 4 times of the sense 2 of the verb *avançar*

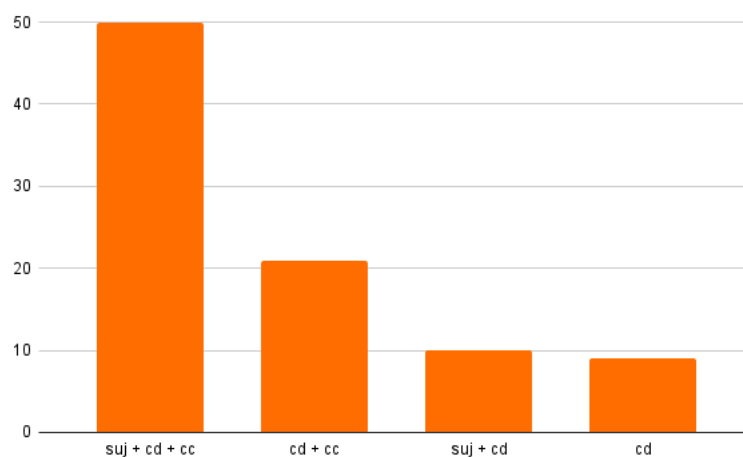


Figure 13: Syntactic structures with a longer appearance than 4 times of the sense 2 of the verb *avançar*

in the corpus (Figure 14). All of them have a similar frequency of appearance. These structures are *agt + pat + tmp* (with an appearance of 14 times) (26), *agt + pat + mnr + tmp* (with an appearance of 11 times), *agt + pat* (with an appearance of 10 times), *agt + pat + mnr* (with an appearance of 9 times), *pat* (with an appearance of 9 times), *pat + loc* (with an appearance of 7 times) and *agt + pat + loc + tmp* (with an appearance of 6 times).

- (26) *El dia del seu comiat (argM cc tmp) Bartomeu (arg0 suj agt) ja va avançar que es crearia la Superlliga (arg1 cd pat).*

The day of his farewell (argM cc tmp) *Bartomeu* (arg0 suj agt) already advanced *that the Superleague would be created* (arg1 cd pat).

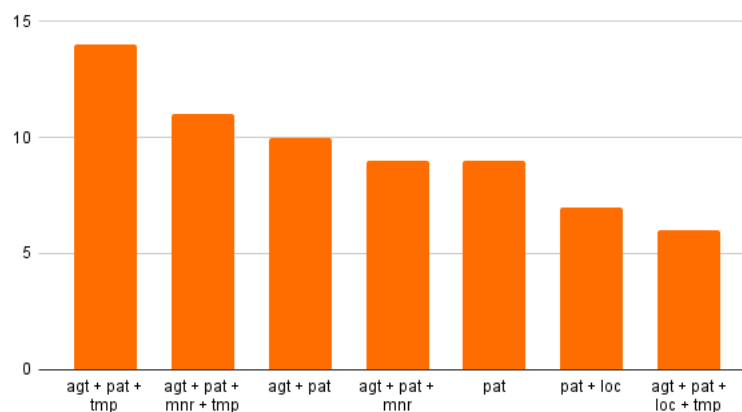


Figure 14: Thematic structures with a longer appearance than 4 times of the sense 2 of the verb *avançar*

These results have shown that, in a sports news context, sense 2 of *avançar* both the subject and the direct object are rarely omitted. What is more, the *arg1 cd pat* is almost never omitted. Another observation that can be done is that even if *avançar* is a verb of trajectory, it is rarely accompanied by an argument of *location*. Moreover, just one occurrence containing the *destination* argument (27) and another containing an argument of *origin* (28) have been found. Thus, it can be corroborated that for this verb sense and this context the trajectory component is rarely activated in the predicate.

(27) Les rondes que avancen *fins a la final de la Lliga de Campions*.
The rounds that advance *to the final of the Champions League*.

(28) *Des de l'Ajuntament* han avançat que la nova pista tindrà un total de vuit carrils.
From the town council it has been announced that the new track will have a total of eight lanes.

After analysing the results, 2 differences with Ancora's analysis of the verb sense have been found. The most relevant difference is that AnCora does not propose an structure for pronominal sentences with the verb *avançar*, such us sentence 29

(29) Els osonencs s'han avançat amb un 0-2.
Those from Osona have gone ahead with a 0-2.

However, in the corpus 35 cases of pronominal sentences belonging to the sense 2 of the verb *avançar* have been found. In this type of sentences, the reflexive pronoun has been

analysed as an *arg1 cd pat*. Thus, in 29, the element *els osonencs* has been analysed as an *arg0 suj agt* and the reflexive pronoun (*s'*) as an *arg1 cd pat*.

The second difference found is that in AnCora's canonical structure for sense 2, there is no adjunct of *manner*. However, in this corpus the adjuncts of 33 occurrences have been classified as *argM cc mnr*.

With these statistical results, a canonical structure for sense 2 of the verb *avançar* has been proposed. This structure can be seen in Table 7

Arguments structure	Syntactic structure	Thematic structure
arg0	suj	agt
arg1	cd	pat
arg2	cc	loc
argM	cc	tmp
argM	cc	mnr

Table 7: Proposal of the canonical structure of the sense 2 of the verb *avançar* in the specialized field of sports news.

In this proposal of a canonical structure for sense 2 of the verb *avançar*, neither the arguments *arg3 cc ori* nor *arg4 cc des* have been added. This is because both have appeared just one time in the corpus. Thus, we have considered that they do not belong to the inner sense 2 of the verb *avançar* in the specialized context of sports news. Two other adjuncts that have appeared in the corpus but have not been considered in the canonical structure, due to their low appearance are *cau* and *fin*.

5.4 Results of the structures of the verb *entrar*

The results collected for the verb *entrar* have shown that all the occurrences except for one belong to the sense 1 of the verb (Table 3 from Section 4.4). Thus, it can be said that sense 2 is not typical of a sports news context.

There are 7 arguments structures for sense 1 that appear more than 4 times in the corpus. These structures can be seen in Figure 15 and they are *arg2* (it appears 30 times) (30), *arg1 + argM* (it appears 21 times), *arg1 + arg2* (it appears 17 times), *arg2 + argM* (it appears 17 times), *arg1 + arg2 + argM* (it appears 16 times), *arg1* (it appears 14 times) and *argM* (it appears 7 times).

(30) Entrant a la porteria.

Entering the goal.

Regarding syntactic structures for sense 1, the most common ones are *cc* (it appears 50 times) (30), *suj + cc* (it appears 42 times), *suj* (it appears 14 times) and *suj + cpred + cc* (it appears 11 times). The frequency of these structures can be graphically seen in Figure 16.

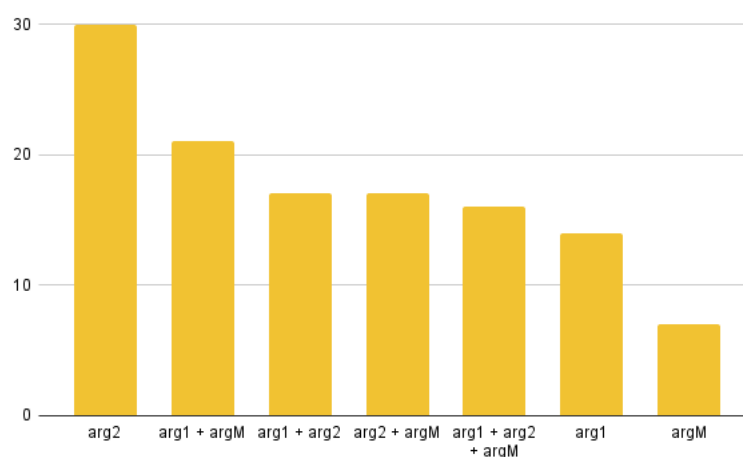


Figure 15: Arguments structures with a longer appearance than 4 times of the sense 1 of the verb *entrar*

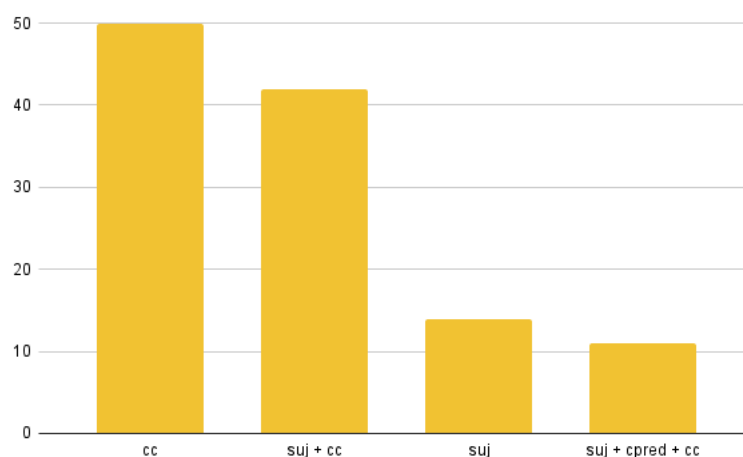


Figure 16: Syntactic structures with a longer appearance than 4 times of the sense 1 of the verb *entrar*

Finally, 8 thematic structures of *entrar*'s sense 1 appear more than 4 times in the corpus (Figure 17). These thematic structures are *loc* (with an appearance of 30 times) (30), *tem + loc* (with an appearance of 17 times), *tem* (with an appearance of 14 times), *tem + tmp* (with an appearance of 11 times), *tem + atr + mnr* (with an appearance of 8 times), *loc + tmp* (with an appearance of 8 times), *tem + loc + tmp* (with an appearance of 6 times), and *loc + mnr* (with an appearance of 5 times).

