Parsing and Evaluation.
Improving Dependency Grammars Accuracy

Anàlisi Sintàctica Automàtica i Avaluació. Millora de qualitat per a Gramàtiques de Dependències

Marina Lloberes Salvatella

Parsing and Evaluation.
Improving Dependency Grammars Accuracy

Anàlisi Sintàctica Automàtica i Avaluació. Millora de qualitat per a Gramàtiques de Dependències

A Dissertation Submitted in Partial Fulfilment of the Requirements
of the Degree of Doctor of Philosophy with the International Mention

Marina Lloberes Salvatella

Supervisors:
Irene Castellón Masalles
Lluís Padró Cirera

Ciència Cognitiva i Llenguatge
Departament de Filologia Catalana i Lingüística General
Universitat de Barcelona
May 2016

UNIVERSITAT DE BARCELONA

Abstract

Because parsers are still limited in analysing specific ambiguous constructions, the research presented in this thesis mainly aims to contribute to the improvement of parsing performance when it has knowledge integrated in order to deal with ambiguous linguistic phenomena. More precisely, this thesis intends to provide empirical solutions to the disambiguation of prepositional phrase attachment and argument recognition in order to assist parsers in generating a more accurate syntactic analysis. The disambiguation of these two highly ambiguous linguistic phenomena by the integration of knowledge about the language necessarily relies on linguistic and statistical strategies for knowledge acquisition.

The starting point of this research proposal is the development of a rule-based grammar for Spanish and for Catalan following the theoretical basis of Dependency Grammar (Tesnière, 1959; Mel'čuk, 1988) in order to carry out two experiments about the integration of automatically-acquired knowledge. In order to build two robust grammars that understand a sentence, the FreeLing pipeline (Padró et al., 2010) has been used as a framework. On the other hand, an eclectic repertoire of criteria about the nature of syntactic heads is proposed by reviewing the postulates of Generative Grammar (Chomsky, 1981; Bonet and Solà, 1986; Haegeman, 1991) and Dependency Grammar (Tesnière, 1959; Mel'čuk, 1988). Furthermore, a set of dependency relations is provided and mapped to Universal Dependencies (Mcdonald et al., 2013).

Furthermore, an empirical evaluation method has been designed in order to carry out both a quantitative and a qualitative analysis. In particular, the dependency parsed trees generated by the grammars are compared to real linguistic data. The quantitative evaluation is based on the Spanish Tibidabo Treebank (Marimon et al., 2014), which is large enough to carry out a real analysis of the grammars performance and which has been annotated with the same formalism as the grammars, syntactic dependencies. Since the criteria between both resources are different, a process of harmonization has been applied developing a set of rules that automatically adapt the criteria of the corpus to the grammar criteria. With regard to qualitative evaluation, there are no available resources to evaluate Spanish and Catalan dependency grammars qualitatively. For this reason, a test suite of syntactic phenomena about structure and word order has been built. In order to create a representative repertoire of the languages observed, descriptive grammars (Bosque and Demonte, 1999; Solà et al., 2002) and the SenSem Corpus (Vázquez and

Fernández-Montraveta, 2015) have been used for capturing relevant structures and word order patterns, respectively.

Thanks to these two tools, two experiments have been carried out in order to prove that knowledge integration improves the parsing accuracy. On the one hand, the automatic learning of language models has been explored by means of statistical methods in order to disambiguate PP-attachment. More precisely, a model has been learned with a supervised classifier using Weka (Witten and Frank, 2005). Furthermore, an unsupervised model based on word embeddings has been applied (Mikolov et al., 2013a,b). The results of the experiment show that the supervised method is limited in predicting solutions for unseen data, which is resolved by the unsupervised method since provides a solution for any case. However, the unsupervised method is limited if it only learns from lexical data. For this reason, training data needs to be enriched with the lexical value of the preposition, as well as semantic and syntactic features. In addition, the number of patterns used to learn language models has to be extended in order to have an impact on the grammars.

On the other hand, another experiment is carried out in order to improve the argument recognition in the grammars by the acquisition of linguistic knowledge. In this experiment, knowledge is acquired automatically from the extraction of verb subcategorization frames from the SenSem Corpus (Vázquez and Fernández-Montraveta, 2015) which contains the verb predicate and its arguments annotated syntactically. As a result of the information extracted, subcategorization frames have been classified into subcategorization classes regarding the patterns observed in the corpus. The results of the subcategorization classes integration in the grammars prove that this information increases the accuracy of the argument recognition in the grammars.

The results of the research of this thesis show that grammars' rules on their own are not expressive enough to resolve complex ambiguities. However, the integration of knowledge about these ambiguities in the grammars may be decisive in the disambiguation. On the one hand, statistical knowledge about PP-attachment can improve the grammars accuracy, but syntactic and semantic information, and new patterns of PP-attachment need to be included in the language models in order to contribute to disambiguate this phenomenon. On the other hand, linguistic knowledge about verb subcategorization acquired from annotated linguistic resources show a positive influence positively on grammars' accuracy.

Resum

Aquesta tesi vol tractar les limitacions amb què es troben els analitzadors sintàctics automàtics actualment. Tot i els progressos que s'han fet en l'àrea del Processament del Llenguatge Natural en els darrers anys, les tecnologies del llenguatge i, en particular, els analitzadors sintàctics automàtics no han pogut traspassar el llindar de certes ambiguïtats estructurals com ara l'agrupació del sintagma preposicional i el reconeixement d'arguments. És per aquest motiu que la recerca duta a terme en aquesta tesi té com a objectiu aportar millores significatives de qualitat a l'anàlisi sintàctica automàtica per mitjà de la integració de coneixement lingüístic i estadístic per desambiguar construccions sintàctiques ambigües.

El punt de partida de la recerca ha estat el desenvolupament de d'una gramàtica en espanyol i una altra en català basades en regles que segueixen els postulats de la Gramàtica de Dependèndencies (Tesnière, 1959; Mel'čuk, 1988) per tal de dur a terme els experiments sobre l'adquisició de coneixement automàtic. Per tal de crear dues gramàtiques robustes que analitzin i entenguin l'oració en profunditat, ens hem basat en l'arquitectura de FreeLing (Padró et al., 2010), una llibreria de Processament de Llenguatge Natural que proveeix una anàlisi lingüística automàtica de l'oració. Per una altra banda, s'ha elaborat una proposta eclèctica de criteris lingüístics per determinar la formació dels sintagmes i les clàusules a la gramàtica per mitjà de la revisió de les propostes teòriques de la Gramàtica Generativa (Chomsky, 1981; Bonet and Solà, 1986; Haegeman, 1991) i de la Gramàtica de Dependències (Tesnière, 1959; Mel'čuk, 1988). Aquesta proposta s'acompanya d'un llistat de les etiquetes de relació de dependència que fan servir les regles de les gramàtques. A més a més de l'elaboració d'aquest llistat, s'han establert les correspondències amb l'estàndard d'anotació de les Dependències Universals (Mcdonald et al., 2013).

Alhora, s'ha dissenyat un sistema d'avaluació empíric que té en compte l'anàlisi quantitativa i qualitativa per tal de fer una valoració completa dels resultats dels experiments. Precisament, es tracta una tasca empírica pel fet que es comparen les anàlisis generades per les gramàtiques amb dades reals de la llengua. Per tal de dur a terme l'avaluació des d'una perspectiva quantitativa, s'ha fet servir el corpus Tibidabo en espanyol (Marimon et al., 2014) disponible només en espanyol que és prou extens per construir una anàlisi real de les gramàtiques i que ha estat anotat amb el mateix formalisme que les gramàtiques. En concret, per tal com els criteris de les gramàtiques i del corpus no són coincidents, s'ha dut a terme un procés d'harmonització de cri-

teris per mitjà d'unes regles creades manualment que adapten automàticament l'estructura i la relació de dependència del corpus al criteri de les gramàtiques. Pel que fa a l'avaluació qualitativa, pel fet que no hi ha recursos disponibles en espanyol i català, hem dissenyat un reprertori de test de fenòmens sintàctics estructurals i relacionats amb l'ordre de l'oració. Amb l'objectiu de crear un repertori representatiu de les llengües estudiades, s'han fet servir gramàtiques descriptives per fornir el repertori d'estructures sintàctiques (Bosque and Demonte, 1999; Solà et al., 2002) i el Corpus SenSem (Vázquez and Fernández-Montraveta, 2015) per capturar automàticament l'ordre oracional.

Gràcies a aquestes dues eines, s'han pogut dur a terme dos experiments per provar que la integració de coneixement en l'anàlisi sintàctica automàtica en millora la qualitat. D'una banda, s'ha explorat l'aprenentatge de models de llenguatge per mitjà de models estadístics per tal de proposar solucions a l'agrupació del sintagma preposicional. Més concretament, s'ha desenvolupat un model de llenguatge per mitjà d'un classificador d'aprenentatge supervisat de Weka (Witten and Frank, 2005). A més a més, s'ha après un model de llenguatge per mitjà d'un mètode no supervisat basat en l'aproximació distribucional anomenat *word embeddings* (Mikolov et al., 2013a,b). Els resultats de l'experiment posen de manifest que el mètode supervisat té greus limitacions per fer donar una resposta en dades que no ha vist prèviament, cosa que és superada pel mètode no supervisat pel fet que és capaç de classificar qualsevol cas. De tota manera, el mètode no supervisat que s'ha estudiat és limitat si aprèn a partir de dades lèxiques. Per aquesta raó, és necessari que les dades utilitzades per entrenar el model continguin el valor de la preposició, trets sintàctics i semàntics. A més a més, cal ampliar el número de patrons apresos per tal d'ampliar la cobertura dels models i tenir un impacte en els resultats de les gramàtiques.

D'una altra banda, s'ha proposat una manera de millorar el reconeixement d'arguments a les gramàtiques per mitjà de l'adquisició de coneixement lingüístic. En aquest experiment, s'ha optat per extreure automàticament el coneixement en forma de classes de subcategorització verbal d'el Corpus SenSem (Vázquez and Fernández-Montraveta, 2015), que conté anotats sintàcticament el predicat verbal i els seus arguments. A partir de la informació extreta, s'ha classificat les diverses diàtesis verbals en classes de subcategorització verbal en funció dels patrons observats en el corpus. Els resultats de la integració de les classes de subcategorització a les gramàtiques mostren que aquesta informació determina positivament el reconeixement dels arguments.

Els resultats de la recerca duta a terme en aquesta tesi doctoral posen de manifest que les regles de les gramàtiques no són prou expressives per elles mateixes per resoldre ambigüitats complexes del llenguatge. No obstant això, la integració de coneixement sobre aquestes ambigüitats pot ser decisiu a l'hora de proposar una solució. D'una banda, el coneixement estadístic sobre l'agrupació del sintagma preposicional pot millorar la qualitat de les gramàtiques, però per afirmar-ho cal incloure informació sintàctica i semàntica en els models d'aprenentatge automàtic i capturar més patrons per contribuir en la desambiguació de fenòmens complexos. D'una altra banda, el coneixement lingüístic sobre subcategorització verbal adquirit de recursos lingüístics anotats influeix decisivament en la qualitat de les gramàtiques per a l'anàlisi sintàctica automàtica.

CONTENTS

LIST OF FIGURES

List of Tables

Chapter 1

Introduction

Automatic syntactic analysis of written texts commonly known as Parsing has been one of the main interests in Natural Language Processing (NLP) since the early prototypes of language technologies appeared in the mid 20th century. This interest originates from the necessity of the language technologies to jump from shallow linguistic analysis usually obtained from textual strings to a deeper and more robust analysis provided by layers of linguistic information representing more abstract concepts of language.

In particular, automatic syntactic analysis provides a pre-processing step for language applications that work with deep linguistic information such as programs that require certain kind of semantic representation. Consequently, a large number of language technologies directly benefit from the improvements at the level of automatic syntactic analysis. For example, parsing is present in other areas of NLP such as Semantic Role Labelling (Gildea and Jurafsky, 2002; Surdeanu and Turmo, 2005; Màrquez et al., 2008), Machine Translation (Hutchins and Somers, 1992; Koehn, 2010), Information Extraction (Cowie and Lehnert, 1996; Freitag, 2000), Information Retrieval (Baeza-Yates and Ribeiro-Neto, 1999; Manning et al., 2008; Büttcher et al., 2010) and Sentiment Analysis (Turney, 2002; Titov and McDonald, 2008)

As a result of this, the task of automatic syntactic analysis has been linked to the advances of language technology. Parsers have been progressively integrating frameworks of language analysis (e.g. Phrase Structure Grammar, Head-Driven Phrase Structure Grammar and other Unification Grammars, Dependency Grammar, etc.), language processing approaches (e.g. rule-based, data-driven and unsupervised parsing methods) and strategies for knowledge representation (e.g. automatic acquisition, automatic supervised learning, unsupervised learning by distributional word representations, word embeddings, neural networks, etc.).

Parsing has been one of the basic areas of NLP since the earliest parsers were developed (Hays, 1964; Gaifman, 1965). However, the great success of parsing began in the late 1990s and extends to nowadays. During this period, the advances in parsing and, specifically, dependency parsing increased significantly because of several conferences, competitions and discussion groups, like the SIGNLL Conference on Computational Natural Language Learning from 2006 to 2009 which focused on dependency parsing (Buchholz and Marsi, 2006; Nivre et al., 2007; Surdeanu et al., 2008; Hajič et al., 2009), the International Conference on Parsing Technologies, the International Conference on Dependency Linguistics, as well as the SIGPARSE (the ACL Special Interest Group on Natural Language Parsing).

As a consequence of this intense activity, nowadays the state of the art of parsers is satisfactory and the list of dependency parsers that are available is large. Among these dependency parsers, some can be distinguished due to their contribution to the field, such as the statistical parser based on support vector machines of Yamada and Matsumoto (2003), the MSTParser (McDonald et al., 2005), the Standford Parser (De Marneffe et al., 2006), the MaltParser (Nivre et al., 2006), the tree adjoining grammar parser of Carreras et al. (2008), the transition-based non-projective parser of Bohnet and Nivre (2012), and the transition-based parser with spinal trees of Ballesteros and Carreras (2015).