These results have proved that this verb sense clearly focalises the *location*. Moreover, many times *arg2 cc loc* is the only argument appearing with the verb. However, in some cases (42 cases) this argument has been omitted. Thus, it can be said that in a sports

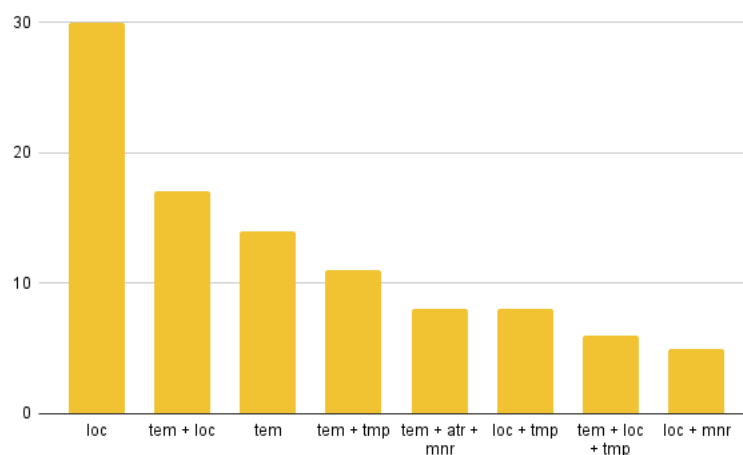


Figure 17: Thematic structures with a longer appearance than 4 times of the sense 1 of the verb *entrar*

news context an ellipsis of this argument can also occur.

After analysing these quantitative results a canonical structure for sense 1 of the verb *entrar* has been proposed. It can be seen in Table 8

Arguments structure	Syntactic structure	Thematic structure
arg1	subj	tem
arg2	cc	loc
argM	cpred	atr
argM	cc	tmp
argM	cc	mnr
argM	cc	fin

Table 8: Proposal of the canonical structure of the sense 1 of the verb *entrar* in the specialized field of sports news.

This canonical structure is quite similar to the AnCora’s structure for their sense 1 of the verb. However it has to be noticed that in AnCora’s structure for their sense 1 there is no *argM cpred atr* 31. However, looking at their canonical structure for what they consider to be sense 6, this type of argument does appear. VerbNet sense corresponding to AnCora’s sense 1 and 6 is the same sense (*escape-51.1*). It is because of this reason, we consider that the addition of the *argM cpred atr* in our canonical structure does not have to do with different senses, but with the possible realisations of the same sense.

- (31) Ha entrant *plorant* a la porteria local.
 He/she has entered the home goal *in tears*.

On the other hand, in AnCora’s structure for their sens 6, *arg3 cc ori* and *arg4 cc des* appear. However, in our structure do not. It is because no occurrences with an *arg4 cc*

des have appeared in the corpus and just one with an *arg3 cc ori* has been found (32). Thus, it seems that for the verb *entrar* in the sports domain, the locative element is just activated by an *arg2 cc loc*.

- (32) Entrant *des del darrere*.
Entering *from behind*.

5.5 Results of the structures of the verb *sortir*

All the occurrences of the verb *sortir* found in the corpus belong to the sense 1 of the verb (Table 3 from Section 4.4). Thus, as in the case of the verb *entrar* it can be said that sense 2 is not typical of a sports news context.

The most common arguments structures for sense 1 that can be seen in Figure 18 are *arg1 + argM* (it appears 34 times), *arg3* (it appears 27 times), *arg1 + arg2 + argM* (it appears 18 times), *argM* (it appears 14 times), *arg3 + argM* (it appears 11 times), *arg1* (it appears 10 times), *arg1 + arg3* (it appears 8 times), *arg1 + arg3 + argM* (it appears 7 times), *arg2 + argM* (it appears 6 times), and *arg2* (it appears 6 times).

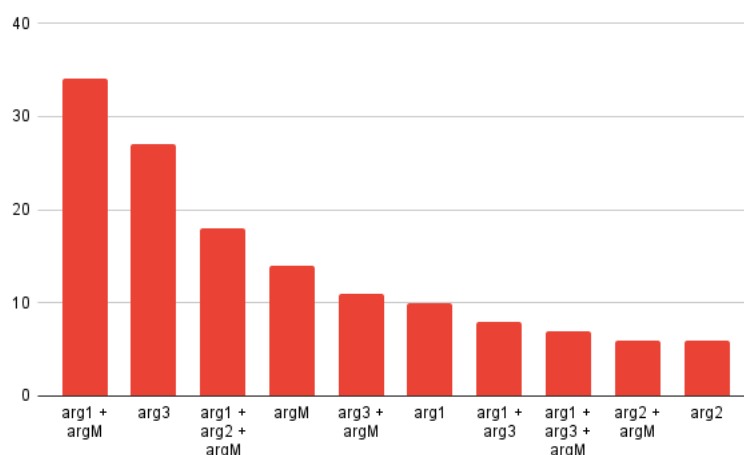


Figure 18: Arguments structures with a longer appearance than 4 times of the sense 1 of the verb *sortir*

Regarding syntactic structures for sense 1, the most common ones are *cc* (it appears 56 times), *subj + cc* (it appears 51 times), *subj + cpred + cc* (it appears 12 times), *subj* (it appears 9 times), *cpred + cc* (it appears 6 times), and *subj + ci + cpred* (it appears 5 times). The frequency of these structures can be graphically seen in Figure 19.

Finally, 7 thematic structures of *sortir*'s sense 1 appear more than 4 times in the corpus (Figure 20). These thematic structures are *ori* (with an appearance of 27 times) (33), *tem + mnr* (with an appearance of 10 times), *tem* (with an appearance of 9 times), *tem + ori* (with an appearance of 8 times), *loc* (with an appearance of 6 times), *tem + loc + tmp* (with an appearance of 5 times) and *tem + mnr + tmp* (with an appearance of 5 times).

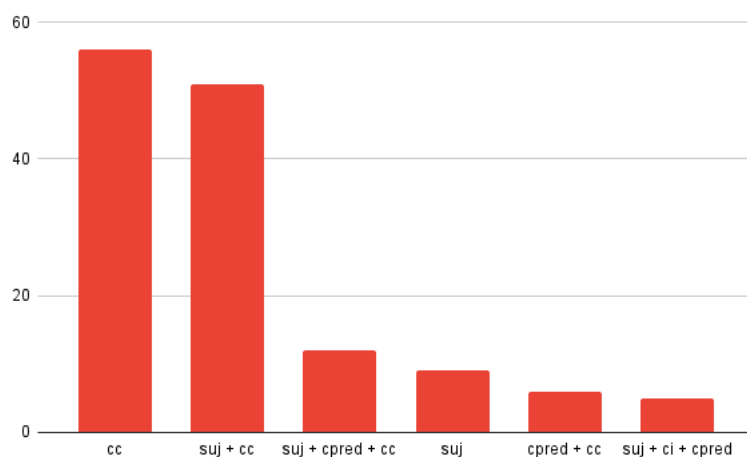


Figure 19: Syntactic structures with a longer appearance than 4 times of the sense 1 of the verb *sortir*

- (33) *Sortir de la meva zona de confort.*
Getting out of my comfort zone.

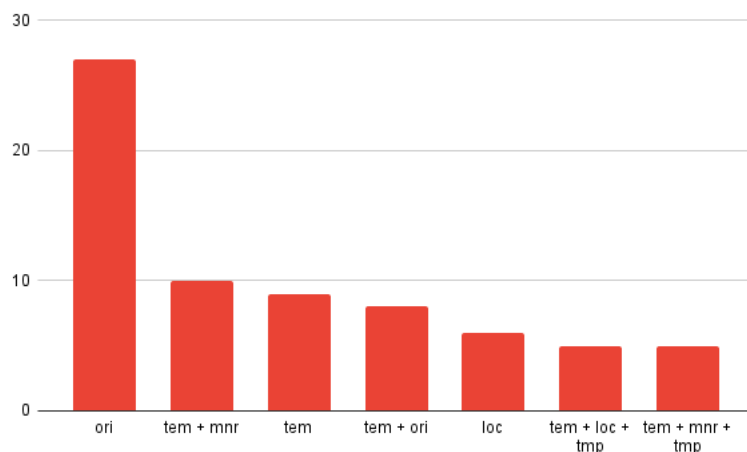


Figure 20: Thematic structures with a longer appearance than 4 times of the sense 1 of the verb *sortir*

These results have shown that, as in the case of *entrar*, this sense of *sortir* clearly focalises one argument. However, the argument this verb focalises is not the same argument as the one focalised by *entrar*. Looking at the graphics of the structures of *sortir*'s sense 1, it can be easily seen that the argument focalised is *arg3 cc ori*. Moreover, as *location* in *entrar*, the argument *origin* in *sortir* can easily be omitted (95 times out of 148 this

argument does not appear with the predicate). Thus, it can be said that when the argument *arg3 cc ori* does not appear in the sentence, it is a case of ellipsis.

Related to this previous point, the results have also shown that the argument *origin* does not appear together with the argument *location*.

One of the adjunct appearing the most is the *argM cc mnr* (it appears 37 times). However, this adjunct does not appear in any of AnCora's senses of the verb *sortir*. Thus, it could be a case of an adjunct just belonging to the verb in a context of sports news.

After analysing these quantitative results a canonical structure for sense 1 of the verb *sortir* has been created. It can be seen in Table 9

Arguments structure	Syntactic structure	Thematic structure
arg1	subj	tem
arg2	ci	ben
arg2	cc	loc
arg3	cc	ori
arg4	cc	des
argM	cpred	atr
argM	cc	tmp
argM	cc	mnr
argM	cc	fin
argM	cc	adv

Table 9: Proposal of the canonical structure of the sense 1 of the verb *sortir* in the specialized field of sports news.

In this proposal of a canonical structure for the structure of *sortir*'s sense 1, there is an argument that does not appear in the most common structures presented. This argument is the *arg2 ci ben* (34). Despite of this, it has been introduced in the canonical structure because it appears more than 4 times in the corpus (it appears 7 times in total). As in sense 2 of *arribar*, for sense 1 of the verb *aconseguir*, it is hard to determine weather phrases, such as *sortir bé* in 34, present a specific structure or, in fact, they are a case of a singular expression.

- (34) Les coses *els* surtin bé.
 Things go well *for them*.

Another argument that does not appear in the most common structures found in the corpus, but appears in our proposal of the canonical structure is the *arg4 cc des*. As in the previous case, it appears more than 4 times (it appears 5 times in the corpus). Thus, it can be considered to belong to the canonical structure of sense 1 of the verb *sortir* in a sports news context.

6 Preliminary experiment: Applying the analysis to information extraction

After knowing how the 5 verbs studied (*aconseguir*, *arribar*, *avançar*, *entrar* and *sortir*) behave in a sports news context, a preliminary experiment has been designed to capture the event argument roles. With this preliminary experiment we want to see if we can create some templates that will allow us to carry out information extraction, exactly event argument extraction (EAE), by using the structures of the corpus we have already analysed.

The most decisive role in the verb's sense activation have been selected. Thus, in the case of *aconseguir*, we have selected the *patient*, in *arribar* we have selected the *destination*, in the case of *avançar* we have selected the *agent*, in the case of *entrar* we have selected the *location*, and finally, in the case of *sortir* we have selected the *origin*. It is true that for the verb *avançar* we could have selected the *patient* as well, because it also plays an important role on activating the verb sense. However, as we already selected it for *aconseguir*, for *avançar* we have considered the *agent*.

In order to create the templates we follow the methodology of Sainz et al. (2022) which use textual entailment for EAE. Based on that framework, in our work we firstly, present the input (a sentence from our corpus that contains the argument in its context), secondly, the argument itself is defined, and thirdly the template containing the verbalization is proposed. To obtain the entailment probability for each verbalization, we have used the xlm-roberta-large-xnli⁷ model. This model can be used for 15 different languages, yet Catalan is not one of them. Since the base model (xlm-roberta-large (Conneau et al., 2019)) was pre-trained on 100 different languages, one of them being Catalan, we have thought that the model could show results for Catalan as well. Moreover, the Hugging Face description of xlm-roberta-large-xnli⁸ indicates that the model could show effectiveness on the 100 languages of the xlm-roberta-large.

For each sentence of each verb, we propose a correct templates vs. some adverse templates to see if the probability results are coherent, expecting to have high entailment probabilities with the correct templates and low probabilities with the adverse ones:

- Grammatical templates
- Agrammatical templates
- Negative templates
- Templates containing prepositions, pronouns or adverbs different from the ones in the principal sentence.

Exactly, the adverse templates that are proposed for each sentence of every verb are

⁷A pre-trained model distributed by Hugging Face: <https://huggingface.co/joeddav/xlm-roberta-large-xnli>

⁸<https://huggingface.co/joeddav/xlm-roberta-large-xnli>

- For the verb *aconseguir*, 3 type of adverse templates are proposed: one containing negation, another containing agrammaticality and another one containing both, negation + agrammaticality (10).
- For the verb *arribar*, 6 type of adverse templates are proposed: one containing negation, another containing agrammaticality, another containing negation + agrammaticality, another containing prepositions different from the ones in the original sentences, another containing prepositions different from the ones in the original sentences + agrammaticality, and another one containing agrammaticality + negation + prepositions different from the ones in the original sentences (11).
- For the verb *avançar*, 5 type of adverse templates are proposed: one containing negation, another containing agrammaticality, another containing pronouns different from the ones in the original sentences, another containing negation + agrammaticality + pronouns different from the ones in the original sentences, and another containing a preposition which did not appear in the original sentence (12).
- For the verb *entrar*, 5 type of adverse templates are proposed: one containing negation, another containing agrammaticality, another containing prepositions different from the ones in the original sentences, another containing negation + agrammaticality + prepositions different from the ones in the original sentences, and another containing an adverb which did not appear in the original sentence (13).
- Finally, for the verb *sortir*, 4 type of adverse templates are proposed: one containing prepositions different from the ones in the original sentences, another containing negation + prepositions different from the ones in the original sentences, another containing agrammaticality + prepositions different from the ones in the original sentences, and another containing negation + agrammaticality. (14).

In the following tables (10, 11, 12, 13 and 14) the probabilities for each of the hypothesis can be seen. These tables contain the information related to the types of templates proposed for each sentence, the templates themselves, the main sentences (to which we compare the verbalizations), the verbalizations and the entailment probability for each verbalization. Each of the tables contain all this information for just one verb.

In the tables, it can be seen that some of the prepositions used in the templates have a similar meaning than the original preposition of the main sentence, and others have a completely different meaning. With this, we wanted to see if those templates containing prepositions which did not alter the sentence's meaning got higher entailment probabilities than the ones containing prepositions that changed the meaning of the sentence completely. For the prepositions *des de* (from) and *de* (from) in the sentences containing the verb *arribar*, we expected to have very low probabilities. However, they have shown entailment probabilities higher than 0.9992 (Table 11).

The same happens with the prepositions *de* (from), *des de* (from), *amb* (with) and *per* (for) for the verb *entrar* (they have shown entailment probabilities higher than 0.9994)(Table 13) and the preposition *a* (to) and *cap a* (towards) for the verb *sortir* (which have shown

entailment probabilities higher than 0.9987) (Table 14). The only preposition which we expected to perform a low probability and which has performed it is the preposition *contra* (against), for the verb *entrar* (it has shown an entailment probability of 0.0561) (Table 13).

Type	Template	Sentence + Verbalization	Probability
Agrammatical	"Un equip ha aconseguit {arg}."	L'Igalada Rigat ha aconseguit els tres punts. Un equip ha aconseguit els tres punts.	0.9998
	"Un equip aconseguir {arg}."	L'Igalada Rigat ha aconseguit els tres punts. Un equip aconseguir els tres punts.	0.9997
Negative	"Un equip no ha aconseguit {arg}."	L'Igalada Rigat ha aconseguit els tres punts. Un equip no ha aconseguit els tres punts.	0.0017
Negative Agrammatical	"Un equip no aconseguir {arg}."	L'Igalada Rigat ha aconseguit els tres punts. Un equip no aconseguir els tres punts.	0.0012
Agrammatical	"Algú aconseguiria {arg} llavors."	Xaus aconseguiria l'empat després d'una jugada per la banda dreta. Algú aconseguiria l'empat llavors.	0.9998
	"Ell aconseguir {arg} llavors."	Xaus aconseguiria l'empat després d'una jugada per la banda dreta. Ell aconseguir l'empat llavors.	0.9997
Negative	"no aconseguiria {arg} en aquell moment."	Xaus aconseguiria l'empat després d'una jugada per la banda dreta. no aconseguiria l'empat en aquell moment.	0.0003
Negative Agrammatical	"no aconseguir {arg} després d'una jugada per la banda dreta."	Xaus aconseguiria l'empat després d'una jugada per la banda dreta. no aconseguir l'empat després d'una jugada per la banda dreta.	0.0017

Table 10: Entailment probability of verbalizations containing the argument *patient* for the verb *aconseguir*.

Type	Template	Sentence + Verbalization	Probability
Agrammatical	"Un equip arriba a {arg}."	El Rayo arriba a les semifinals. Un equip arriba a les semifinals.	0.9997
	"Un equip arribar a {arg}."	El Rayo arriba a les semifinals. Un equip arribar a les semifinals.	0.9997
Negative	"Un equip no arriba a {arg}."	El Rayo arriba a les semifinals. Un equip no arriba a les semifinals.	0.0035
Negative Agrammatical	"Un equip no arribar a {arg}."	El Rayo arriba a les semifinals. Un equip no arribar a les semifinals	0.0032
Different preposition (similar meaning)	"Un equip arriba fins {arg}."	El Rayo arriba a les semifinals. Un equip arriba fins les semifinals.	0.9997
Different preposition (different meaning)	"Un equip arriba des de {arg}."	El Rayo arriba a les semifinals. Un equip arriba des de les semifinals.	0.9992
Different preposition (different meaning) Agrammatical	"Un equip arribar de {arg}."	El Rayo arriba a les semifinals. Un equip arribar de les semifinals.	0.9995
Agrammatical	"una cosa ha arribat al {arg}."	El partit ha arribat al minut cinquanta-vuit. una cosa ha arribat al minut cinquant-vuit.	0.9994
	"el partit arribar al {arg}."	El partit ha arribat al minut cinquanta-vuit. el partit arribar al minut cinquanta-vuit.	0.9996
Negative	"alunga cosa no ha arribat al {arg}."	El partit ha arribat al minut cinquanta-vuit. alguna cosa no ha arribat al minut cinquanta-vuit.	0.0041
Negative Agrammatical	"això no arribar des del {arg}."	El partit ha arribat al minut cinquanta-vuit. això no arribar des del minut cinquanta-vuit.	0.0097
Different preposition (different meaning)	"això arriba fins al {arg}."	El partit ha arribat al minut cinquanta-vuit. això arriba fins al minut cinquanta-vuit.	0.9996
Different preposition (similar meaning)	"això arriba des del {arg}."	El partit ha arribat al minut cinquanta-vuit. això arriba des del minut cinquanta-vuit.	0.9995
Different preposition (different meaning)	"això arriba del {arg}."	El partit ha arribat al minut cinquanta-vuit. això arriba del minut cinquanta-vuit.	0.9995

Table 11: Entailment probability of verbalizations containing the argument *destination* for the verb *arribar*.