1.1 Research Motivation

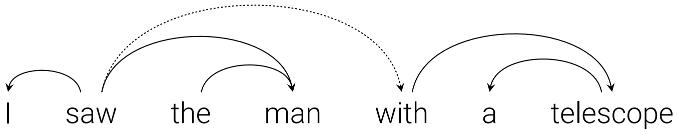
Despite the successful performance of parsers, they are still limited in resolving particular natural language phenomena because they have not yet found a satisfactory methodology to handle the structural ambiguities inherent in natural language. In particular, prepositional phrase attachment or PP-attachment (1-a) and argument recognition (1-b) are two problematic language constructions which have been widely studied in NLP, but they still need to be resolved satisfactorily. In addition, coordination has been noticed as a highly ambiguous construction (1-c), but it has not been as widely handled in NLP as the rest of ambiguous phenomena.

(1) a. I saw the man *with a telescope*
b. He visto *a mis amigos*
'$\emptyset_{1sg}$ saw my friends'
c. Els alumnes parlen *de la vaga i de pressa* convencen els sindicats
'The students talk about the strike and quickly convince the unions'

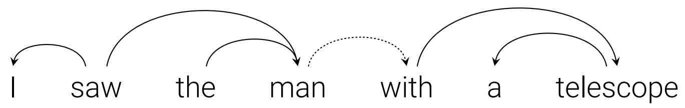
The sentence in (1-a) contains a structural ambiguity in the PP-attachment as it has two interpretations which are represented by a different syntactic structure and which parsers cannot distinguish often (2). While the sentence can express that the subject ('I') used a telescope in order to perform the action of the sentence (2-a), it also can refer to the fact that the object of the sentence ('a man') was using a telescope (2-b).

(2) I saw the man *with a telescope*

a. I saw with a telescope the man



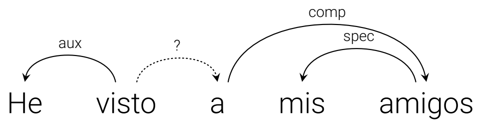
b. I saw the man who was using a telescope



This problem has been addressed from the point of view of statistical knowledge learning and several supervised (Hindle and Rooth, 1993; Ratnaparkhi et al., 1994; Collins and Brooks, 1995; Stetina and Nagao, 1997; Olteanu and Moldovan, 2005; Merlo and Ferrer, 2006; Agirre et al., 2008), unsupervised (Ratnaparkhi, 1998; Pantel and Lin, 2000; Šuster, 2012; Belinkov et al., 2014) and semi-supervised (Gala and Lafourcade, 2006) methods. Despite the good results of these studies, there are few proposals oriented to the disambiguation of the PP-attachment in parsing and, consequently, accuracy is still low.

In the case of the ambiguity in (1-b), there is one interpretation possible of the prepositional phrase as a direct object. However, a parser cannot easily distinguish a direct object, an indirect object, a prepositional object and an adjunct when they are expressed by a prepositional phrase (3). This limitation is observed by the studies of Carroll et al. (1998); Zeman (2002) which point to the necessity of adding in a parser subcategorization knowledge that has automatically been acquired or learned.

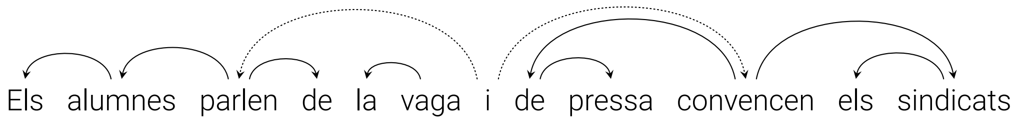
(3) He visto a mis amigos
‘$\emptyset_{1sg}$ saw my friends’



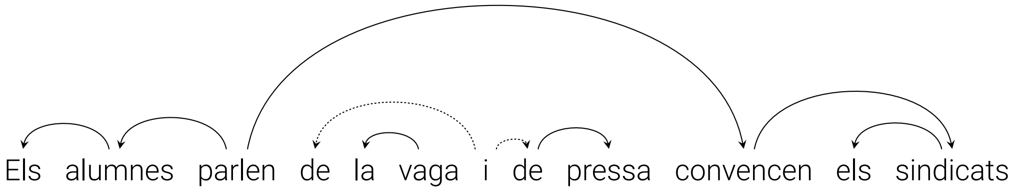
Concerning the coordinating construction such as (4), the sentence can only be interpreted in one way as well. However, since coordination can occur at any sentence level and coordinated elements can be very diverse, the probability of finding ambiguities is greater. In (4), the only correct answer is the structure of coordinated sentences (4-a). Despite this, a parser can choose as the correct answer the analysis of (4-b) in which the two prepositional phrases are coordinated because it is not able to discriminate the right structure.

(4) Els alumnes parlen *de la vaga i de pressa* convencen els sindicats
‘The students talk about the strike and quickly convince the unions’

a. Els alumnes parlen i convencen



b. Els alumnes parlen de la vaga i de pressa


Els alumnes parlen de la vaga i de pressa convencen els sindicats

In addition to this diverse scenario, another discussion surrounding the parsers' performance is the evaluation of their accuracy. The NLP community is aware of the necessity of parsing evaluation tasks in order to measure the quality of the tools. For this reason, an empirical assessment of a parser is carried out during the development of the tool (Lin, 1998b; Calvo and Gelbukh, 2006; Bick, 2006; Buchholz and Marsi, 2006; Nivre et al., 2007; Hajič et al., 2009; Ballesteros and Carreras, 2015).

Despite the effort required for evaluating parsers, in general, results usually only refer to a quantitative analysis and exclude a qualitative explanation of the errors. For example, in the three Conference on Computational Natural Language Learning (CoNLL) contests Buchholz and Marsi (2006); Nivre et al. (2007); Hajič et al. (2009) as well as the evaluation of specific parsers such as Lin (1998b); Calvo and Gelbukh (2006); Bick (2006) and Ballesteros and Carreras (2015), only quantitative results are provided, and error analysis is skipped. Then, these kinds of evaluation tasks are not exhaustive since they do not answer all the questions.

Therefore, in order to determine the quality of a parser, the evaluation task needs to be global, i.e., to explain the errors from a quantitative and qualitative point of view (McEnery and Wilson, 1996). The quantitative analysis relies on statistically informative data and provides an approximate explanation about the real spectrum. On the other hand, quantitative analysis offers an in-depth description rather than a quantification of the data and provides an exhaustive description of the data. Both perspectives complement each other and make it possible to determine the real performance of a parser. As well as to establish the priorities and the strategies in the following iterations of development in order to successfully parse written texts.

On the other hand, despite the extensive development in parsing, this area tends to focus on particular languages such English. Consequently, other languages such as Spanish are represented to a lesser extent or are even barely represented such as Catalan. From the point of view of statistical parsing, there are several parsers available in both languages because these languages were present in the CoNLL contest about multilingual data-driven parsing, Spanish in 2006 (Buchholz and Marsi, 2006), Catalan in 2007 (Nivre et al., 2007) and both languages in 2009 (Hajič et al., 2009). However, the number of contributions falls with regard to rule-based parsing in Spanish and Catalan. While some authors developed Spanish rule-based parsers (Tapanainen and Järvinen, 1997; Ferrández and Moreno, 2000; Bick, 2006; Calvo and Gelbukh, 2006; Marimón, 2010; Gamallo, 2015), only a rule-based parser is available in Catalan (Alsina et al., 2002).

1.2 Aim of this Thesis

Because parsers are still limited in analysing specific ambiguous constructions, the research presented in this thesis mainly aims to contribute to **the improvement of parsing performance when it has knowledge integrated about the language of highly ambiguous linguistic phenomena**. More precisely, this thesis intends to provide empirical solutions to the disambiguation of prepositional phrase attachment and argument recognition in order to assist parsers in generating a more accurate syntactic analysis.

The disambiguation of these two highly ambiguous linguistic phenomena by the integration of knowledge about the language necessarily relies on the strategies for knowledge acquisition. In particular, types of knowledge can be described and they are closely associated with acquisition strategies. There are two kinds of knowledge according to this: linguistic knowledge and statistical knowledge. Linguistic knowledge refers to the information obtained from language resources already linguistically processed which contain the answer to the problem aimed to be solved. Statistical knowledge corresponds to the information contained in a language model learned by applying machine learning techniques, which basically can be divided into supervised and unsupervised learning.

In this thesis, both kinds of knowledge, linguistic and statistical, are explored in order to determine how they contribute to the improvement in parsing accuracy. Linguistic knowledge about verb subcategorization is addressed to improve the accuracy of argument recognition in parsing (Lloberes et al., 2015a). Statistical knowledge acquired from applying supervised learning and word embeddings techniques is implemented in order to disambiguate PP-attachment in parsing.

On the other hand, since parsing evaluation does not offer a global interpretation of the results, the research of this thesis also aims to **design a global evaluation task method for parsing that takes care of the analysis of errors quantitatively and qualitatively**. A method for global parsing evaluation is required in order to provide a complete empirical analysis of the contribution of parsing of the linguistic and statistical knowledge learned.

For this reason, a quantitative and a qualitative analysis of errors which calibrates the contribution of knowledge integration in parsing is necessary in order to measure how the performance of parsing is influenced by the integration of knowledge. While the quantitative analysis will measure the amount of correct answers of the integration of knowledge, the qualitative analysis will provide rich information about the details and the reasons for the integration errors (Lloberes et al., 2014, 2015b).

Both goals of this thesis are tested in a rule-based dependency grammar for Spanish and one for Catalan specifically developed for this proposal. The grammars follow the dependency formalism (Tesnière, 1959; Mel'čuk, 1988) and the grammar rules provide a robust parse tree of a sentence in which the complete syntactic structure is built and every link of the structure is labelled with a dependency relation (Lloberes et al., 2010). In particular, Spanish and Catalan languages are chosen for the development of the grammars because the repertoire of rule-based parsers in these languages is limited or almost non-existent.

The development of the grammars needs to be guided by linguistic criteria which determine the most appropriate syntactic representation of linguistic phenomena. In this research, controversial constructions in both languages are reviewed in order to provide a repertoire of empirically and linguistically motivated criteria concerning syntactic structure. In addition, a list of dependency relations labels needs to be created in order to define the relations established between nodes of a dependency tree.

1.3 Main Hypothesis

The main goal of this thesis leads us to formulate a set of specific questions which conduct the research and can be defined as follows:

1. Is syntactic information of dependency grammars rules expressive enough in order to provide an appropriate dependency parse tree?
2. Does statistical knowledge about prepositional phrase attachment improve the performance of the dependency grammars?
3. Is linguistic knowledge about verbal subcategorization informative enough for solving recognition of arguments by the dependency grammars?

These research questions can be formulated in a set of statements which are expressed in the following hypothesis and which will be answered at the end of this thesis as a result of the research.

Hypothesis 1
Syntactic grammar rules provide an acceptable solution for the majority of constructions except for ambiguous syntactic phenomena. Grammar rules can provide an acceptable syntactic analysis of a sentence on their own. However, rules are limited in making decisions on ambiguous constructions in which more than one structure is possible to be generated by the grammar.

Hypothesis 2
Statistical knowledge integrated in the grammar improves the accuracy of the grammar's performance. The integration of information about PP-attachment disambiguation learned from language models makes a difference to the syntactic analysis generated by the grammars. In particular, the implementation of automatic learning techniques in PP-attachment disambiguation leads to the discussion of the following statements:

> Hypothesis 2.1
> **Unsupervised learning makes it possible to more consistently capture unpredicted data**, while supervised learning techniques are limited in that regard.

Hypothesis 2.2

Language models learned by simple information such as lexical information provide a language representation of the PP-attachment which is not precise enough to disambiguate it. Therefore, enriched vectors with more complex information such as syntactic and semantic information ensure an improvement in the disambiguation task.

Hypothesis 3

Linguistic knowledge added in a rule-based grammar contributes to an improvement of the grammar's performance. The addition of syntactic-semantic information by means of verbal subcategorization frames extracted from linguistic annotated resources ensures a significant improvement of the analysis generated by the grammar. This hypothesis can be expressed more precisely by the following two points:

Hypothesis 3.1

Subcategorization information has a great impact on highly ambiguous arguments, whereas the recognition of low ambiguous arguments tends to remain stable because the grammar rules themselves are expressive enough to capture these arguments.

Hypothesis 3.2

Fine-grained subcategorization frame classes are able to capture arguments more precisely than coarse-grained subcategorization classes.

Previous and current sections have set out to explain the motivation for this research, the aim of this thesis and our main hypotheses. Next, the organization of this thesis is described.

1.4 Thesis Structure

This thesis consists of 9 parts namely:

- Cahpter 1. Introduction
- Chapter 2. Trends in Parsing (State of the Art)
- Chapter 3. Natural Language Ambiguity in Parsing (State of the Art)
- Chapter 4. Methodology
- Chapter 5. FreeLing Dependency Grammars (Development)
- Chapter 6. Dependency Grammars Evaluation (Development)
- Chapter 7. Exploring PP-attachment (Experiments)

- Chapter 8. Improving Argument Recognition (Experiments)
- Chapter 9. Conclusions

The present explanation (§1) is a general introduction to the research carried out in this thesis. A general overview of parsing in the area of Natural Language Processing is provided in order to detect the necessities that are still a problem and which motivate the research of this thesis (§1.1). Accordingly, a main and a secondary goal are established in order to solve these issues (§1.2): (1) improvements of the performance in parsing with knowledge about highly ambiguous linguistic phenomena, and (2) the design of a global evaluation task method for parsing that takes care of the analysis of errors quantitatively and qualitatively. From the main goal, three hypothesis are stated in order to validate in the conclusions of this thesis and which are formulated as follows (§1.3): (1) syntactic grammar rules provide an acceptable solution for the majority of constructions except for ambiguous syntactic phenomena, (2) statistical knowledge integrated in the grammar also improves the accuracy of the grammar performance, and (3) linguistic knowledge added in a rule-based grammar contributes to the improvement of the grammar performance. Finally, the section 1.4 provides an outline of the structure of this study.