Type	Template	Sentence + Verbalization	Probability
Agrammatical	"{arg} avançava algú."	Jan Escala avançava el Calafell. Jan Escala avançava algú.	0.9985
	"{arg} avançava un equip."	Jan Escala avançava el Calafell. Jan Escala avançava un equip.	0.9970
Negative	"{arg} avançar algú."	Jan Escala avançava el Calafell. Jan Escala avançar algú.	0.9982
	"{arg} no avançava un equip."	Jan Escala avançava el Calafell. Jan Escala no avançava un equip.	0.0011
Agrammatical	"{arg} no avançar un equip."	Jan Escala avançava el Calafell. Jan Escala no avançar un equip.	0.0016
	Different pronoun (different meaning)	"{arg} avançava ningú."	Jan Escala avançava el Calafell. Jan Escala avançava ningú.
Different pronoun (different meaning)		"{arg} va avançar alguna cosa."	l'equip visitant va avançar la seva línia de pressió. l'equip visitant va avançar alguna cosa.
	Agrammatical	"{arg} va avançar res."	l'equip visitant va avançar la seva línia de pressió. l'equip visitant va avançar res.
Negative		"{arg} avançar algú."	l'equip visitant va avançar la seva línia de pressió. l'equip visitant avançar algú.
	Negative	"{arg} no va avançar res."	l'equip visitant va avançar la seva línia de pressió. l'equip visitant no va avançar res.
Negative		"{arg} no va avançar la seva línia de pressió."	l'equip visitant va avançar la seva línia de pressió. l'equip visitant no va avançar la seva línia de pressió.
	Preposition added (different meaning)	"{arg} va avançar cap a alguna cosa."	l'equip visitant va avançar la seva línia de pressió. l'equip visitant va avançar cap a alguna cosa.

Table 12: Entailment probability of verbalizations containing the argument *agent* for the verb *avançar*.

Type	Template	Sentence + Verbalization	Probability
Different preposition (different meaning)	"Algú havia entrat al {arg} llavors."	L'ariet uruguaià havia entrat al camp just després del descans. Algú havia entrat al camp llavors.	0.9995
	"Algú havia entrat del {arg} llavors."	L'ariet uruguaià havia entrat al camp just després del descans. Algú havia entrat del camp llavors.	0.9996
Different preposition (different meaning)	"Algú havia entrat des del {arg} llavors."	L'ariet uruguaià havia entrat al camp just després del descans. Algú havia entrat des del camp llavors.	0.9996
	Negative	"Algú no havia entrat al {arg} llavors."	L'ariet uruguaià havia entrat al camp just després del descans. Algú no havia entrat al camp llavors.
Agrammatical	"Algú entrar al {arg} llavors."	L'ariet uruguaià havia entrat al camp just després del descans. Algú entrar al camp llavors.	0.9996
	Negative	"Algú no entrar al {arg} llavors."	L'ariet uruguaià havia entrat al camp just després del descans. Algú no entrar al camp llavors.
Different adverb (different meaning)		"Algú havia entrat {arg} mai."	L'ariet uruguaià havia entrat al camp just després del descans. Algú havia entrat camp mai.
	Different preposition (different meaning)	"van entrar a {arg}."	4 jugadors de la base ja van entrar a les convocatòries. van entrar a les convocatòries.
Different preposition (different meaning)		"van entrar amb {arg}."	4 jugadors de la base ja van entrar a les convocatòries. van entrar amb les convocatòries.
	Different preposition (different meaning)	"van entrar contra {arg}."	4 jugadors de la base ja van entrar a les convocatòries. van entrar contra les convocatòries.
Different preposition (different meaning)		"van entrar per {arg}."	4 jugadors de la base ja van entrar a les convocatòries. van entrar per les convocatòries.

Table 13: Entailment probability of verbalizations containing the argument *location* for the verb *entrar*.

Type	Template	Sentence + Verbalization	Probability
	"els jugadors surten des de {arg}."	Els jugadors surten des de la banqueta. els jugadors surten des de la banqueta.	0.9994
Different preposition (similar meaning)	"ells surten de {arg}."	Els jugadors surten des de la banqueta. ells surten de la banqueta.	0.9995
Different preposition (different meaning)	"ells surten a {arg}."	Els jugadors surten des de la banqueta. ells surten a la banqueta.	0.9993
Different preposition (different meaning)	"ells surten fins a {arg}."	Els jugadors surten des de la banqueta. ells surten fins a la banqueta.	0.9987
Negative			
Different preposition (different meaning)	"els jugadors no surten a {arg}."	Els jugadors surten des de la banqueta. els jugadors no surten a la banqueta.	0.0003
Agrammatical			
Different preposition (different meaning)	"ells sortir a {arg}."	Els jugadors surten des de la banqueta. ells sortir a la banqueta.	0.9991
	"alguna cosa sortirà de {arg}."	El recorregut sortirà de la plaça dels Gemans Trillas de Coma-ruga. alguna cosa sortirà de la plaça dels Gemans Trillas de Coma-ruga.	0.9995
Different preposition (similar meaning)	"ell sortirà des de {arg}."	El recorregut sortirà de la plaça dels Gemans Trillas de Coma-ruga. ell sortirà des de la plaça dels Gemans Trillas de Coma-ruga.	0.9996
Agrammatical			
Different preposition (different meaning)	"ell sortir cap a {arg}."	El recorregut sortirà de la plaça dels Gemans Trillas de Coma-ruga. ell sortir cap a la plaça dels Gemans Trillas de Coma-ruga.	0.9987
Different preposition (different meaning)	"alguna cosa sortirà a {arg}."	El recorregut sortirà de la plaça dels Gemans Trillas de Coma-ruga. alguna cosa sortirà a la plaça dels Gemans Trillas de Coma-ruga.	0.9994
Negative			
Different preposition (different meaning)	"alguna cosa no sortirà a {arg}."	El recorregut sortirà de la plaça dels Gemans Trillas de Coma-ruga. alguna cosa no sortirà a la plaça dels Gemans Trillas de Coma-ruga.	0.0149
Negative			
Agrammatical	"ell no sortir de {arg}."	El recorregut sortirà de la plaça dels Gemans Trillas de Coma-ruga. ell no sortir de la plaça dels Gemans Trillas de Coma-ruga.	0.0015

Table 14: Entailment probability of verbalizations containing the argument *origin* for the verb *sortir*.

As it can be seen in all the previous tables, some good results have been achieved with negative verbalizations when the principal sentence was positive, because they have shown very low entailment probabilities. In Table 10 it can be seen that the highest probability for these type of verbalizations for the verb *aconseguir* is 0.0017, in Table 11 it can be seen that the highest probability for the verb *arribar* is 0.0097, in Table 12 it can be seen that the highest probability for the verb *avançar* is 0.0016, in Table 13 it can be seen that the highest probability for the verb *entrar* is 0.0034, and in Table 14 it can be seen that the highest probability for the verb *sortir* is 0.0149.

On the other hand, we would also like to point out that, as in Sainz et al. (2022), agrammaticalization has not have influence on the entailment probability either. When looking at verbalizations conatining agrammitaclity, but not negation, it can be seen that in Table 10 (*aconseguir*) the lowest probability is 0.9997, in Table 11 (*arribar*) it is 0.9995, in Table 12 (*avançar*) it is 0.9982, in Table 13 (*entrar*), it is 0.9996, and in Table 14 (*sortir*), the lowest probability is 0.9987.

Unfortunately, we can conclude that this experiment has not worked. This might be due to 2 different aspects: The first one is that the model used for the experiment was not trained on Catalan. And the second aspect is that in most of the verbalizations we have focused on the meaning changed by prepositions. Since prepositions are grammatical words, which by themselves they do not have much meaning, some language models struggle on understanding these type of functional words (Kim et al., 2019). Thus, it could be the case of this language model too. When there will be more resources for Catalan, this experiment can be taken up again.

7 Conclusion and Future Work

As far as we now, this has been the first study on specialized use of verbs in Catalan in the sports domain. With this study, we have been able to analyse how some trajectory verbs behave in a corpus specialised in sport and we have been able to detect some differences between the general use of the analysed verbs (the one captured in AnCora's lexicon) and their use in the specialised context. We have seen that there are verb senses which appear most often in the sports domain, than other senses that rarely appear or do not even appear in the corpus. We have been able to establish which senses typically appear in the sports domain. These senses are the ones related to the VerbNet classes *obtain-13.5.2* (for *aconseguir*), VerbNet's *escape-51.1* (for *arribar*), VerbNet's *roll-51.3.1* (for *avançar*), VerbNet's *escape-51.1 for entrar* and VerbNet's *appear-48.1.1* for *sortir*.