The state of the art of parsing focuses on two aspects: main trends in parsing (§2) and natural language ambiguities (§3).

The chapter about Trends in Parsing (§2) focuses on three main aspects of parsing: Theoretical Frameworks in Parsing (§2.1), Methodological Frameworks in Parsing (§2.2) and Language Diversity in Parsing (§2.3). The chapter 2.1 presents the main linguistic theories applied in parsing. In particular, the theories presented correspond to Constituency Grammars (§2.1.1) focusing on Phrase Structure Grammar (§2.1.1.1), and Dependency Grammars (§2.1.2) paying special attention to Meaning-Text Theory (§2.1.2.1), Link Grammar (§2.1.2.2) Constraint Dependency Grammar (§2.1.2.3) and Extensible Dependency Grammar (§2.1.2.4). Furthermore, Unification Grammars (§2.1.3) and, specifically, Head-Driven Phrase Structure Grammar (§2.1.3.1) are described. The chapter concludes with a discussion about the implementation of the frameworks presented in parsing (§2.1.4). Chapter 2.2 about the methodological approaches used in parsing highlights the concepts of projective paring (§2.2.1) and deterministic parsing (§2.2.2). Next, the distinction between parsing following a rule-based strategy (§2.2.3) or a statistical-based approach (§2.2.4) is provided. Finally, the distribution of parsing tools and resources in languages is discussed and parsers and grammars developed in the languages concerned with this study are listed (§2.3).

The chapter 3 presents the state of the art of the two linguistic phenomenon studied in the experiments carried out. Firstly, in section §3.1 the phenomenon of ambiguity in natural language is explained according to the problems that are involved in parsing. Then, the problem of the prepositional phrase attachment is presented (§3.2) by defining it from a linguistic point of view (§3.2.1), and by discussing the studies about automatic learning in order to disambiguate the attachment of the prepositional phrase (§3.2.2). On the other hand, the limitations in recognizing arguments in parsing are explained in (§3.3), where the problem is defined (§3.3.1) and concrete proposals for automatic recognition are described (§3.3.2).

Following the state of the art, methdology is explained (§4). This chapter focuses on the methodology followed in the development part of this research and the steps followed and resources used in order to achieve the goals of this thesis.

In the development part, the tools that are developed to carry out this research are detailed: FreeLing Dependency Grammars (§5) and Dependency Grammars Evaluation (§6).

The chapter 5 relies on the dependency grammars specifically developed for this research. The chapter is organized into the explanation of the parser where the grammars are implemented (§5.1) and the architecture of the grammars (§5.2) consisting of attachment rules (§5.2.1) and labelling rules (§5.2.2). After this technical description, a detailed discussion of the syntactic criteria elaborated for the grammars developed in this thesis is provided (§5.3). The section is opened by a theoretical overview about the nature of syntactic heads (§5.3.1). In particular, the most controversial syntactic structures concerning the head selection are explained from the point of view of the Generative Grammar and the Dependency Grammar: auxiliary (§5.3.2), prepositional phrase (§5.3.3), subordinate clauses in §5.3.4 (substantive and adverbial clauses in §5.3.4.1, relative clause in §5.3.4.2, free relative clause and indirect question in §5.3.4.3), and coordination (§5.3.5). After every discussion, the criterion established in the grammars developed is explained. Apart from the structural criteria, the set of dependency relations defined for the grammars is listed and mapped to the Universal Dependencies (§5.4). The last section of the chapter is dedicated to the explanation of the development process of the grammars (§5.5).

The following chapter (§6) focuses on the method to measure the accuracy of the dependency grammars developed by carrying out an empirical evaluation. Firstly, general aspects of a design for evaluation task are described and the main method of the evaluation performed in this research is explained (§6.1). Next, the statistical metrics used to calculate the quality of the grammar are described (§6.2). Furthermore, the databases used in this task are detailed (§6.3). In particular, several tasks carried out to adapt a syntactically annotated resource to the criteria of the dependency grammars presented in this study in order to perform a quantitative evaluation (AnCora Corpus in §6.3.1.1 and Tibidabo Treebank in §6.3.1.2) are presented (§6.3.1). On the other hand, the test suite developed in this research in order to analyse the grammars from a qualitative point of view is described (§6.3.2). Finally, an evaluation task is carried out with the grammar developed in §5 and the results of this first evaluation are analysed from the quantitative and qualitative points of view (§6.4).

After the development, the experiments of this research are presented. This thesis presents two experiments that explore PP-attachment disambiguation (§7) and a proposal for argument recognition (§8).

Chapter §7 is the first chapter aimed at explaining if the integration of knowledge in parsing improves the accuracy of the grammars. In particular, in this experiment an automatic learning approach is chosen in order to first approach PP-attachment disambiguation. Firstly, a preliminary experiment based on a supervised classifier is presented (§7.1). Secondly, the framework followed in this experiment is presented and the distributional method is described mentioning particular works that are taken into account in the experiment (§7.2). Next, the general framework of distributional vectors is described (§7.2.1) in order to proceed with the explanation of the first of a series of experiments about learning language models of PP-attachment by learning the distributions of words. The first experiment of a series of experiments following an unsupervised method by learning word embeddings is detailed. The training data (§7.2.2) for naive supervised classifiers (§7.2.2.1) and for word embeddings models developed (§7.2.2.2), and test data are described (§7.2.2.3). The explanation of the development and the results of naive supervised classifiers (§7.2.3) and of language models learned by word embeddings (§7.2.4) is provided. In addition, a detailed description of the integration process in the grammars of the PP-attachment knowledge learned (§7.3). Finally, an exhaustive evaluation task of several versions of the grammars using the different information of models learned is presented (§7.4) together with an explanation of the evaluation experiments carried out (§7.4.1), a description of the results (§7.4.2) and a deep analysis of these results (§7.4.3).

Chapter §8, Improving Argument Recognition, presents the second approach of knowledge integration in the grammars. In this experiment, the acquisition of linguistic knowledge and, more precisely, syntactic-semantic knowledge is explored (§8.1). The chapter describes an initial subcategorization classes created for the dependency grammars developed (§8.1.1), and the redesign of these classes by creating a new lexicon from a syntactically annotated corpus (§8.1.2). Next, the integration of subcategorization information in the grammars is explained (§8.1.3). The last section of this chapter relies on the evaluation task of the grammars (§8.2) using several versions of the subcategorization classes (§8.2.1). Then, the results of the evaluation experiments are described quantitatively and qualitatively taking into account the accuracy (§8.2.2), the precision (§8.2.3) and the recall metrics (§8.2.4). A detailed analysis of the results related is provided (§8.2.5). Finally, section §8.3 focuses on a general comparison of the results of the grammars developed in this research and the parsers and grammars of the state of the art.

The last part corresponds to the conclusions of the research of this thesis (§9). A recapitulation of this research by reviewing every aspect considered in this study. Then, a global analysis of the results of the experiments is performed and the validation of the hypothesis is provided. Following this analysis, the contributions of this research are listed. In addition, the new aspects and problems that the experiments surfaced is discussed for further research.

Chapter 2

Trends in Parsing

Parsing is the field of Natural Language Processing (NLP) and Computational Linguistics (CL) which takes care of the task of automatically analysing the syntactic structure of input sentences. In particular, the problem that automatic syntactic analysis aims to solve, known as parsing problem, is to offer target mapping of the input sentences to their corresponding syntactic analysis. Then, the principles, the strategies and the tools to solve the parsing problem are the areas which this field is responsible for.

The repertoire of principles, strategies and tools implemented in parsing refers respectively to the theoretical frameworks (i.e. linguistic theories), the methodological frameworks (i.e. rule-based, data-driven) and the programs in which all this apparatus is implemented to deal with the parsing problem (i.e. parsers).

In this chapter, these three basic principles will be presented in detail. Firstly, the main theoretical linguistic frameworks which have been used in parsing will be described (§2.1). This includes the list of linguistic grammars around the concepts of constituency (§2.1.1), dependency (§2.1.2) and unification (§2.1.3). Furthermore, within the explanation of every framework, particular grammars used in parsing will be explored (e.g. Phrase Structure Grammar, Meaning-Text Theory, Head-Driven Phrase Structure Grammar, among others). Finally, the suitability of the most influential grammars and their application in parsing will be discussed at the end of this section (§2.1.4).

The second part of this chapter will be dedicated to the methodological frameworks used in parsing (§2.2). Firstly, the projective and deterministic strategies will be introduced (§2.2.1 and §2.2.2). After this description, the section will focus on the distinction between rule-based approaches (§2.2.3) and data-driven approaches (§2.2.4). Likewise in the theoretical frameworks explanation, the main strategies of both methods will be presented (e.g. Context-Free Grammar, Constraint Grammar, Transition-Based and Graph-Based strategies).

This chapter will be concluded with a discussion about language diversity in parsing (§2.3). Whether languages tend to be equally represented in parsing will also be explored. Also, the language distribution diversity will be compared among the several theoretical and methodological frameworks presented in the previous sections.

2.1 Theoretical Frameworks in Parsing

In this section, the main linguistic frameworks implemented in parsing will be explained. Specifically, three general frameworks have been closely related to the automatic syntactic analysis: Generative Grammar (GG), Dependency Grammar (DG) and Unification Grammar (UG).

These frameworks are syntactic theories with a strong philosophical basis to explain the principles of language. Despite this, their conceptualization in formal grammars makes it possible to process their principles (e.g. every sentence has a syntactic structure) and the mechanisms for representing these principles (e.g. the syntactic structure of a sentence can be represented with links between words) in computer programs like parsers.

Furthermore, every linguistic framework offers different perspectives about the target syntactic analysis of the sentence. Choosing a particular framework determines the steps, the strategies, the representation of the information and the solution as a syntactic analysis that the parsing task performs. Therefore, the selection of the theoretical framework is essential in order to provide an answer to the parsing problem according to our goals.

In the following sections, the general theoretical basis of GG (§2.1.1), DG (§2.1.2) and UG (§2.1.3) will be detailed. Every framework will be extended with an explanation of particular formal grammars that have been applied in parsing. In particular, Phrase Structure Grammar will be presented as the most typical constituency grammar with a parsing correlate (§2.1.1.1). Among dependency-based grammars, the Meaning-Text Theory (§2.1.2.1), the Link Grammar (§2.1.2.2), the Constraint Dependency Grammar (§2.1.2.3) and the Extensible Dependency Grammar (§2.1.2.4) will be described. Concerning the unification approach, the Head-Driven Phrase Structure Grammar (§2.1.3.1) will be explained as one of the unification grammars most frequently used in parsing.

Finally, the implementation of these formalisms in parsing will be discussed at the end of this section (§2.1.4). It will be argued that their success is due to their expressiveness of the linguistic description and their adequacy in fulfilling the automatic syntactic analysis goals.

2.1.1 Constituency Grammars

Constituency-based grammars are build on the concept of constituency. This has been an underlying concept in the study of language and it was articulated explicitly in the mid-1960s by the studies of Chomsky (1965) in the framework of GG.

From a generative point of view, language is an abstract structure of linguistic signs organized in a hierarchy. This hierarchy is formed by the combination of atomic linguistic structures themselves or with more complex linguistic structures to create new complex structures (**constituents** or phrases). This process operates until the highest unit of the hierarchy is reached (the sentence). Broadly speaking, lexical units are classified into parts of speech (e.g. nouns, verbs, adjectives, etc.), the parts of speech are grouped into constituents or phrases (e.g. noun phrase, verb phrase, adjective phrase, etc.) and constituents combine to form clauses and/or a sentence (e.g. main clause, relative clause, etc.). Therefore, the GG considers that syntax is concerned with the **linguistic structures** and their **combination principles**.

In order to explain the principles by which a hierarchy is built, one of the main issues that the GG has to face is the distinction between the set of parts of speech with full meaning and the set of parts of speech whose meaning expresses some grammatical feature. Despite the fact that we can identify word classes intuitively, there is no consensus on a general solution among authors because this is not a trivial issue. Then, the classification problem relies on the definition of which parts of speech are **lexical categories** and which ones are **functional categories**. The former category refers to word classes that are semantically full, may have inflection, are morphologically independent, may accept complements, and are open sets of lexical units. The latter category groups word classes that work as relational elements, tend not to be independent either phonologically or morphologically, and are close sets of lexical units.

If the distinction of the parts of speech between lexical or functional categories is unsystematic, the set of units which are able to appear in the **syntactic nucleus** position is also unclear. The GG argues that the lexical unit determines the syntactic category of the syntactic nucleus and endocentric categories necessarily contain a nucleus. According to this, lexical categories are the only ones that are endocentric and able to be in the nucleus position, whereas functional categories are not. Since this is an unsolved issue, we will come back to this discussion in §5.3 in order to detail the syntactic criteria and the set of syntactic nuclei for the development of the proposal presented in this dissertation.

The step from a syntactic structure to a more complex one is due to the above mentioned combination principles. Because language is inherently linear, the syntactic structures are not combined randomly, but follow this principle of linearity. The GG explains the combination of language units, constituents and more complex syntactic structures due to the operations performed by **phrase structure rules**. These rules operate explicitly to describe the principles that structure the natural language in the sentence by the use of a formal language (§2.1.1.1).

Aside from the traditional GG, other trends in the GG and some post-generative grammars take the notion of constituency as one of the central concepts that describe the grammar of the language. That is the reason why the following frameworks can be considered as constituency-based: Phrase Structure Grammar (PSG) (Chomsky, 1959), Government and Binding (Chomsky, 1981), Minimalist Program (Chomsky, 1995) and several grammars proposed in the general framework of Unification Grammar (UG) (Functional Unification Grammar, Definite-Clause Grammars, Lexical-Functional Grammar, Generalized Phrase Structure Grammar and Head-Driven Phrase Structure Grammar).

The PSG will be described in detail in the following subsection (§2.1.1.1) as one of the constituency-based grammars widely implemented in parsing. In addition, the UG will be presented, but in a separate section (§2.1.3). Although constituency is one of the central concepts of the UG, this framework is built on different principles and properties which make it independent from the constituency-based grammars.