With this study we have also detected some new structures that did not appear in AnCora: A new structure containing an adjunct *co-agt* has been found for the verb *aconseguir* and a pronominal structure has been found for sense 2 of the verb *avançar*.

Finally, we have been able to see some differences related to adjuncts: for sense 1 of *sortir* and sense 2 of *avançar*, adjuncts of manner have been seen in the corpus. However, in the description of AnCora, such adjuncts do not appear for these verb senses. Same happens with the locative adjunct and the verb *aconseguir*. The adjunct has appeared in the corpus but not in the description of AnCora. On the other hand, the adjunct *argM cpred atr* appears in AnCora for the verb *aconseguir*, yet it has not appeared in our corpus. The last difference related to adjuncts that we have seen in this study is that in AnCora *aconseguir* and *arribar*'s sense 2, the *argM cc mnr* is introduced by the preposition *amb* (with). However, in the corpus we have found other prepositions introducing the adjunct for the same verbs.

Regarding omission, we have been able to see that in a sports context, some arguments of some verbs are omitted. Moreover, in many cases we have seen that these arguments are the ones which determine the trajectory (*arg2 cc loc*, *arg3 cc ori* and *arg4 cc des*). In many cases, we have been able to see that the omission was a case of ellipsis, and not underspecification, because the argument omitted in the sentences was the argument which is normally focalised by the verb. We have seen that for the verb *arribar*, the argument which has suffered ellipsis is *arg4 cc des*, the argument for the verb *entrar* has been *arg2 cc loc* and for the verb *sortir*, it has been *arg3 cc ori*. Since *aconseguir* and *avançar* do not focalise any argument related to trajectory, no cases of omission of *origin*, *destination* or *location* by ellipsis have been found for such verbs.

With all this information in mind, we have been able to create canonical structures for each verb sense belonging to the sports domain. However, we would like to point out that we have had some challenges on selecting some arguments for the canonical structures. These arguments appeared in AnCora's canonical structure, yet they did not appear much in the corpus, or did not appear at all. In the end, we have opted to just put those arguments which had a representation in the corpus of 5 occurrences, at least.

Another challenge has been deciding whether some phrases, such as *arribar a un acord* (to reach an agreement) and *sortir bé o malament* (to go right or wrong), belonged to a

specific structures or they were a case of a concrete syntagmatic realization.

One of our sub-objectives of our work was to see if the analysed structures can help to perform tasks on event argument extraction (EAE). However, due to lack of resources, we have only performed a preliminary experiment on a set of templates using a language model which was not trained for Catalan. It is because of this reason that a proposal for future work would be an exhaustive experiment using a language model trained for Catalan. Moreover, an analysis of why language models struggle with functional words, such as prepositions, and an analysis of why agrammaticality is not detected by *xlm-roberta-large-xnli* could be carried out too.

Moreover, in the future, all these proposals could be used to create an automatic database annotator specialised in the sports domain. Moreover, this could also help to create sport news generators or improve automatic translators in the sports domain. In order to carry out these proposals, the corpus could be extended and other verbs from the class could be annotated as well.

References

- Andreína Adelstein. Unidad léxica y valor especializado: estado de la cuestión y observaciones sobre su representación. *Estudios de lingüística del español*, 31:0001–256, 2001.
- Andreína Adelstein and María Teresa Cabré. Representación lexicográfica y terminográfica de las unidades terminológicas. In *Terminologia e indústrias da língua: actas do VII Simpósio Ibero-Americano de Terminologia*, pages 103–116. Instituto de Linguística Teórica e Computacional, 2003.
- Andreína Adelstein et al. *Unidad léxica y significado especializado: modelo de representación a partir del nombre relacional madre*. Universitat Pompeu Fabra, 2007.
- Olga Babko-Malaya. Propbank annotation guidelines. URL: <http://verbs.colorado.edu>, 2005.
- David Stéphane Belemkoabga, Aurélien Bossard, Abdallah Essa, Christophe Rodrigues, and Kévin Sylla. Neural network-based generation of sport summaries: A preliminary study. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 147–154, Held Online, September 2021. INCOMA Ltd. URL <https://aclanthology.org/2021.ranlp-1.18>.
- Manuel Bertran, Oriol Borrega, M Antònia Martí, and Mariona Taulé. Ancorapipe: A new tool for corpora annotation. Technical report, Working paper 1: TEXT-MESS 2.0 (Text-Knowledge 2.0). Available at [http ...](http://...), 2010.
- Jean-Paul Boons. La notion sémantique de déplacement dans une classification syntaxique des verbes locatifs. *Langue française*, (76):5–40, 1987.
- María Teresa Cabré. La terminología: Representación y comunicación. 1999.
- María Teresa Cabré. Sumario de principios que configuran la nueva propuesta teórica. (2001). *La terminología científico-técnica: reconocimiento, análisis y extracción formal y semántica*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, pages 19–25, 2001.
- María Teresa Cabré and Andreína Adelstein. *¿Es la terminología lingüística aplicada?*, pages 387–393. Univerbook, S.L, 2001.
- M Teresa Cabré i Castellví. *La terminología: representación y comunicación: elementos para una teoría de base comunicativa y otros artículos*. Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra., 1999.
- Anna Joan Casademont. On the elements activating the transmission of specialized knowledge in verbs. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 20(1):92–116, 2014.

- José Luis Cifuentes Honrubia. Sintaxis y semántica del movimiento: aspectos de gramática cognitiva. Technical report, INSTITUTO DE CULTURA [JUAN GIL-ALBERT], 1999.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019. URL <http://arxiv.org/abs/1911.02116>.
- Jacqueline Dervillez-Bastuji. *Structures des relations spatiales dans quelques langues naturelles: introduction à une théorie sémantique*, volume 13. Librairie Droz, 1982.
- David Dowty. Thematic proto-roles and argument selection. *language*, 67(3):547–619, 1991.
- Markus Egg. Semantic underspecification. *Language and Linguistics Compass*, 4(3):166–181, 2010.
- Ch. Leclère J.-P. Boons, A. Guillet. La structure des phrases simples en français. 1976.
- Ray Jackendoff. Toward an explanatory semantic representation. *Linguistic inquiry*, 7(1): 89–150, 1976.
- Ray S Jackendoff. Semantic structures. 1990.
- Christopher R Johnson, Myriam Schwarzer-Petruck, Collin F Baker, Michael Ellsworth, Josef Ruppenhofer, and Charles J Fillmore. Framenet: Theory and practice. Technical report, International Computer Science Institute, 2016.
- Kyo Kageura and Bin Umino. Methods of automatic term recognition: A review. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 3(2):259–289, 1996.
- Najoung Kim, Roma Patel, Adam Poliak, Alex Wang, Patrick Xia, R. Thomas McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. Probing what different NLP tasks teach machines about function word comprehension. *CoRR*, abs/1904.11544, 2019. URL <http://arxiv.org/abs/1904.11544>.
- Paul Kingsbury, Martha Palmer, and Mitch Marcus. Adding semantic annotation to the penn treebank. In *Proceedings of the human language technology conference*, pages 252–256. San Diego, California, 2002.
- Karin Kipper, Martha Palmer, and Owen Rambow. Extending propbank with verbnet semantic predicates. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas (AMTA-2002)*, Tiburon, CA, USA, October, pages 6–12, 2002.
- Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. A large-scale extension of verbnet with novel verb classes. In *Proceedings of the 12th EURALEX International Congress, Turin, Italy*, 2006.

- Béatrice Lamiroy. *Léxico y gramática del español: Estructuras verbales de espacio y de tiempo*, volume 2. Anthropos Editorial, 1991.
- Ronald W Langacker. *Foundations of cognitive grammar. Part 1*, volume 1. Stanford university press, 1987.
- Dany Laur. La relation entre le verbe et la préposition dans la sémantique du déplacement. *Langages*, (110):47–67, 1993.
- Alessandro Lenci, Nuria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, et al. Simple: A general framework for the development of multilingual lexicons. *International Journal of Lexicography*, 13(4):249–263, 2000.
- Pierre Lerat. Qu'est-ce qu'un verbe spécialisé? le cas du droit. *Cahiers de lexicologie*, (80):201–211, 2002.
- Beth Levin. *English verb classes and alternations: A preliminary investigation*. University of Chicago press, 1993.
- Beth Levin. Aspect, lexical semantic properties, and argument expression. In *Annual Meeting of the Berkeley Linguistics Society*, volume 26, pages 413–429, 2000.
- Beth Levin and Malka Rappaport Hovav. *The semantic determinants of argument expression: A view from the English resultative construction*. na, 2004.
- Beth Levin, Malka Rappaport Hovav, and Samuel Jay Keyser. *Unaccusativity: At the syntax-lexical semantics interface*, volume 26. MIT press, 1995.
- M-C L'Homme. Le statut du verbe en langue de spécialité et sa description lexicographique. *Cahiers de lexicologie*, 73(2):61–84, 1998.
- Changliang Li, Yixin Su, Ji Qi, and Min Xiao. Using gan to generate sport news from live game stats. In *International Conference on Cognitive Computing*, pages 102–116. Springer, 2019.
- Adela Ljajić, Ertan Ljajić, Petar Spalević, Branko Arsić, and Darko Vučković. Sentiment analysis of textual comments in field of sport. In *24th International Electrotechnical and Computer Science Conference (ERK 2015), IEEE, Slovenia*, 2015.
- Mercè Lorente. Les unitats lèxiques verbals dels textos especialitzats. redefinició d'una proposta de classificació. *Estudis de lingüística i de lingüística aplicada en honor de M. Teresa Cabré Castellví*, 2:365–380, 2007.
- Mercè Lorente Casafont. Tipología verbal y textos especializados. In *Cuestiones conceptuales y metodológicas de la lingüística*, pages 143–153. Servicio de Publicaciones= Servizo de Publicacións, 2001.