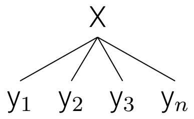
2.1.1.1 PHRASE STRUCTURE GRAMMAR

The PSG uses the mechanisms of the Context-Free Grammar (CFG), the Type 2 in the Chomsky hierarchy of formal grammars (Chomsky, 1959). Both grammars are so close that the terms usually are used interchangeably. The basic operator of the grammar is a set of **rewrite rules** (*R*) named phrase structure rules that follow the schema in (1).

(1) $\mathrm{X} \rightarrow \mathrm{y}_n$

where both symbols *X* and *y* represent syntactic categories and *n* is a finite natural number. The symbol '→' expresses the projection of *y* to *X* (structurally more complex), i.e. *X* is rewritten to *y*. This projection is not restricted to the syntactic context where *y* occurs because phrase structure rules are context-independent. Furthermore, the projection expressed by rule (1) can be represented as a syntctic tree such as (2).

(2)



Among the symbols that take part in the rule in (1), three types are distinguished. The leaves of the tree are occupied by a finite set of nodes which do no accept child nodes and are called **terminal symbols** (Σ). The intermediate nodes of the tree are filled with a different kind of symbol known as **non-terminal symbols** (*N*), which also are a finite set and are able to have child nodes depending from them. The root of the tree is defined by a **start symbol** (*S*).

Terminal symbols are the ones that cannot occur on the left side of the rule and they correspond to the atomic syntactic structures which do not decompose into other syntactic categories (3-a), non-terminal symbols can appear on both sides of the rule and are complex syntactic categories like phrases (3-b), and the start symbol is one of the symbols of the set of non-terminals which corresponds to the concept of the root of the sentence by default. The rewrite rules (3-c) make it possible to group the atomic syntactic units into complex constituents and to structure them in the shape of a constituent tree (3-d).

(3) I read a new paper about parsing

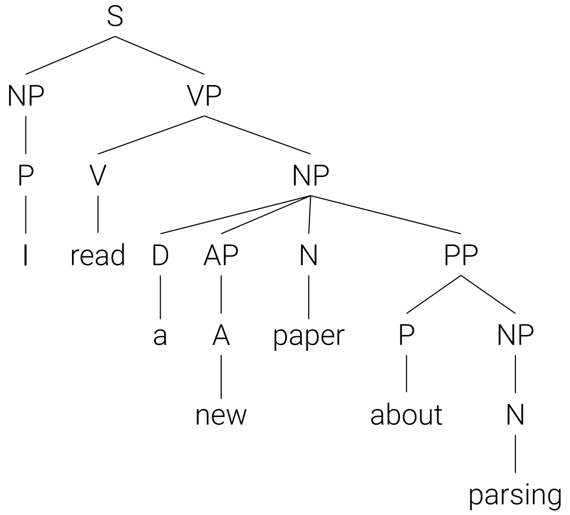
a. Σ: { I, read, a, new, paper, about, parsing }

b. *N*: { A, DT, N, P, V, AP, NP, PP, VP, S }

c. *R*:

$$\left\{\begin{array}{lcl|lcl|lcl} \text{I} & \leftarrow & \text{P} & \text{P} & \leftarrow & \text{NP} & \text{NP VP} & \leftarrow & \text{S} \\ \text{read} & \leftarrow & \text{V} & \text{V NP} & \leftarrow & \text{VP} & & & \\ \text{a} & \leftarrow & \text{D} & & & & & & \\ \text{new} & \leftarrow & \text{A} & \text{A} & \leftarrow & \text{AP} & & & \\ \text{paper} & \leftarrow & \text{N} & \text{D AP N PP} & \leftarrow & \text{NP} & & & \\ \text{about} & \leftarrow & \text{P} & \text{P NP} & \leftarrow & \text{PP} & & & \\ \text{parsing} & \leftarrow & \text{N} & \text{N} & \leftarrow & \text{NP} & & & \end{array}\right\}$$

d.



As mentioned, phrase structure rules are independent of the context (1). However, certain syntactic combinations are possible and some other ones appear very strange due to the context in which the lexical unit occurs. This principle was already observed by Chomsky (1959). For this reason, a grammar with context-dependent rules was defined in the hierarchy of formal grammars. This class of grammar was assigned to the Type 1 of the hierarchy and given the name: Context-Sensitive Grammar (CSG).

Although CSG solves the problem from a theoretical point of view, it has not proven to be powerful enough in parsing to capture most of the language regularities. Alternatively, the CFG evolved into grammars of context-free rules with constraints (in which context can be one of the features of the context, for example). The notion of constraint will appear again in the explanation of one of the grammars of the dependency approach that implements the mechanism of constrains to tighten the action of the rules. That was named Constraint Dependency Grammar (§2.1.2.3). Finally, this topic will be extended in the methodological frameworks section and specifically in the framework of the Constraint Grammar (§(3-c)) because, in fact, a constraint is used as a methodological mechanism in parsing.

2.1.2 Dependency Grammars

As opposed to the GG, the framework presented in this section and generally known as Dependency Grammar (DG) puts aside the concept of constituency and reinforces the principles and mechanisms of dependency between lexical units.

When defining the principles about the structure of a sentence, a concept like dependency is central to the syntactic theory in order to describe how a node is linked to another one in the tree. Although language grammars have referred to this concept throughout history, the first linguist who systematized this concept was Tesnière in the framework of the Structural Syntax in the late 1950s (Tesnière, 1959). Later the foundations of dependency syntax formulated by Tesnière were extended by the work of Mel'čuk (1988) in the framework of Meaning-Text Theory (§2.1.2.1).

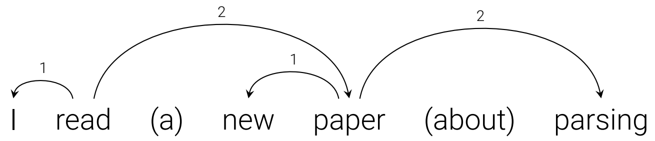
According to this author (Mel'čuk, 1988), syntax is the intermediate structure between meaning and morphology since the theory of language explains the principles of **meaning-to-text** (i.e. from semantics to morphology). Therefore, the three language levels are closely related. In this context, linguistic dependency is expressed across the levels of the language.

In the dependency approach, the meaning is the base of the language and it is expressed through a network where the nodes correspond to meanings and the arcs are predicate-to-argument relations (e.g. the sentence 'I read a new paper about parsing' in (3) represented with semantic dependencies in (4)).

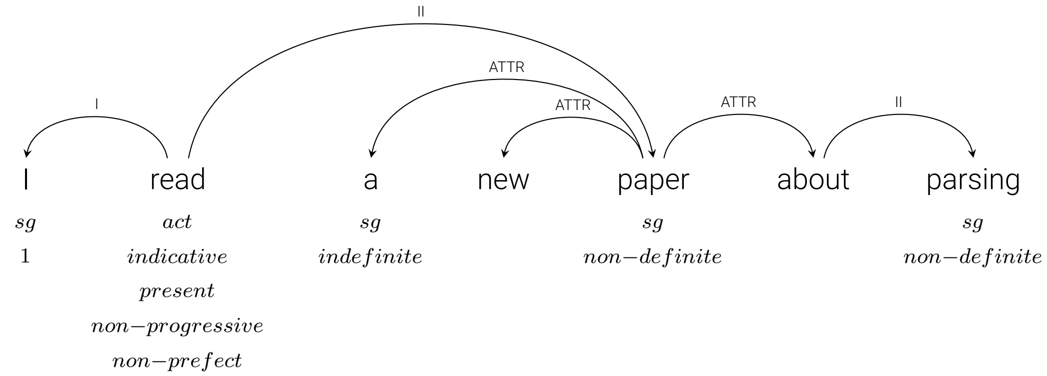
The syntactic level distinguishes between two structures, a deep structure where the universal syntactic dependency relations are manifested (5) and a surface structure in which the syntactic dependency relations are language specific (6).

Concerning the morphology, it is expressed by a deep morphological structure in which the nodes are the lexico-morphological forms of the words of the sentence and the arcs are expressed by the linear order of the words in the sentence (7). On the other hand, there is also a surface morphological structure, but this structure is not universal and there are languages which do not have it. If this morphological structure is present, it is not present in every lexical unit necessarily. For these reasons, Mel'čuk (1988) argues that is not a completely connected structure (i.e. there may be non-connected lexical units) and this structure is not explicitly specified in the linguistic description.

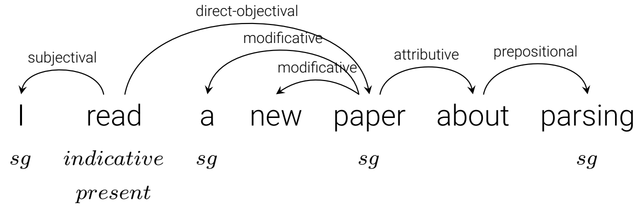
(4) I read a new paper about parsing



(5) I read a new paper about parsing



(6) I read a new paper about parsing



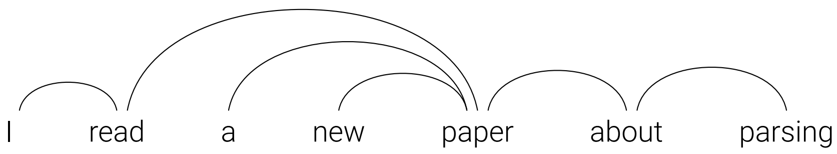
(7) I$_{sg,1}$ read$_{ind,pres,sg,not-3}$ a$_{sg}$ new$_{sg}$ paper$_{sg}$ about parsing$_{sg}$

Despite the fact the four dependency types are interdependent, they can be analysed separately as observed in the examples (4)–(7). From this point on, the focus will be placed on the syntactic dependencies without the distinction of deep versus surface, because this is the preferred language representation of the proposal presented in this dissertation.

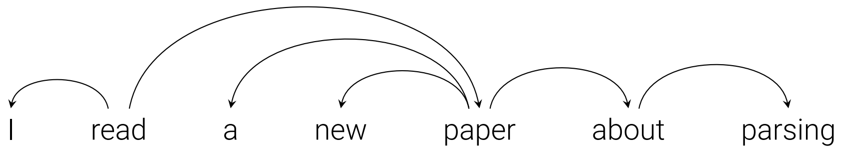
Dependency syntax is based on two assumptions (Tesnière, 1959; Mel'čuk, 1988), which are extensible to the other types of linguistic dependencies (Mel'čuk, 1988). The sentence is organized internally over a **syntactic structure** and this syntactic structure is a net of lexical units linked by **syntactic relations**. Unlike constituency-based frameworks which consider the structure to be built over constituents, syntactic dependencies happen between lexical units directly.

Then, the syntactic structure is a system of syntactic relations where the whole lexical units of the sentence are linked, i.e. **connectedness of the syntactic structure** (8). These connections between the lexical units are always directed (9) in the way that a lexical unit (governor) dominates another one (dependent) but not the other way around (10) because dependency relations are asymmetric, i.e. **directedness of the syntactic structure**.

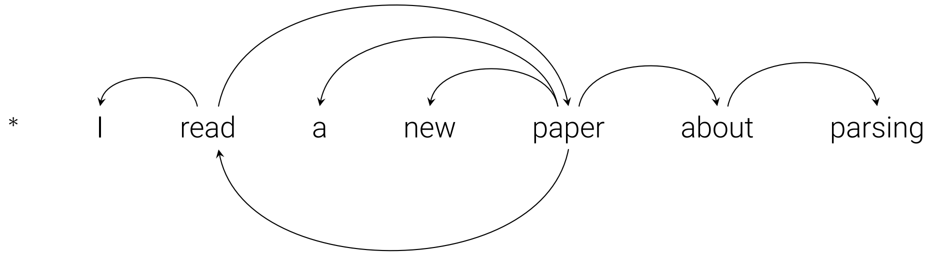
(8) I read a new paper about parsing



(9) I read a new paper about parsing



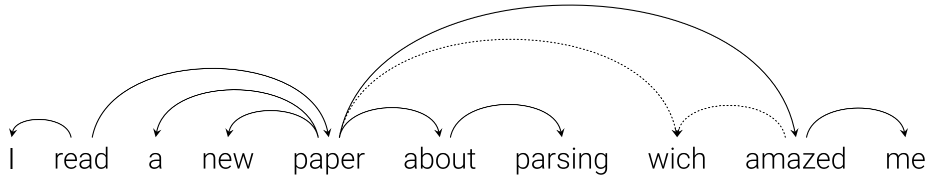
(10) I read a new paper about parsing



Furthermore, the syntactic structure is a **strict hierarchy** in which every lexical unit is connected to a single governor, except the top node which is non-governed (9). Consequently, the structure of the sentence is an acyclic directed graph resembling a tree structure.

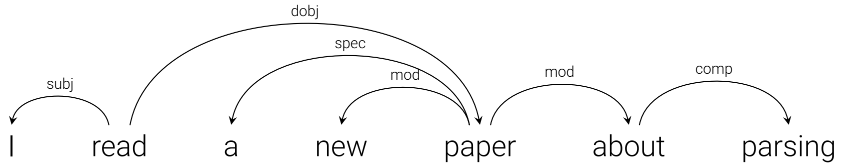
Despite this, several syntactic phenomena hardly fit with the single-head property as described. For example, relative pronouns have a double function in the sentence: marker of the relative clause and anaphoric particle that points to the antecedent of the clause (11). For this reason, it is difficult to state that relative clauses are single-headed and several implementations of the syntactic dependencies in parsing choose a multiple-headed solution (De Marneffe et al., 2006).

(11) I read a new paper about parsing which amazed me



Tesnière (1959) and Mel'čuk (1988) reinforce the important role of the syntactic relations. Their work shows that links are defined by **meaningful syntactic relations** that make explicit the semantics of every lexical unit of the sentence. That is the reason why the same kind of phrase can establish different syntactic relations because in every link has its own function according to the syntax and the semantics. In (12), there are two noun phrases ('I' and 'a new paper') that have different syntactic relations regarding the verb 'to read'. While the former is a subject, the latter is the direct object of the verb.

(12) I read a new paper about parsing



The dependency approach uses the tree structure to represent the syntactic dependencies. The syntactic dependency tree is formed by a finite set of nodes labelled with the lexical units of the sentence and linked by a finite set of arcs labelled with the syntactic relations, as shown in (12).

Since the main contribution of Tesnière (1959), new studies on syntactic dependency appeared and coexisted with the GG such as Word Grammar (Hudson, 1984), Functional Generative Description (Sgall et al., 1986) and Meaning-Text Theory (Mel'čuk, 1988). The latter is specially relevant because it was the most influential dependency grammar that simplified and systematized Tesnière (1959) theoretical framework.

Other DGs followed the same path as the Meaning-Text Theory (MTT). They proposed new aspects in the dependency theory to overcome the limitations of MTT. Examples of this are: Dependency Unification Grammar (Hellwig, 1986) which is also based on the unification principles (§2.1.3), Link Grammar (LG) which incorporates the optionality of the dependency directedness (Sleator and Temperley, 1991), Constraint Dependency Grammar (CDG) which uses constraints to restrict new connections between lexical units (Maruyama, 1990) and which influenced Weighed Constraint Dependency Grammar (Schröder et al., 2001), Functional Dependency Grammar which also makes use of constraints (Tapanainen and Järvinen, 1997), and Extensible Dependency Grammar (XDG) which aims to make more explicit the semantic-syntactic interface (Debusmann et al., 2004).

Among the above mentioned dependency-based grammars, the MTT, the LG, the CDG and the XDG will be described in the following subsections (§2.1.2.1, §2.1.2.2, §2.1.2.3, §2.1.2.4, respectively) because their contributions were applied in parsing.

2.1.2.1 Meaning-Text Theory

The MTT (Mel'čuk, 1988) appeared at a moment in the area of language technology there was a need of a theoretical framework to overcome the gap that existed between syntax and semantics after the GG. The main idea of the MTT is that language is a mapping from the meaning of a sentence to its form or text (i.e. the phonetics). Mel'čuk (1988) finds in Tesnière (1959) postulates on dependency an initial description of this idea. For this reason, the MTT takes up the dependency tradition, but it updates this linguistic framework in a formal description of the language, as shown next.

First of all, the MTT formally describes the **dependency structures** regarding the levels of the language. Specifically, it makes explicit the interdependence of semantics and syntax, as well as the link between syntax and morphology, as discussed in the previous section §2.1.2 in the representation of the natural language dependencies (4)–(7).

Besides the capability of language to be organized in different but connected levels of dependency structures, lexical units (i.e. lexemes, constructions and idioms, in terms of the MTT) contain the core information (i.e. rules and conditions) that determine their combination options. The set of lexical units of a language and their combination properties are stored in a **lexicon**, which the MTT calls an Explanatory Combinatorial Dictionary.

Finally, the MTT also contributes to formal description of the lexical relations between the lexical units as a consequence of the combination principles by means of **lexical functions** (Mel'čuk, 1996). While there are some lexical units that are semantically-driven selected (i.e. selected by their meaning and independently from the other lexical units), there are other lexical selections that are lexically-driven and that are formally explained by lexical functions, such as *Magn* or *Oper* in the example (13).

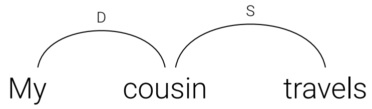
(13) a. Intensifier: **Magn**(sleep$_v$) = *deeply*, *heavily*, *like a dog*
b. Light or support verbs: **Oper**$_1$(give) = *talk*

Because formalization of the linguistic description is one of the main goals of the MTT, this framework is appropiate for NLP applications.

2.1.2.2 Link Grammar

The LG is a framework developed exclusively for the area of parsing (Sleator and Temperley, 1991). The particularities of the LG approach are the capability of creating **undirected links** (14), the possibility of **circular links** and the abandonment of the concept of the root of the sentence.

(14) My cousin travels



Although the authors do not claim this framework to be dependency-based, both frameworks are closely connected because both are context free (Gaifman, 1965; Sleator and Temperley, 1991). Like the dependency approach, the LG represents the syntactic relations in a structure of links very similar to the dependency structure and every link of the structure is labelled as in syntactic dependencies. In addition, the authors argue that it is possible to generate a link grammar from a dependency grammar, which shows that their framework can be mapped to syntactic dependencies.

2.1.2.3 Constraint Dependency Grammar

The CDG states that dealing with natural language ambiguities and proposing the best syntactic analysis for them is a constraint satisfaction problem (Maruyama, 1990). It assumes that the sentence is a fully connected graph or a **constraint network** where the nodes are variables in the form of words and the links are constraints that allow the variables to be connected (15). Then, a possible analysis that matches all the constraints is a parse tree.

(15) The athlete competes

a. *Link:* The ← athlete
Constraint:
`word(pos(x))=D ⇒ (word(lab(x))=DET, word(mod(x))=N, pos(x)<mod(x))`
Meaning: A determiner (D) modifies a noun (N) on the right with the label DET

b. *Link:* athlete ← competes
Contraint:
`word(pos(x))=N ⇒ (word(lab(x))=SUBJm word(mod(x))=V, pos(x)<mod(x))`
Meaning: A noun (N) modifies a verb (V) on the right with the label SUBJ

Firstly, every word is potentially linked to any other word of the sentence. However, all the links are disambiguated by the technique of constraint propagation. Constraints are activated to block inconsistent dependencies and propagated through the network. Therefore, the best matches are proposed as the solution for the syntactic tree.

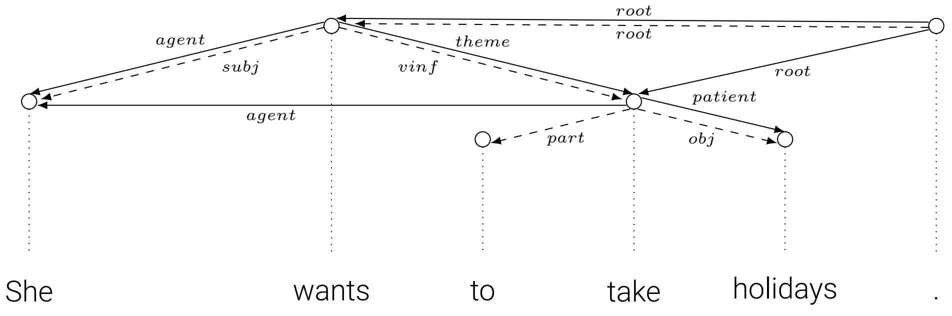
In cases where the constraints cannot disambiguate completely the sentence, weighted constraints can be used instead, which is the approach proposed by Schröder et al. (2001).

2.1.2.4 Extensible Dependency Grammar

The dependency approach states that syntax has a correlation with semantics as opposed to the traditional constituency approaches (Mel'čuk, 1988). Nevertheless, authors like Debusmann et al. (2004) point out that most of the dependency formalisms do not pay enough attention to the syntax-semantic interface. For this reason, they developed the Extensible Dependency Grammar (XDG) to describe formally the interconnection between syntax and semantics.

This formal grammar considers the language as a modular object with **multiple dimensions**. In particular, three dimensions are described relating to syntax and semantics (16): Immediate Dominance (which is represented by unordered and non-projective syntactic dependency trees), Linear Precedence (which represented by ordered and projective syntactic dependency trees) and Predicate-Argument Structure (which is represented by an acyclic graph).

(16) She wants to take holidays.



In the example (16), every node represented by a circle is related to a lexical unit by a projection line (dotted line). Every node is linked to another one by an edge labelled and dashed in the syntactic representation, and a labelled and continuous edge in the semantic representation. Every edge corresponds to a syntactic relation in the syntactic representation (e.g. *subj*, *obj*, *vinf*, *part*, *root*) and a semantic role in the semantic representation (e.g. *agent*, *theme*, *pacient*, *root*).

2.1.3 Unification Grammars

Diverse declarative formalisms were developed in the areas of theoretical linguistics and computational linguistics during the 1980s. This is the case of Categorial Grammar (Ajdukiewicz, 1935; Bar-Hillel, 1953), Generalized Phrase Structure Grammar (Gazdar et al., 1985), Head Grammar (Pollard, 1984) and Head-Driven Phrase Structure Grammar (Pollard, 1985), Lexical Functional Grammar (Bresnan, 1982) and Functional Unification Grammar (Kay, 1983).

Despite the differences among them, they have been classified into the general framework of the Unification Grammar because they are built on two common properties. Firstly, linguistic objects are defined by mathematical objects called **feature structures**. Secondly, the feature structures are merged by an operation of **unification**. Then, the representation of language which the UG proposes is formed by **lexical entries** represented in terms of feature structures, and **rules** that determine the possible combinations of the feature structures (from the most simple ones to the most complex ones).

The feature structures are a set of features and their values which can be variable (17-a), constant (17-b) or another feature structure (17-c). The use of complex features makes it possible to describe lexical entries with syntactic and semantic ones (e.g. agreement, verb predicate structure, thematic roles, etc.) such as in (17-d).

(17) a.

$$\begin{bmatrix} number : \alpha \\ gender : \text{feminine} \end{bmatrix}$$

b.

$$\begin{bmatrix} cat : \text{noun} \\ number : \text{singular} \\ gender : \text{femine} \\ lemma : \text{woman} \end{bmatrix}$$

c.

$$\begin{bmatrix} cat : \text{noun} \\ agreement : \begin{bmatrix} number : \text{singular} \\ gender : \text{feminine} \end{bmatrix} \\ lemma : \text{woman} \end{bmatrix}$$

d.

$$\begin{bmatrix} cat : \text{verb} \\ subject : \begin{bmatrix} cat : \text{np} \\ role : \text{experiencer} \end{bmatrix} \\ lemma : \text{rest} \end{bmatrix}$$

The unification of features happens by combining the information of two features to obtain a feature structure that includes all the information. Therefore, the unification of two feature structures α (18-a)–(19-a) and β (18-b)–(19-b) in a more general feature structure γ (18-c)–(19-c) is true if $\alpha \subseteq \gamma$ and $\beta \subseteq \gamma$, or $\gamma = \alpha \cup \beta$ (18)–(19).

(18) a.

$$\begin{bmatrix} cat : \text{np} \end{bmatrix}$$

b.

$$\begin{bmatrix} agreement : \begin{bmatrix} number : \text{singular} \end{bmatrix} \end{bmatrix}$$

c.

$$\begin{bmatrix} cat : \text{np} \\ agreement : \begin{bmatrix} number : \text{singular} \end{bmatrix} \end{bmatrix}$$

(19) a.

$$\begin{bmatrix} cat : \text{np} \\ agreement : \begin{bmatrix} number : \text{singular} \\ gender : \text{feminine} \end{bmatrix} \end{bmatrix}$$

b.

$$\begin{bmatrix} agreement : \begin{bmatrix} number : \text{singular} \\ case : \text{nominative} \end{bmatrix} \end{bmatrix}$$

c.

$$\begin{bmatrix} cat : \text{np} \\ agreement : \begin{bmatrix} number : \text{singular} \\ gender : \text{feminine} \\ case : \text{nominative} \end{bmatrix} \end{bmatrix}$$

Alternatively, the unification can be **reentrant**. Specifically, a feature structure is reentrant if two feature share a common value (20). In order to make the shared values explicit, they are coindexed in the feature structure as follows: one of the structures keeps the shared values with an index added and the other one just contains the same index pointing to the shared values (20).

(20) a.

$$\begin{bmatrix} cat : \text{np} \\ role : \text{experiencer} \\ agreement : \ \boxed{1} \ \begin{bmatrix} number : \text{singular} \\ person : \text{third} \end{bmatrix} \\ lemma : \text{woman} \end{bmatrix}$$

b.

$$\left[\begin{array}{ll}
cat: \text{vp} & \\
agreement: \boxed{1} & \\
subject: & \left[\begin{array}{l} cat: \text{np} \\ role: \ \text{experiencer} \end{array}\right] \\
lemma: \text{rest} &
\end{array}\right]$$

In the following subsection the Head-Driven Phrase Structure Grammar (HPSG) will be presented (§2.1.3.1) in order to observe the principles of the UG applied in a particular grammar. Particularly, the HPSG will be described since it is one of the frameworks most widely used in parsing.

2.1.3.1 Head-Driven Phrase Structure Grammar

The Head-Driven Phrase Structure Grammar or HPSG (Pollard, 1985) is considered a formalism within the unification grammar framework (e.g. uses the unification operation, feature structures, etc.) and it is a successor of the Generalized Phrase Structure Grammar (Gazdar et al., 1985). One of the contributions of the HPSG is the establishment of the principles and the mechanisms that make explicit the connection between syntax and semantics of the linguistic signs.

The linguistic sign is the basic linguistic unit of the language and it has subtypes such as words (lexical) and phrases (non-lexical sign). The information of the sign is expressed by a feature structure that defines orthographic (PHON), syntactic (SYNSEM), semantic (SYNSEM) and pragmatic aspects of that sign (21).

(21)

$$\left[\begin{array}{ll}
PHON & \langle dreams \rangle \\
SYNSEM & \left[\begin{array}{ll}
synsem & \\
CAT & \left[\begin{array}{ll}
category & \\
HEAD & verb \\
VALENCE & \left[\begin{array}{ll}
SUBJ & \left[\begin{array}{ll}
synsem & \\
CAT|HEAD & noun \\
CONT\ \boxed{1} & \left[\begin{array}{l} ref-index \\ PER\ \ 3rd \\ NUM\ \ sg \\ TEN\ \ pres \end{array}\right]
\end{array}\right] \\
COMP & \langle\rangle
\end{array}\right]
\end{array}\right] \\
CONT & \left[\begin{array}{l} context \\ DREAMER\ \ \boxed{1} \end{array}\right]
\end{array}\right]
\end{array}\right]$$

This framework is based on principles, grammar rules and lexicon entries. Concerning the principles, language works according to universal principles (e.g. Head Feature Principle, Binding Inheritance Principle and Subcategorization Principle) and language specific principles (e.g. Constituent Order Principle). The grammar rules are classified into three types, constituency rules (i.e. to form constituents from lexical units, to create new constituents from other constituents), linear precedence rules (i.e. to state the surface word order) and lexical rules (i.e. to make generalizations about the lexical entries, to express the several subcategorization frames given a lexical entry). Finally, the lexical entries contain the majority of the syntactic and semantic information, as observed in (21). As a consequence of the richness of the lexical entries, the set of grammar rules is simplified considerably.