- Mercè Lorente Casafont. Terminología y fraseología especializada: del léxico a la sintaxis. In *Panorama actual de la terminología*, pages 159–180. Comares, 2002b.
- Mercè Lorente Casafont. Verbos y discurso especializado. *Estudios de Lingüística del español*, 16, 2002a.
- Peter Hugoe Matthews. *The concise Oxford dictionary of linguistics*. Oxford University Press, 2014.
- Kurt McKee. feedparser - parse atom and rss feeds in python. 2021.
- Maite Melero, Blanca C. Figueras, Mar Rodríguez, and Marta Villegas. European language equality: Report on the catalan language. 2022.
- George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
- Darija Omrčen. Analysis of gender-fair language in sport and exercise. *Rasprave: časopis Instituta za hrvatski jezik i jezikoslovlje (Zagreb)*, 43:143–161, 2017.
- Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106, 2005.
- Janine Pimentel. Description de verbes juridiques au moyen de la sémantique des cadres. *Terminologie/Ontologie: Théories et applications (Toth 2011)*, pages 145–166, 2012.
- James Pustejovsky. The generative lexicon. cambridge (mass.), 1995.
- Malka Rappaport and Beth Levin. The lexical semantics of verbs of motion. i. roca, ed., thematic structure: its role in grammar. *Mouton de Gruyter*, pages 247–269, 1992.
- Malka Rappaport Hovav and Beth Levin. Building verb meanings. *The projection of arguments: Lexical and compositional factors*, pages 97–134, 1998.
- Leonard Richardson. Beautiful soup documentation. *April*, 2007.
- Oscar Sainz, Itziar Gonzalez-Dios, Oier Lopez de Lacalle, Bonan Min, and Eneko Agirre. Textual entailment for event argument extraction: Zero-and few-shot with multi-source learning. *arXiv preprint arXiv:2205.01376*, 2022.
- Sunita Sarawagi. *Information extraction*. Now Publishers Inc, 2008.
- Jeffrey O Segrave, Katherine L McDowell, and James G King. Language, gender, and sport: A review of the research literature. *Sport, rhetoric, and gender*, pages 31–41, 2006.
- F. Serzisko. On bounding in ik. b. rudzka-ostyn, ed., topics in cognitive linguistics. pages 429–445, 1988.

- Leonard Talmy. Lexicalization patterns: Semantic structure in lexical forms. *Language typology and syntactic description*, 3(99):36–149, 1985.
- Mariona Taulé, M Antònia Martí, and Oriol Borrega. Ancora 2.0: Argument structure guidelines for catalan and spanish. 2011.
- Gloria Vázquez, Ana Fernández, and M ANTÒNIA Martí. Clasificación verbal. *Alternancias de diátesis. Quaderns de Sintagma*, 3, 2000.
- Xiaozhi Wang, Ziqi Wang, Xu Han, Zhiyuan Liu, Juanzi Li, Peng Li, Maosong Sun, Jie Zhou, and Xiang Ren. HMEAE: Hierarchical modular event argument extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5777–5783, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1584. URL <https://aclanthology.org/D19-1584>.
- Jan Wijffels. Udpipes natural language processing - text annotation. 2022.

A List of Catalan trajectory verbs

Abaixar, abalançar-se, abandonar, abocar, abonar, abraonar-se, abreviar, aclamar, acomiadar, acompanyar, aconseguir, aconsellar, acudir, adjudicar, advertir, afaitar, aixecar, alçar, allunyar, allunyar-se, anar, anar-se'n, anticipar, anunciar, apartar, apartar-se, avançar, apoderar-se, aportar, apregonar, anpregonar-se, apropar, apropar-se, apropiar-se, aprovisionar, aproximar, aproximar-se, arrebatat, arrambar, arrambar-se, arriar, arribar, arriar, arriar-se, arrossegar, arrossegar-se, arxivar, ascendir, assignar, atansar, atansar-se, atènyer, aterrar, atorgar, baizar, banderejar, bolcar, bordejar, botar, buidar, caminar, cantar, carregar, carregar, carretejar, catapultar, caure, cedir, citar, comadrejar, comentar, comissar, comprar, comunicar, concedir, conducir, conferenciar, conferir, confessar, confiscar, consiltar, contar, contestar, conversar, córrer, costejar, creuar, cridar, dardar, deambular, declara, decomissar, deixar, deportar, derramar, descarregar, desertar, detenir, detenir-se, dialogar, dictar, difondre, dipositar, dir, dirigir-se, discursejar, discutir, distribuir, divulgar, donar, elevar, embarcar, embargar, embeinar, embocar, embossar, embotellar, embutxacar, emetre, emmagatzemar, empetar, emperxar, empresonar, encaixar, encaixonar, encastar, endentar, endinsar, endinsar-se, enfilear, enfonsar, enfonsar-se, enforollar, enfundar, engabiar, engolar-se, engranar, ensacar, enterrar, entrar, entregar, envasar, enviar, errar, esbarriar, escalar, escamotejar, escampar, escapar, escapar-se, escapolar-se, escolar-se, escoltar, escórrer-se, esmentar, esmunyir-se, espargir, esquiar, esquitllar-se, etzibar, evacuar, expedir, explicar, expulsar, extirpar, extraditar, facilitar, ficar, fugar-se, fugir, furtar, gatejar, gitar, graponar, guardar, guillar, heure, hissar, importar, incrustar, indemnitzar, indicar, informar, injectar, intercalar, internar-se, introduir, introduir-se, juxtaposar, llançar, llançar-se, llegar, llegir, llençar, lliurar, manifestar, marxar, mormolar, moure, mudar-se, muntar, muntar-se, murmurar, mussitar, narrar, navegar, nedar, obsequiar, pobternir, pagar, parlar, partir, passar, passejar, pedalejar, penetrar, penjar, planar, portar, posar, posicionar, precipitar-se, predicar, pregonar, preguntar, prendre, presentar-se, prestar, proclamar, projectar, proporcionar, propulsar, proveir, radiar, rampinyar, raptar, rebre, recobrar, recollir, recompensar, recordar, recórrer, regalar, regresar, relatar, remetre, remolcar, repartir, respondre, retirar, retirar-se, retornar, retransmetre, robar, rondar, saltar, sermonar, sermonejar, servir, situar, sobreposar, sortir, submergir, submergir-se, subministrar, tafanejar, telefonar, telegrafiar, televisar, tirar, tirar-se, tornar, tragar, transbordar, transferir, transmetre, transportar, traslladar, travessar, trepr, treure, ubicar, usurpar, vagabundejar, vagar, vendre, venir, vessar, viatjar, volar, xafardejar, xiuxiuejar.

B AnCorra's correspondences between Arguments, Thematic Roles and Syntactic Functions

Attribute <func> value / function	Attribute <arg> value	Attribute <tem> value / thematic role	Example	
su _j / subject	arg0	agt / agent	Juan lee una novela	
		cau / cause	El viento abrió la puerta	
		exp / experiencer	Juan sueña	
	arg1	src / source	Juan sudaba	
		pat / patient	Clara es amada por todos	
	arg2	tem / theme	Juan llegó tarde	
		ins / instrument	Una lona cubre el coche de Juan	
	argL	loc / locative	El diario abordó la noticia	
		none	El cadáver fue levantado a la 1 de la tarde	
	cd / direct object	arg1	pat / patient	Juan lee una novela
arg2		tem / theme	El viento abrió la puerta	
		atr / attribute	Juan tiene un coche blanco	
argL		ext / extension	El paro subió 15.891 personas	
creg / prepositional object	arg1	none	Puso punto final a la discusión	
		tem / theme	Clara se rió del chiste	
	arg2	loc / locative	Juan intervino ante la comisión	
		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	
		eim / 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	
		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	
		eim / 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		
	arg3	ori / origin	Juan ha regresado de las vacaciones	
		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
	et / textual element	none	none	¿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

Figure 21: AnCorra's correspondences between Arguments, Thematic Roles and Syntactic Functions.