2.1.4 Implementation in Parsing

In the previous sections, the most important frameworks in parsing (i.e. the GG focusing on the PSG, the DG, the UG with special attention to the HPSG) have been described in detail. Based on the characterization of these frameworks, the advantages of their implementation will be discussed in this section and complemented with examples.

Nowadays, parsing and DG are closely related (even nowadays the term 'parsing' is used meaning 'dependency parsing' quite often). Furthermore, dependency parsing has also become the preferred approach by most of the NLP tools like Information Extraction or technologies that require deep, robust and certain kinds of semantic representations.

Several reasons lead to the DG being the most successful framework in parsing and in NLP. Although an analysis of them would need an exhaustive study, the three main types will be described briefly in the following lines (Mel'čuk, 1988).

The dependency approach integrates semantics into the linguistic description. The syntactic dependency hierarchy is formalized as a structure of nodes linked and labelled with their dependency relations. The resulting tree of a sentence is **close to the semantic dependency structure** since the syntactic realization of the the verb predicate arguments is already expressed. On the other hand, the constituency approach considers that syntax is autonomous. Consequently, semantics is barely analysed in the GG.

The fact that DG labels every link of the syntactic structure with **a syntactic relation provides a more robust analysis** than a pure constituency representation. As Mel'čuk (1988) argues, if the syntactic relations are added in the constituency analysis, the representation of constituents is not useful. For this reason, some grammars of the unification approach, such as LFG and HPSG, skip the representation of non-terminal nodes and use the syntactic relations in the output, which make them similar to the dependency approach.

In addition, the dependency approach breaks the tradition of representing linear order in syntax. Consequently, **the description of the word order becomes a more simple task** than in the constituency approach. The order can be represented according to the position of a syntactic dependent with respect to its syntactic governor. Furthermore, Mel'čuk (1988) states that dependency structure satisfies the projection principle (Chomsky, 1981) since it prevents branches to cross. However, this point needs to be specified because there are unbounded dependencies and discontinuities in natural language (i.e. 'The restaurant will be open on the weekends, where we usually go'). Therefore, it would be more precise to argue that projectivity is respected most of the times and some exceptions violate this principle.

First parsers based on the dependency approach appeared in the 1960s (Hays, 1964; Gaifman, 1965) and later on in the early 1990s (Maruyama, 1990). However, the big success of the dependency approach begins in the late 1990s and the early 2000s. Among the extensive list of dependency parsers, some can be distinguished due to their contribution to the field, such as the statistical parser based on support vector machines of Yamada and Matsumoto (2003), MSTParser (McDonald et al., 2005), Standford Parser (De Marneffe et al., 2006), MaltParser (Nivre et al., 2006), the tree adjoining grammar parser of Carreras et al. (2008), the transition-based non-projective parser of Bohnet and Nivre (2012), and the transition-based parser with spinal trees of Ballesteros and Carreras (2015).

In addition, such contributions have been reinforced by several conferences, competitions and discussion groups, like the SIGNLL Conference on Computational Natural Language Learning from 2006 to 2009 which focused on dependency parsing (Buchholz and Marsi, 2006; Nivre et al., 2007; Surdeanu et al., 2008; Hajič et al., 2009), the International Conference on Parsing Technologies, the International Conference on Dependency Linguistics, as well as the SIGPARSE (the ACL Special Interest Group on Natural Language Parsing).

Concerning phrase structure parsers, although PSG is not one of the most used frameworks nowadays, some statistical parsers appeared in the late 1990s and early 2000s that work with phrase structure trees (Charniak and Johnson, 2005; Collins, 1999; Klein and Manning, 2003). As De Marneffe et al. (2006) remark, although these parsers are able to generate high accuracy trees, constituency annotated treebanks from which parsers were trained were slowly converted to the dependency representation. Consequently, parsers working with phrase structure representations are not the main trend in parser development any more.

Among the UG framework, HPSG has also been quite successful in parsing because similarly to the DG it provides deep and robust representations close to the semantic level. Furthermore, this grammar operates with the unification mechanism and is organized with modules, which make it suitable to be implemented in parsing. So there have been several initiatives in this field like the English Resource Grammar (Copestake and Flickinger, 2000), the work on HPSG parse disambiguation of Toutanova et al. (2002), the JACY Japanese Grammar (Siegel and Bender, 2002), the German HPSG grammar GG (Müller and Kasper, 2000), the Portuguese Resource Grammar (Costa and Branco, 2010), among others.

Recapitulation

This section has been focused on the theoretical framework implemented in parsing. The main linguistic approaches (constituency, dependency and unification) have been presented in detail, explaining the principles and the properties of the main linguistic theories (Generative Grammar, Dependency Grammar and Unification Grammar) as well as the principal formal grammars that have their correlate in parsing (e.g. the Phrase Structure Grammar in the constituency approach, the Meaning-Text Theory as one of the most important dependency approaches for NLP, and the Head-Driven Phrase Structure Grammar of the unification approach). In the second part of this section, the theoretical reasons that lead to the formal grammars success in parsing have been argued and complemented with examples about particular parsers. The diversity of formal grammars applied in parsing make explicit that the automatic analysis of the syntactic structure of linguistic data is not a straightforward task and every framework contributes differently to the syntactic description.

2.2 Methodological Frameworks in Parsing

In this section, the parsing task will be presented from the methodological side, i.e., concerning the several methods that have been used in order to provide a solution to the parsing problem. However, before presenting the main methods used in parsing, the concept of knowledge and its relation to parsing needs to be introduced.

In this context, knowledge is equivalent to the combinatorial principles and to the syntactic and semantic properties of the words that make it possible for the words to be organized in a syntactic structure. Since parsers are interpreters of pieces of information, they need certain kinds of knowledge that indicates to them how to build the syntactic structure of an input sentence in order to build an appropriate syntactic analysis.

From a methodological point of view, the knowledge which a parser works with is managed essentially in two different ways. A parser can take a rule-based approach (i.e. the knowledge consists of formal rules that determine the grammaticality of the analysis of the input sentences) or a statistical approach (i.e. the knowledge refers to machine learning strategies to inform about the well-formedness of the input sentences). In addition, a third approach is also present in some parsers, which combines features of both previous methods. Because the hybrid approach is not consolidated in the area of parsing, several implementations following this approach will be mentioned in a generic way.

On the other hand, two another methodological aspects must be considered due to their importance in the parser internal representation of the input sentences and in the output representation. They refer to the projectivity principle (i.e. respecting or violating the linear order of the sentence in the parser internal representation or the output representation of the input sentences) and the determinism in the linguistic representation (i.e. building a unique or multiple linguistic representations).

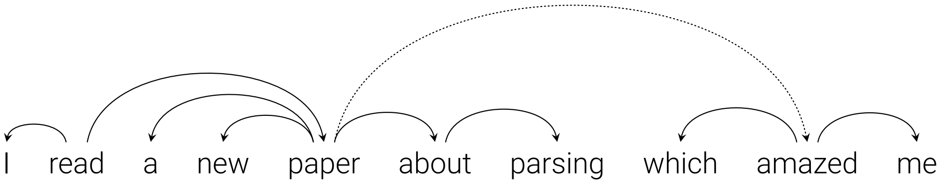
Likewise the selection of a particular linguistic framework for the construction of a parser (§2.1), choosing a rule-based or a statistical approach, a way of representing the linear order and the determinism, completely determines the architecture of the parser. It restricts the repertoire of algorithms to be implemented, as well as it influencing on the representation of the output. Finally, it also conditions the appropriateness of the results of the syntactic analysis.

In the following subsections, the concepts of projectivity and determinism applied to parsing will be explained (§2.2.1 and §2.2.2). The main part of this section will be dedicated to both rule-based and statistical methods in §2.2.3 and §2.2.4 respectively. For every approach, the main trends will be exposed as well as some hybrid approaches developed. Specific examples of parsers will be provided in order to illustrate each approach.

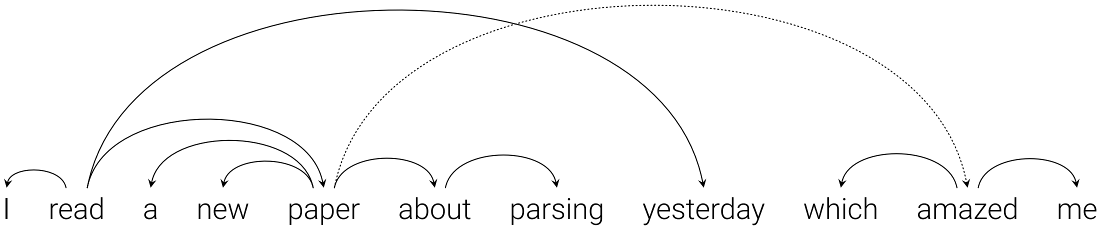
2.2.1 Projective Strategies

The **projective strategies** refer to the implementation of the projection principle (Chomsky, 1981) in parsing. Basically, this principle states that lexical units project to more complex units (i.e. constituents) in order to combine with other projected lexical units in the syntactic structure. In the syntactic tree representation, lexical entries are projected into more complex syntactic structures by projection lines which preserve the linear order of the sentence (22), i.e., there are no crossing lines in the tree representation.

(22) I read a new paper about parsing which amazed me



(23) I read a new paper about parsing yesterday which amazed me

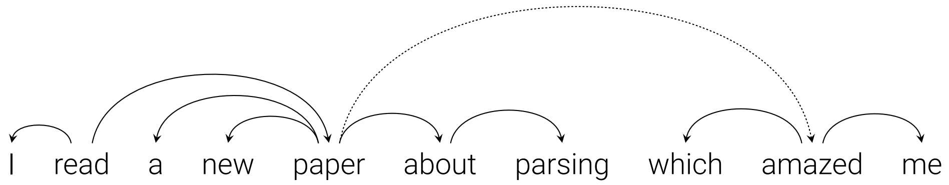


Consequently, a projective parser preserves the projection principle (22), so it prevents tree arcs to cross (Nivre, 2003; Carreras, 2007). On the other hand, a parser violating the projection principle (23) so that it accepts crossing tree arcs is a non-projective parser (Tapanainen and Järvinen, 1997; McDonald et al., 2005; Attardi, 2006; Bohnet and Nivre, 2012).

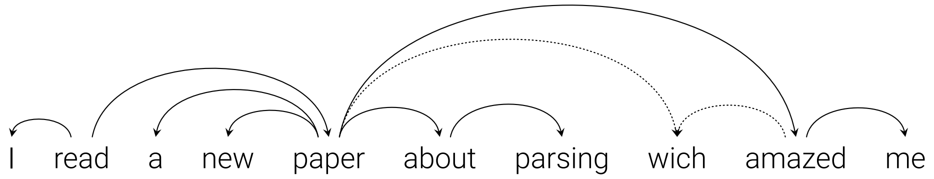
2.2.2 Deterministic Methods

The **deterministic methods** decide how the parser builds the syntactic tree regarding structural ambiguities. For example, the sentence like 'Visiting relatives can be so boring' can be interpreted as the relatives who visit me are very boring or as the fact of visiting relatives is a boring thing to do. Depending on the kind of method, the parser manages the ambiguities in a different way. In particular, two kinds of automata can be implemented to deal with ambiguities, a deterministic finite automaton or a non-deterministic finite automaton (Rabin and Scott, 1959), which are applied in the deterministic strategy and the non-deterministic strategy respectively.

(24) I read a new paper about parsing which amazed me



(25) I read a new paper about parsing which amazed me



The deterministic strategy attempts to solve the ambiguity by providing a unique solution (24). Deterministic parsers only retrieve a tree structure as the appropriate solution to the parsing problem (Briscoe, 1987; Tapanainen and Järvinen, 1998; Yamada and Matsumoto, 2003; Nivre, 2004; Bick, 2006). On the other hand, the structural ambiguity is not a problem for the non-deterministic strategy because all the possible solutions are retrieved (25). Therefore, the output of non-deterministic parser contains the set of alternative tree structures as the appropriate solution to the parsing problem (Luque et al., 2012; Björkelund and Nivre, 2015).

2.2.3 Rule-based Approach

A rule-based approach uses **formal rules** as a source of knowledge that inform whether a sentence is possible in the formal language described by the set of rules, so that it is possible to associate to the input sentence a syntactic structure according to what the rules state. The set of rules is finite and they use a formal language to express the properties of the natural language. The set of rules are grouped into a **grammar**, which can be compared to the traditional concept of a grammar of the language (i.e. the set of rules that express the principles and the syntactic and semantic properties of a particular language).

A parser, called **rule-based parser**, uses the grammar as a source of knowledge. Then, a rule-based parser interprets the information of the formal rules and make decisions concerning the grammaticality of the input sentence. The basic operation is **matching** the information provided in the rules and the input sentences. Broadly speaking, if the rules state that a given input sentence is possible, there is a match. Consequently, the parser provides an analysis of the sentence according the properties of the rules. On the contrary, if the rules do not recognize the input sentence, there is no possible match and the parser cannot provide a syntactic analysis of the input sentence. Therefore, the parsing problem cannot be solved. In this case, a syntactic analysis cannot be provided unless some operation is performed on it.

Additionally, **probabilities** can be associated to every rule of the grammar. They replace the default order in which rules are applied with the order which the numeric value of the probability states (e.g. rules containing a probability closer to 0 is applied before rather than the ones with a probability closer to higher numeric values). Therefore, they are essentially used to prioritize how to resolve linguistic phenomena and two or more competing rules. These probabilities can be learned from linguistic resources or they can be weights assigned manually or acquired automatically.

There are some aspects of some linguistic formal grammars that have been used as a resource for processing information. In particular, two main linguistic frameworks, the Context-Free Grammar (CFG) and the Constraint Grammar (CG), contributed to the rule-based approach due to the **rewriting rules system** of the CFG and the **constraints** introduced by the CG to restrict the number of times that a rule can be applied. Both methods are orthogonal, since it is possible to make use of context-free grammars with constraints.

As described previously in §2.1.1.1, the CFG was identified as a constituency-based approach, like the Phrase Structure Grammar (§2.1.1.1). However, some parsing algorithms available for CFG can be also used to parse linguistic data with other linguistic frameworks like syntactic dependencies (Gaifman, 1965; Hays, 1964). The CFG uses a basic operator which is a rewriting rule independent from the context (e.g. $\mathrm{X} \rightarrow \mathrm{y}_n$). A strict CFG is mainly developed manually, but other extensions of this grammar can mix manual and data-driven methods or can be exclusively data-driven, such as the Probabilistic Context-Free Grammar (Collins, 1999; Charniak and Johnson, 2005).

On the other hand, the CG (Karlsson, 1990) considers that the rules operate as a constraint satisfaction solver (Maruyama, 1990). As explained in §2.1.2.3 about the Constraint Dependency Grammar framework, the rules do not rewrite a terminal or non-terminal symbol to a more complex symbol until the start symbol is reached. The rules in this grammar are formed by a set of constraints that restrict the probability of the rule to be applied (e.g. identify the sequence of the input sentence as *subject* if it has the feature *noun* and it *precedes a verb*). Therefore, a rule applies if the constraints are matching with the input sentence. Alternatively, the constraints can discard certain matchings (e.g. discard the sequence of the input sentence as *pronoun* if it is *followed by a noun*). Both operations are carried out by constraint propagation through a constraint network where the nodes are the words and the arcs the constraints (Tapanainen and Järvinen, 1998; Bick, 2006).

After applying this technique, some unresolved cases may still persist. Therefore, other techniques can be used to overcome these ambiguities, like the use of **weights for the constraints**. The Weighted Constraint Grammar proposes weighted rules in order to make less strict the rules using bare constraints and, consequently, to handle the gradation concerning the most prototypical structures to the less prototypical (Lin, 1998b; Schröder et al., 2001).

2.2.4 Statistical-based Approach

The statistical-based approach uses **machine learning** to represent the knowledge that will be used to parse new sentences. Broadly speaking, machine learning applied to language aims to recognize patterns in a language and classify them from big linguistic databases. Therefore, the statistical-based approach presents an extra problem to solve. Apart from the parsing problem, a **learning problem**, i.e., to induce a parsing model that allows parsing of the input sentences, has to be solved.

According to this, the learning task can be performed mainly in two ways according to the kind of data used for learning. If the model is learned from unannotated or unseen linguistic data, the learning is said to be **unsupervised**. Otherwise, annotated data is used to induce the model, which is called **supervised** learning or data-driven.

The unsupervised parsing aims to solve the parsing problem by **inducing grammar rules**, their probabilities and syntactic trees from unseen or unannotated linguistic data (Klein, 2005). Two of the strategies applied in this task are the **lexical attraction** of associated words (Yuret, 1998) and **bootstrapping**, i.e., to guide learning starting from the simplest structures and incrementally increasing the complexity of the sentences (Spitkovsky et al., 2010). Unsupervised learning does not require annotated data, so it is independent of the problems related to the linguistic annotation. However, there are some limitations associated with this approach. The main observation is that the results are not satisfactory and this area still needs a lot of progress in the future (Spitkovsky et al., 2010).

On the other hand, the data-driven approach requires exclusively **annotated data** for the learning task. Consequently, there is a considerable manual effort in developing and maintaining annotated resources, as well as parsers based on supervised learning which rely on the domains and genres used in the training data (Rimell et al., 2009). Nevertheless, data-driven parsers are the statistical parsers with the best scores.

In the data-driven approach, the majority of methods proposed can be grouped into two major strategies (McDonald and Nivre, 2011), which are named **transition-based** and **graph-based** respectively. Therefore, the majority of data-driven parsers follow one of these methods to a greater or lesser extent.

A transition-based model tries **to predict the next transition** (i.e. the next parser action) from one state to another one among the set of possible transitions given a transition history (Kudo and Matsumoto, 2000; Yamada and Matsumoto, 2003; Nivre, 2004; Nivre et al., 2006; Attardi, 2006; Bohnet and Nivre, 2012; Ballesteros and Carreras, 2015). Consequently, this method allows parsers to perform **local inferences**. The goal of a transition-based parser, then, is to find the best transitions of the input sentences according to language model learned until a condition is met. For this reason, this approach is essentially **deterministic**.

The most frequent strategy followed in transition-based parsing is the strategy based on **shift-reduce parsing**. In general terms, a shift-reduce parser performs three basic operations on a buffer (where the words of the input sentence are stored). The **transition shift** removes the current word from the buffer and it moves this word to the top of a stack, the **transition left-arc** adds a left arc to the dependency tree, and the **transition right-arc** adds a right arc to the dependency tree.

On the other hand, the graph-based strategy approaches the learning task from a different perspective. The parser defines the set of **possible graph candidates** (i.e. the possible syntactic structures expressed by a graph). The learning tasks rely on inducing a model to score all the possible arcs, so it carries out **global inferences**. Finally, the parsing task consists of predicting the best graph, i.e., the **graph with the highest score**, given an input sentence and according to the learned model (Eisner, 1996; McDonald et al., 2005; Carreras, 2007; Koo et al., 2007).

The most basic strategy in graph-based parsing is known as **arc-factored** or **edge-factored** (Eisner, 1996; McDonald et al., 2005; Carreras, 2007; Koo et al., 2007). In general terms, an arc-factored parser assigns a score (i.e. weights or probabilities) to every possible arc of the graph. The scores are obtained by computing some parameters associated with every arc of the graph and the two nodes of the arc (e.g. lemmas of parent-child, part of speech tags of both nodes, distance in number of words between parent and child, direction of the arc, etc.). To solve the parsing problem, then, the parser has to find the tree with the highest score among the best factored-arcs graph given an input sentence.

Recapitulation

Throughout this section, several parsing approaches have been presented from the methodological point of view. Firstly, two strategies about how the parser builds the tree concerning the projection principle and the structural ambiguities. According to these two strategies, parsers can be projective or non-projective and deterministic or non-deterministic. After this explanation, several methods have been described that use different strategies to manage the knowledge which the parsers work with. As observed, they are rule-based or statistical-based. While the former approach is defined by context-free rules or rules with constraints, the latter relies on unsupervised learning or supervised learning (transition-based or graph-based). The fact that there are multiple techniques and most of them can be combined shows that parsers are sophisticated tools in order to cope with the complexity of language.

2.3 Language Diversity in Parsing

Parsing is a relevant field of the NLP since most of language technologies that work with semantic information or that need robust linguistic representations use automatic syntactic analysers (i.e. Information Extraction, Machine Translation, Semantic Parsing). For this reason, this field grew at the same high-speed as language technologies, so a wide-range of linguistic frameworks (§2.1) and an intricate set of methods (§2.2) have been applied.

Despite this intense development of the field, languages are not equally distributed in terms of parsers and parsing resources. That is to say, while there are significant advances in particular languages, other ones are less represented or are barely represented.

English is in the group of most represented languages in parsing. There are a wide range and a large number of linguistic resources available in English related to syntax, which provides a robust base for developing parsers in this language. For example, large syntactically annotated corpora have been created for parsing among other purposes, such as the PennTreebank (2,881,188 tokens) of the University of Pennsylvania (Marcus et al., 1993), the Prague English Dependency Treebank (1,200,000 tokens) of ÚFAL at Charles University in Prague (Hajič et al., 2012), and the English Web Treebank (254,830 tokens) released by the Linguistic Data Consortium (Bies et al., 2012) and transformed to syntactic dependencies afterwards (Silveira et al., 2014).

Due to this intense activity over the English language, the quantity and the quality of English parsers are remarkable. Well-known statistical parsers have been developed, like Collins PCFG parser (Collins, 1999), Standford Parser (Klein and Manning, 2003), MaltParser (Nivre, 2003), MST Parser (McDonald et al., 2005), the work of Carreras (2007) on a projective graph-based parser, the Berkeley parser (Petrov and Klein, 2007), the non-projective transition-based parser of Bohnet and Nivre (2012), and the transition-based parser for spinal trees of Ballesteros and Carreras (2015).

On the side of rule-based parsers, there have also been important initiatives such as the MINIPAR dependency and generative-principle-based parser (Berwick et al., 1991), the link grammar of Sleator and Temperley (1991), the functional dependency grammar proposed by Tapanainen and Järvinen (1998), the English Resource Grammar in the HSPG framework (Copestake and Flickinger, 2000), the incremental dependency parser of Ait-Mokhtar et al. (2001), and the weighted constraint dependency grammar developed by By (2004).

Other languages that are less predominant than English in the field of parsing also have a notable presence in parsing. In the group of **European languages**, the development around the Czech language is noticeable in the ÚFAL of Charles University in Prague (Collins et al., 1999; Bojar, 2004; Ribarov, 2004; Zeman, 2009), as well as in German (Müller and Kasper, 2000; Rafferty and Manning, 2008), French (Ait-Mokhtar et al., 2001; Candito et al., 2010), Italian (Attardi, 2006) and Portuguese (Silva et al., 2010; Gamallo, 2015). In addition, the increasing interest in **Chinese** and **Arabic** can be also observed in this area due to the works on Chinese by Bikel and Chiang (2000); Levy and Manning (2003); Qian and Liu (2012) and Arabic (Green and Manning, 2010; Marton et al., 2013), among other authors. Finally, it has to be noticed that, although **Basque** is not a wide-spread language compared to English, there is a large contribution in the parsing field concerning this language (Aldezabal et al., 2003; Aduriz et al., 2004; Bengoetxea and Gojenola, 2010).

With regard to **Spanish** and **Catalan** languages, which are the languages of this proposal, their presence in the parsing field is not as strong as it is for the English. Some linguistic resources related to syntax are available in these languages. The first treebank for both languages is the AnCora Corpus (Taulé et al., 2008), which contains 500,000 tokens for each language. There is another syntactically annotated corpus for Spanish, the IULA Spanish LSP Treebank, which is the largest treebank with 590,000 tokens (Arias et al., 2014). In sections §6.3.1.1 and §6.3.1.2, both resources will be described in detail. Recently, another treebank for Spanish has been released, the AnCora-UPF, which contain 100,892 tokens annotated with deep syntactic dependencies (Mille et al., 2013).

Because large treebanks are available in both languages, there are several statistical parsers available. In particular, Spanish and Catalan were two of the languages present in the CoNLL Shared task about multilingual data-driven parsing, Spanish in 2006 (Buchholz and Marsi, 2006), Catalan in 2007 (Nivre et al., 2007) and both languages in 2009 (Hajič et al., 2009). In addition to the parsers of the CoNLL contest, Ballesteros et al. (2014) propose a dependency parser with deep syntactic structures for Spanish, and Agerri et al. (2014) develop a Spanish shift-reduce parser for IXA-Pipeline which is based on a maximum entropy algorithm available in Apache OpenNLP API[1].

On the other hand, there are also some contributions to rule-based parsing, which are much less frequent in Catalan than in Spanish and which have been decreasing year after year.

For the Catalan language, the only robust grammar which we are aware of is CATCG (Alsina et al., 2002). It is essentially a rule-based grammar based on the constraint grammar (Karlsson, 1990) and formed by 227 rules for shallow parsing and 1387 rules for deep parsing, which provides shallow syntactic parse trees and some deep syntactic parse trees when the deep syntactic rules can disambiguate the case.

With regard to the grammars available for Spanish, several frameworks approaches have been used: constraint-based dependency grammars (e.g. HISPAL and Connexor), context-free rules integrated in dependency grammars (e.g. DILUCT and Compression Rules Dependency Parser), or grammars based on unification (e.g. Slot Unification Parser and Spanish Resource Grammar). They are summarized in the following lines.

HISPAL This parser for Spanish was developed jointly with 22 other languages by the Institute of Language and Communication at the University of Southern Denmark (Bick, 2006). It uses a manually defined grammar that follows the framework of the constraint grammar (Karlsson, 1990). Basically, the Spanish rules were imported from the Portuguese grammar and some language specific rules were added. Syntax is handled in two modules in HISPAL. The first module is a set of constrain rules that assign shallow syntactic structures for the non-ambiguous structures from the output of the morphological module of HISPAL. After the shallow syntactic module, another set of constraint rules using semantic-syntactic information are applied to disambiguate the structures and to provide a deep syntactic analysis.

Connexor This parser is one of the NLP tools of the Finnish company Connexor Oy (`www.connexor.com`). This tool implements the functional dependency grammar introduced by Tapanainen and Järvinen (1997) and the constraint grammar (Karlsson, 1990). The parser is built on a single module that performs the syntactic disambiguation of morphologically tagged data by constraint rules that use syntactic and semantic information.

[1] `https://opennlp.apache.org/`

DILUCT This is a robust dependency parser that uses a set of hand-written rules with statistics of lexical attraction words for disambiguating certain phenomena (Calvo and Gelbukh, 2006). The grammar rules operate over lemmatized text morphologically annotated and disambiguated, they define the governor given two words and the words which a syntactic governor has been assigned to are labelled with a syntactic relation. Specifically, DILUCT contains a module for the PP-attachment disambiguation based on the words co-occurrence statistics following the method of lexical attraction of Yuret (1998), which is applied after the parser has finished the processing.

DepPattern Gamallo (2015) proposes a new finite-state method for dependency parsing using compression parsing. This strategy is similar to the approach followed by the shift-reduce parsers since the compressing rules remove the dependent node once the dependency relation is created. Consequently, the input is reduced progressively every time a new rule is applied, and the processing of the parser is simplified when finding new dependencies. In addition, the rules have a set of operations to inherit properties of the dependent that has been removed, to add new features to the preserved nodes and to correct morphological errors. In order to implement successfully the compression method, the grammar solves the easiest phenomena first and the most complex ones are processed last.

Slot Unification Parser It was designed at the University of Alicante (Ferrández and Moreno, 2000). This parser is based on the Slot Unification Grammar, which is an extension of the Definite Clause Grammar (Pereira and Warren, 1986), and the grammar is a set of phrase structure rules following the slot unification formalism. The rules are translated to Prolog, so that the parser can process them given a input sentence. The output of the parser contains morphological, syntactic and semantic information of every constituent. This information is used in the subsequent module that solves linguistic problems. Finally, a module of syntactic analysis builds a semantic interpretation of the information processed.