C Ancora's structure proposals for *aconseguir*, *arribar*, *avançar*, *entrar* and *sortir*

C.1 Aconseguir

Function	Argument	Theme	Constituents
suj	arg0	agt	
cd	arg1	pat	
cpred	argM	atr	
cc	argM	adv	sp(per)
cc	argM	fin	sp(per)
cc	argM	mnr	sp(en), sp(amb)
cc	argM	tmp	sp(en)

Figure 22: Ancora's structure proposal for the active voice of the verb *aconseguir*

Function	Argument	Theme	Constituents
suj	arg1	pat	
cag	arg0	agt	sp(per)
cc	argM	tmp	

Figure 23: Ancora's structure proposal for the passive voice of the verb *aconseguir*

C.2 Arribar

Function	Argument	Theme	Constituents
subj	arg1	tem	
ci	arg2	ben	
cc	arg3	ori	sp(de)
cc	arg4	des	sp(a), sp(fins)
cpred	argM	atr	
cc	argM	tmp	sp(a)
cc	argM	mnr	sp(amb), sp(mitjançper)
cc	argM	fin	sp(per)
cc	argM	adv	

Figure 24: Ancora’s structure proposal for the active voice of sense 1 of the verb *arribar*

Function	Argument	Theme	Constituents
ci	arg2	ben	
cc	arg3	ori	sp(de)
cc	arg4	des	sp(a), sp(fins)

Figure 25: Ancora’s structure proposal for the impersonal voice of sense 1 of the verb *arribar*

Function	Argument	Theme	Constituents
subj	arg1	tem	
creg	arg2	loc	sp(a)
cc	argM	tmp	sp(a), sp(en)
cc	argM	mnr	sp(amb)

Figure 26: Ancora’s structure proposal for the active voice of sense 2 of the verb *arribar*

Function	Argument	Theme	Constituents
creg	arg2	loc	sp(a)
cc	argM	cau	sp(amb)

Figure 27: Ancora’s structure proposal for the impersonal voice of sense 2 of the verb *arribar*

Function	Argument	Theme	Constituents
subj	arg0	agt	
creg	arg1	tem	sp(a)
cc	argM	tmp	
cc	argM	adv	sp(en), sp(amb)

Figure 28: Ancora’s structure proposal for the active voice of sense 3 of the verb *arribar*

Function	Argument	Theme	Constituents
creg	arg1	tem	sp(a)
cc	argM	tmp	sp(durant), sp(en)
cc	argM	adv	sp(per)
cc	argM	loc	sp(a)

Figure 29: Ancora’s structure proposal for the impersonal voice of sense 3 of the verb *arribar*

Function	Argument	Theme	Constituents
subj	arg0	agt	
cd	arg1	pat	
ci	arg2	ben	sp(a)
cc	arg4	des	sp(a)
cc	argM	mnr	

Figure 30: Ancora's structure proposal for the active voice of sense 4 of the verb *arribar*

Function	Argument	Theme	Constituents
subj	arg1	pat	
ci	arg2	ben	sp(a)
cc	arg4	des	sp(a)

Figure 31: Ancora's structure proposal for the passive voice of sense 4 of the verb *arribar*

Function	Argument	Theme	Constituents
subj	arg1	tem	
cc	arg4	efi	sp(a)

Figure 32: Ancora's structure proposal for the active voice of sense 5 of the verb *arribar*

Function	Argument	Theme	Constituents
subj	arg0	agt	
cd	argL		

Figure 33: Ancora's structure proposal for the active voice of sense 6 of the verb *arribar*

Function	Argument	Theme	Constituents
subj	arg1	tem	

Figure 34: Ancora's structure proposal for the active voice of sense 7 of the verb *arribar*

C.3 Avançar

Function	Argument	Theme	Constituents
suj	arg1	tem	
cc	argM	mnr	
cc	argM	adv	sp(sobre)

Figure 35: Ancora's structure proposal for the active voice of sense 1 of the verb *avançar*

Function	Argument	Theme	Constituents
suj	arg0	cau	
cd	arg1	tem	

Figure 36: Ancora's structure proposal for the causative voice of sense 1 of the verb *avançar*

Function	Argument	Theme	Constituents
suj	arg0	agt	
cd	arg1	pat	
cc	arg2	loc	sp(en)
cc	arg3	ori	
cc	arg4	des	
cc	argM	ext	
cc	argM	tmp	
cc	argM	cau	

Figure 37: Ancora's structure proposal for the active voice of sense 2 of the verb *avançar*

Function	Argument	Theme	Constituents
suj	arg0	agt	
cd	arg1	pat	
ci	arg2	ben	sp(a)

Figure 38: Ancora's structure proposal for the benefactive voice of sense 2 of the verb *avançar*

Function	Argument	Theme	Constituents
cd	arg1	ext	
cc	argM	tmp	sp(en)

Figure 39: Ancora's structure proposal for the impersonal voice of sense 2 of the verb *avançar*

Function	Argument	Theme	Constituents
suj	arg1	pat	
cc	argM	tmp	
cc	argM	loc	sp(a)

Figure 40: Ancora's structure proposal for the passive voice of sense 2 of the verb *avançar*

Function	Argument	Theme	Constituents
subj	arg0	agt	
cd	arg1	pat	
ci	arg2	ben	sp(a)
cc	argM	tmp	

Figure 41: Ancora's structure proposal for the active voice of sense 3 of the verb *avançar*

Function	Argument	Theme	Constituents
subj	arg0	agt	
cc	argM	mnr	

Figure 42: Ancora's structure proposal for the active voice of sense 4 of the verb *avançar*

Function	Argument	Theme	Constituents
subj	arg0	cau	
cd	arg1	tem	

Figure 43: Ancora's structure proposal for the active voice of sense 5 of the verb *avançar*

Function	Argument	Theme	Constituents
subj	arg0	cau	
cd	arg1	tem	

Figure 44: Ancora's structure proposal for the active voice of sense 6 of the verb *avançar*

C.4 Entrar

Function	Argument	Theme	Constituents
subj	arg1	tem	
cc	arg2	a	
cc	arg2	loc	sp(en)
cc	argM	loc	sp(en)
cc	argM	adv	sp(con), sp(por)
cc	argM	fin	sp(a)
cc	argM	tmp	
cc	argM	mnr	sp(a)

Figure 45: Ancora's structure proposal for the active voice of sense 1 of the verb *entrar*

Function	Argument	Theme	Constituents
subj	arg1	tem	
cc	arg2	efi	sp(en)

Figure 47: Ancora's structure proposal for the active voice of sense 2 of the verb *entrar*

Function	Argument	Theme	Constituents
ccreg	arg2	loc	sp(en)
cc	argM	adv	sp(amb)

Figure 49: Ancora's structure proposal for the impersonal voice of sense 3 of the verb *entrar*

Function	Argument	Theme	Constituents
cc	argM	loc	sp(a)

Figure 46: Ancora's structure proposal for the impersonal voice of sense 1 of the verb *entrar*

Function	Argument	Theme	Constituents
subj	arg1	tem	
ccreg	arg2	loc	sp(en)
cc	argM	mnr	sp(en)
cc	argM	adv	sp(amb)
cc	argM	fin	

Figure 48: Ancora's structure proposal for the active voice of sense 3 of the verb *entrar*

Function	Argument	Theme	Constituents
subj	arg0	agt	
cd	arg1	pat	
cc	arg2	loc	sp(a)
cc	argM	tmp	
cc	argM	fin	

Figure 50: Ancora's structure proposal for the active voice of sense 4 of the verb *entrar*

Function	Argument	Theme	Constituents
suj	arg1	pat	
cc	arg2	loc	sp(a)

Figure 51: Ancora's structure proposal for the passive voice of sense 4 of the verb *entrar*

Function	Argument	Theme	Constituents
suj	arg1	tem	
creg	argL		sp(en)

Figure 52: Ancora's structure proposal for the active voice of sense 5 of the verb *entrar*

Function	Argument	Theme	Constituents
suj	arg1	tem	
cc	arg2	loc	sp(a)
cc	arg3	ori	sp(de)
cc	arg4	des	sp(en)
cpred	argM	atr	
cc	argM	adv	
cc	argM	tmp	sp(a), sp(entre)
cc	argM	cau	sp(en)
cc	argM	fin	sp(a)

Figure 53: Ancora's structure proposal for the active voice of sense 6 of the verb *entrar*

Function	Argument	Theme	Constituents
suj	arg0	cau	
cd	arg1	tem	

Figure 54: Ancora's structure proposal for the active voice of sense 7 of the verb *entrar*

C.5 Sortir

Function	Argument	Theme	Constituents
suj	arg1	tem	
cc	arg2	loc	sp(a), sp(per)
cc	arg3	ori	sp(de)
cc	arg4	des	sp(en)
cpred	argM	atr	
cc	argM	adv	
cc	argM	tmp	sp(a), sp(entre)
cc	argM	cau	sp(en)
cc	argM	fin	sp(a)

Figure 55: Ancora’s structure proposal for the active voice of sense 1 of the verb *sortir*

Function	Argument	Theme	Constituents
suj	arg0	agt	
creg	argL		sp(en)

Figure 57: Ancora’s structure proposal for the active voice of sense 2 of the verb *sortir*

Function	Argument	Theme	Constituents
suj	arg1	tem	
ci	arg2	ben	sp(a)
cpred	argM	atr	

Figure 59: Ancora’s structure proposal for the active voice of sense 4 of the verb *sortir*

Function	Argument	Theme	Constituents
cc	arg3	ori	sp(de)
cc	argM	adv	

Figure 56: Ancora’s structure proposal for the impersonal voice of sense 1 of the verb *sortir*

Function	Argument	Theme	Constituents
suj	arg1	tem	
cc	argL		

Figure 58: Ancora’s structure proposal for the active voice of sense 3 of the verb *sortir*

Function	Argument	Theme	Constituents
suj	arg1	tem	
creg	arg2	loc	sp(de)

Figure 60: Ancora’s structure proposal for the active voice of sense 5 of the verb *sortir*

D Tables of results

D.1 Tables of results of the verb *aconseguir*

Arguments structure		Syntactic structure		Thematic structure	
arg0 + arg1 + argM	110	subj + cd + cc	109	pat	77
arg1 + argM	81	cd + cc	79	agt + pat	45
arg1	77	cd	75	agt + pat + tmp	32
arg0 + arg1	44	subj + cd	43	pat + tmp	29
		subj	2	agt + pat + mnr	17
		subj + cc	2	agt + pat + tmp + mnr	16
		subj + cag	1	pat + mnr	15
		cd + ci	1	agt + pat + loc	14
				agt + pat + tmp + loc	10
				pat + loc	9
				pat + tmp + mnr	8
				pat + fin	5
				pat + tmp + loc	4
				pat + co-agt	4
				agt + pat + fin	3
				agt + pat + loc + mnr	3
				agt + pat + fin + loc	3
				pat + fin + mnr	3
				pat + fin + tmp	2
				agt + pat + co-agt	2
				agt + pat + tmp + co-agt	2
				pat + adv	1
				agt + pat + adv + tmp	1
				agt + pat + tmp + loc + mnr	1
				agt + pat + fin + loc + tmp + co-agt	1
				agt + pat + tmp + loc + fin	1
				pat + fin + mnr + tmp	1
				agt + pat + tmp + fin	1
				agt + pat + tmp + mnr + co-agt	1
				agt + pat + tmp + loc + co-agt	1

Table 15: Quantitative results of the verb *aconseguir*.

D.2 Tables of results of the verb *arribar*

Arguments structure		Syntactic structure		Thematic structure	
arg1 + argM	121	subj + cc	196	tem + tmp	71
arg1 + arg4 + argM	60	cc	105	des	55
arg4	55	subj	38	tem	38
arg 4 + argM	44	subj + cpred + cc	10	des + mnr	31
arg1	38	subj + ci + cc	6	tem + des + mnr	25
arg1 + arg4	21	cpred + cc	3	tem + des + tmp	23
argM	6	subj + ci	3	tem + mnr + tmp	21
arg1 + arg2 + argM	4	creg + cc	2	tem + des	21
arg1 + arg3	4	cpred	1	tem + mnr	20
arg1 + arg2	3	subj + cpred	1	des + tmp	8
arg3	2			tem + des + atr	4
arg1 + arg3 + argM	2			tem + ori	4
arg1 + arg2 + arg4	1			des + tmp + mnr	3
arg3 + argM	1			tem + des + tmp + mnr	3
arg1 + arg2 + arg3 + argM	1			tem + des + fin	3
arg1 + arg3 + arg4 + argM	1			tem + ben	3
arg1 + arg3 + arg4	1			tem + atr + tmp	2
				des + atr	2
				tem + ben + tmp	2
				ori	2
				tem + atr + tmp + fin	1
				tem + adv	1
				tem + tmp + fin	1
				tem + atr	1
				tem + tmp + loc	1
				tem + atr + fin	1
				tem + atr + adv	1
				tem + des + mnr + fin	1
				tem + des + atr + tmp	1
				tem + ben + fin	1
				tem + ben + tmp + mnr	1
				ori + adv	1
				tem + ori + mnr	1
				tem + ori + tmp	1
				tem + ori + des + atr	1
				tem + ori + des	1
				tem + ben + des	1
				tem + ben + ori + tmp	1
				tem + ori + des + atr	1
				tem + ori + des	1

Table 16: Quantitative results of the verb sense 1 of *arribar*.

Arguments structure		Syntactic structure		Thematic structure	
arg0 + arg1 + argM	8	suj + creg + cc	8	agt + tem + adv	6
arg1 + argM	7	creg + cc	7	tem + mnr	3
arg1	2	creg	2	tem + adv	3
arg0 + arg1	1	suj + creg	1	tem	2
				tem + tmp	1
				agt + tem + tmp	1
				agt + tem	1
				agt + tem + adv + tmp	1

Table 17: Quantitative results of the verb sense 2 of *arribar*.

Arguments structure		Syntactic structure		Thematic structure	
arg1 + arg2	3	cd + ci	3	agt + pat + ben	2
arg0 + arg1 + arg2	2	suj + cd + ci	2	pat + ben	3

Table 18: Quantitative results of the verb sense 3 of *arribar*.

D.3 Tables of results of the verb *avançar*

Arguments structure		Syntactic structure		Thematic structure	
arg1 + argM	5	suj + cc	5	tem + adv	2
argM	4	cc	4	tem	3
arg1	3	suj	3	tem + mnr	2
				tmp	1
				mnr	3
				tem + tmp	1

Table 19: Quantitative results of the verb sense 1 of *avançar*.

Arguments structure		Syntactic structure		Thematic structure	
arg0 + arg1 + argM	37	suj + cd + cc	50	agt + pat + tmp	14
arg0 + arg1 + arg2 + argM	10	cd + cc	21	agt + pat + mnr + tmp	11
arg1 + argM	10	suj + cd	10	agt + pat	10
arg0 + arg1	10	cd	9	agt + pat + mnr	9
arg1	9	cc	4	pat	9
arg1 + arg2	7	suj + cc	2	pat + loc	7
arg0 + arg1 + arg2	3	suj	1	agt + pat + loc + tmp	6
arg1 + arg2 + argM	3			pat + tmp	4
arg2	2			pat + mnr	3
argM	2			agt + pat + loc + mnr + tmp	3
arg0 + argM	1			agt + pat + loc	3
arg1 + arg3	1			loc	2
arg0	1			agt + pat + loc + mnr	2
arg1 + arg4	1			pat + loc + mnr	2
				agt	1
				pat + loc + cau	1
				pat + cau	1
				agt + pat + tmp + cau	1
				agt + des	1
				pat + tmp + cau	1
				pat + ori	1
				pat + mnr + tmp + fin	1
				mnr + tmp	1
				agt + tmp	1
				mnr	1
				agt + pat + tmp + fin	1

Table 20: Quantitative results of the verb sense 2 of *avançar*.

D.4 Tables of results of the verb *entrar*

Arguments structure		Syntactic structure		Thematic structure	
arg2	30	cc	50	loc	30
arg1 + argM	21	subj + cc	42	tem + loc	17
arg1 + arg2	17	subj	14	tem	14
arg2 + argM	17	subj + cpred + cc	11	tem + tmp	11
arg1 + arg2 + argM	16	cpred + cc	4	tem + atr + mnr	8
arg1	14	subj + cpred	1	loc + tmp	8
argM	7	creg	1	tem + loc + tmp	6
arg3	1			loc + mnr	5
				tem + loc + fin	3
				tmp	3
				atr + loc	2
				atr + mnr	2
				tem + atr + loc	2
				tem + loc + mnr	2
				loc + fin	2
				tem + loc + adv	2
				tem + atr	1
				tmp + mnr	1
				ori	1
				tem + adv + tmp	1
				loc + adv	1
				tem + loc + tmp + co-agt	1

Table 21: Quantitative results of the verb sense 1 of *entrar*.

Arguments structure		Syntactic structure		Thematic structure	
arg1 + argM + argL	1	subj + cpred + creg + cc	1	tem + atr + tmp	1

Table 22: Quantitative results of the verb sense 2 of *entrar*.

D.5 Table of results of the verb *sortir*

Arguments structure		Syntactic structure		Thematic structure	
arg1 + argM	34	cc	56	ori	27
arg3	27	su _j + cc	51	tem + mnr	10
arg1 + arg2 + argM	18	su _j + cpred + cc	12	tem	9
argM	14	su _j	9	tem + ori	8
arg3 + argM	11	cpred + cc	6	loc	6
arg1	10	su _j + ci + cpred	5	tem + mnr + tmp	5
arg1 + arg3	8	cpred	4	mnr	5
arg1 + arg3 + argM	7	su _j + cpred	3	tem + loc + tmp	5
arg2 + argM	6	su _j + ci + cc	2	tem + atr	4
arg2	6			tem + ben + atr	4
arg1 + arg2	3			atr	4
arg1 + arg1 + argM	3			ori + tmp	4
arg2 + arg1	1			tem + atr + tmp	3
				tem + tmp	3
				tmp	3
				tem + loc	3
				tem + loc + mnr	2
				tem + ori + tmp	2
				ori + atr	2
				ori + mnr	2
				tem + ben + tmp + fin	2
				loc + mnr	2
				tem + adv	2
				loc + fin	2
				tem + ori + fin	2
				tem + mnr + fin	2
				tem + ori + atr	2
				tem + tmp + fin	1
				tem + loc + mnr + tmp	1
				tem + atr + mnr	1
				atr + loc	1
				loc + des	1
				tem + loc + atr + mnr	1
				tem + atr + adv	1
				tem + des + atr + tmp	1
				ori + atr + tmp	1
				des	1
				tem + des + tmp	1
				fin	1
				ori + mnr + adv	1
				tem + fin	1
				tem + des + mnr	1
				agt + tem + atr	1
				loc + atr + fin	1
				tem + ori + atr + mnr	1
				tem + ben + atr + mnr	1
				atr + tmp	1
				ori + atr + mnr	1
				tem + loc + atr	1
				tem + atr + mnr + tmp	1

Table 23: Quantitative results of the verb sense 1 of *sortir*.