Spanish Resource Grammar It is a broad-coverage manually developed HPSG grammar (Marimón, 2010). The grammar needs linguistic pre-processed data and it uses the FreeLing NLP library (Padró et al., 2010). The grammar works with three components: inflectional rules which provide a morphological analysis of the input sentence and perform NE recognition and classification, the lexicon (with the lexical entries of the grammar containing semantic relations) and syntactic rules (phrase structure rules that combine the tokens of the input sentence).

Recapitulation

A general overview of the main trends in parsing was presented in this chapter. Specifically, parsing has been described from three main axes. Firstly, from the theoretical frameworks perspective, the linguistic theories most frequently implemented in parsing (grammars following the constituency approach, the dependency grammar and the unification grammars) have been described (§2.1.1, §2.1.2, §2.1.3), as well as specific grammars for each one of these approaches (e.g. Phrase Structure Grammar, Meaning-Text Theory, Constraint Dependency Grammar, Head-Driven Phrase Structure Grammar). Furthermore, the suitability of these frameworks for parsing was discussed with particular examples (§2.1.4). From the methodological point of view, the main strategies regarding these grammars have been reviewed by the way the parser of representing the information, also the projection method (§2.2.1) and deterministic strategies (§2.2.2) were described. After this explanation, an exposition about the approaches based on the kind of knowledge that the parsers manages was performed. It was restricted to the rule-based approaches (§2.2.3), context-free rules and constraint rules, and statistical approaches (§2.2.4), unsupervised and supervised. The last part of this chapter presented parsing with regard to language diversity. The unbalanced distribution among languages and, in particular, the status of the Spanish and Catalan languages has been presented in detail (§2.3).

Chapter 3

Natural Language Ambiguity in Parsing

Parsing is one of the Natural Language Processing (NLP) areas that has advanced most over recent years. However, significant efforts are required to increase the performance of current parsers. In fact, the main stopping issue which parsers have to overcome is ambiguity. These programs make use of a very shallow knowledge of the world and of language. Without any kind of knowledge they can only interpret the sentences literally. Therefore, in order to succeed in the interpretation of ambiguous words, phrases or sentences they need to process information with deep and robust knowledge.

The majority of linguistic phenomena can be interpreted relatively successfully by a robust parser. However, some specific linguistic phenomena are ambiguous for parsers. As a consequence of this complexity, syntactic analyses that parsers generate can contain a large number of errors when ambiguous phenomena are processed. In particular, parsers show poor performance in prepositional phrase attachment (i.e. the assignment of the right syntactic head to the prepositional phrase), in argument recognition (i.e. the recognition of the arguments of the verb predicate and the assignment of the right syntactic relations to the verb's dependants), and in coordinating constructions (i.e. the identification of the set of elements participating in the coordination). Although these three phenomena correspond to very concrete problems, the knowledge that the parser has to process is complex and sometimes difficult to manage.

In this chapter, these issues will be the focus of attention. In order to understand the main problematic linguistic phenomena, ambiguity in natural language will be first introduced and it will be explained from the point of view of parsing (§3.1). After establishing the main issues that ambiguity causes in parsing, two major ambiguous linguistic phenomena will be discussed, the prepositional phrase attachment (§3.2) and the argument recognition (§3.3). In both sections, these phenomena will be described from the linguistic point of view in order to explain next how they affect the performance of parsers. Finally, several works about both problems will be reviewed explaining their main contribution, and the methods, algorithms and resources that they have developed. Concerning the coordinating construction, it will be discussed when presenting the syntactic criteria for the FreeLing Dependency Grammars (§5.3).

3.1 Natural Language Ambiguity

Ambiguity is one of the inherent properties of natural language. So that, words, phrases, or sentences may contain more than one meaning with relative frequency (Gillon, 1990; Sennet, 2016). For example, *my colleague whom I was writing an article got up from his desk last Friday afternoon and he said*:

(1) My work here is finished

In that scenario, he could refer to several situations such as:

1. He had finished his contribution in the paper because the article could not be better.
2. He was done with the whole work week and he was going for the weekend, although the paper was not finished yet.
3. He was so exhausted from the project that he was quitting without caring if the paper was ready to submit or not.

As shown in the example (1), ambiguity is meaning-motivated but it is concerned with two major types: **lexical ambiguity** and **structural ambiguity** (Sennet, 2016). If a word has more than one possible interpretation, it is said to be lexically ambiguous (2). Alternatively, a sentence can contain more than one underling structure, then, the sentence is syntactically ambiguous (3).

(2) Safety experts say school bus passengers should be belted

(3) British left waffles on Falklands

The example (2) contains a lexical ambiguity. The two meanings of the sentence rely on the double interpretation of the verb 'to belt'. It can be interpreted as the action of securing the passengers with a seatbelt or the action of hitting the passengers with a belt.

On the other hand, the example (3) has two possible structures that convey different meanings, so it is a case of structural ambiguity. The ambiguity exists between the fact that the British left-wing cannot make a decision about what to do with the Falklands ('left' is an adjective and 'waffles' is a verb), and the fact that the British took some waffle pastries to the Falklands and then they went away ('left' is a verb and 'waffles' is a noun).

In the majority of occasions, the context of the sentence or knowledge of the world makes it possible to select the most probable interpretation, i.e., to **disambiguate** the several possible meanings (Gillon, 1990). For example, the most natural interpretation of (2) is the meaning about the passengers safety. In the case of (3), the most expected interpretation refers to the meaning about the British left-wing waffling. Despite this, sometimes the selection of the most probable interpretation is not possible because pieces of information are lacking, such as some cases of anaphora in which it is not clear which antecedent the pronoun points to (4).

(4) After their father removed the trash from the pool, the kids played in it

For example, in (4) it is not possible to disambiguate if the kids played in the trash or in the pool, since the pronoun ‘it’ may refer indistinctly to either of both entities without providing any hint about its true antecedent.

In terms of Natural Language Processing (NLP), ambiguity has a much broader sense, which makes it more complex to solve than in human language. From this point of view, any NLP application process by default the information literally and almost without knowledge of language (Jurafsky and Martin, 2000). For this reason, any natural language string (not only natural language ambiguous strings) is ambiguous by NLP tools, unless some deep knowledge about language is provided. Therefore, NLP tasks are oriented to solve ambiguity by means of adding layers of linguistic knowledge (Jurafsky and Martin, 2000).

(5) Sony buys Michael Jackson’s stake in lucrative music catalog

When processing a sentence like (5), the majority of NLP applications need to know that the sentence is formed by string of nine words as follows: (1) *Sony*, (2) *buys*, (3) *Michael Jackson*, (4) *’s*, (5) *stake*, (6) *in*, (7) *lucrative*, (8) *music*, and (9) *catalog*. Once the words are recognized, the program needs to know the grammatical category of every word and to deal with the words that accept more than one interpretation. For example, in (5), the morphological analysis should indicate that ‘stake’ and ‘catalog’ can be interpreted as verbs as well as nouns.

Furthermore, if the word forms are generalized by their lemma in order to simplify the subsequent automatic language analysis tasks, it will pop up that ‘catalog’ is an orthographic variant of ‘catalogue’. Frequently, it is useful to classify the named entities mentioned in the sentence (e.g. *Sony* as a company proper and *Michael Jackson* as a person proper name).

Finally, a syntactic analysis layer can decide how the words are grouped into phrases and how these phrases are combined between them forming the sentence structure. At this level, the system has to deal with ambiguities such as the double attachment of the sequence ‘in lucrative music catalog’ to the main verb ‘to buy’ or to the previous noun ‘stake’.

Therefore, the methods, algorithms and resources integrated in the NLP tools are mechanisms to represent knowledge about the world and about language, so that they provide the information that the application needs in order to choose the most appropriate interpretation among all the possible interpretations (Jurafsky and Martin, 2000).

Parsing is not exempt from ambiguity among the other applications of NLP (Carroll, 2003), as observed in (5). In general, nowadays parsers can deal with ambiguity quite successfully and their performance can be considered highly accurate. Despite the huge achievements in the area of parsing, some ambiguities strongly remain because of the complexity of the linguistic phenomena involved. Specifically, the scientific community has been extensively working on proposing solutions for the prepositional phrase attachment (Ratnaparkhi et al., 1994; Stetina and Nagao, 1997; Pantel and Lin, 2000; Belinkov et al., 2014) and the argument recognition problem (Carroll et al., 1998; Zeman, 2002). However, they are issues to be resolved as it will be exposed in the following sections §3.2 and §3.3 respectively.

3.2 Prepositional Phrase Attachment

The prepositional phrase attachment (PP-attachment) is a linguistic phenomenon in which parsers have a lot of problems for generating a right analysis. For more than two decades, the disambiguation of the PP-attachment has been the focus of many researchers working in parsing as, well as in word sense disambiguation and in lexical meaning representation. The explosion of unsupervised methods like distributional semantics and word embeddings in NLP offers new ways of disambiguating the PP-attachment and shows improved solutions to the previous supervised and unsupervised works.

In this section, the concept of PP-attachment will be delimited by explaining the ambiguities related to the prepositional phrase (PP) from a linguistic point of view, and describing the problems that these ambiguities cause to the parsers when trying to provide an accurate analysis of a sentence containing a PP (§3.2.1). Next, a selection of works about PP-disambiguation will be presented (§3.2.2). This part focuses almost exclusively on the proposals based on learning tasks since the PP-attachment disambiguation has been addressed from a statistical point of view. For this reason, the main statistical approaches and learning algorithms applied in the disambiguation of the PP-attachment will be presented.

3.2.1 Definition of the problem

The prepositional phrase may appear complementing or modifying the majority of syntactic heads. That is, in verb phrases 'I think *about buying tickets for the jazz festival*', in noun phrases like 'I got addicted to the tea *of the month*', in adjective phrases 'I am happy *for your achievements*', or modifying adverbs 'He didn't answer at all convincingly'.

The **wide range of possible attachments of the prepositional phrase** can explain why PP-attachment ambiguities are very common, happen frequently and are difficult to solve. The following sentence (6) adapted from the famous Chomsky (1957) example has several interpretations.

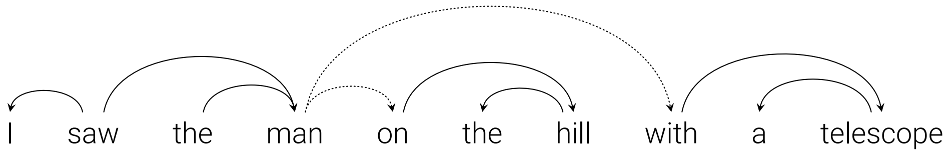
(6) I saw the man *on the hill with a telescope*

In particular, the sentence illustrated in the previous example (6) has five possible meanings:

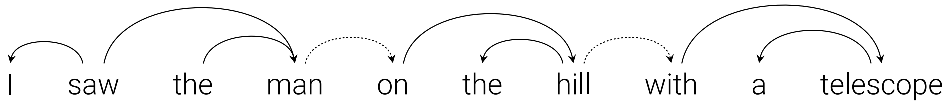
1. I saw the man. The man was on the hill. The man had a telescope. (7-a)
2. I saw the man. The man was on the hill. The hill had a telescope. (7-b)
3. I saw the man. The man was on the hill. I saw him using a telescope. (7-c)
4. I saw the man. I was on the hill. The hill had a telescope. (7-d)
5. I saw the man. I was on the hill. I saw him using a telescope. (7-e)

Every meaning of the sentence in (6) is conveyed by different syntactic structures concerning the PP-attachment, as showed in (7).

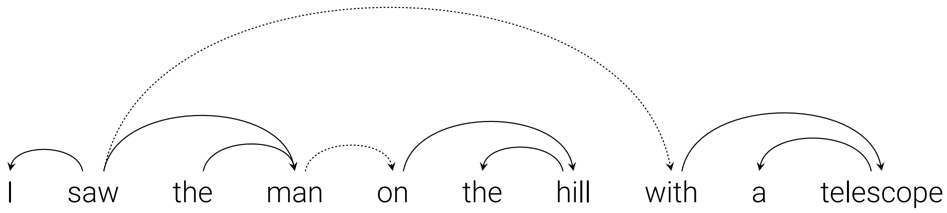
(7) a. I saw the man. The man was on the hill. The man had a telescope.



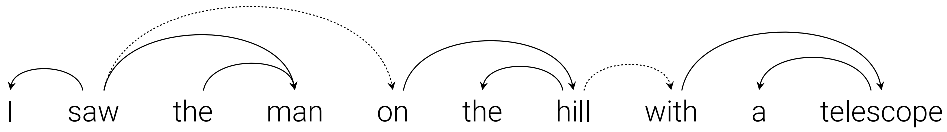
b. I saw the man. The man was on the hill. The hill had a telescope.



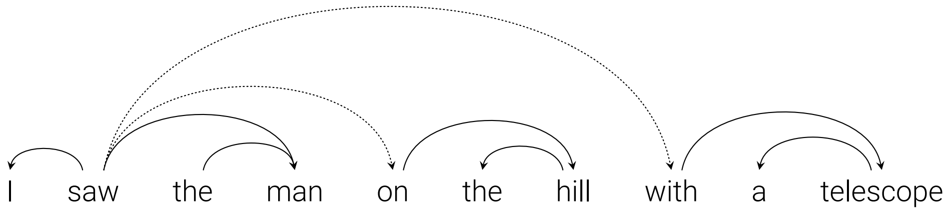
c. I saw the man. The man was on the hill. I saw him using a telescope.



d. I saw the man. I was on the hill. The hill had a telescope.



e. I saw the man. I was on the hill. I saw him using a telescope.



In order to solve the parsing problem, the parser can retrieve all the possible analyses (non-deterministic approach) or a single analysis (deterministic approach) of the ambiguous sentence (6), as explained in §2.2.2. In any case, if the sentence or sentences retrieved are among the possible analyses in (7), the parsing problem can be considered solved.

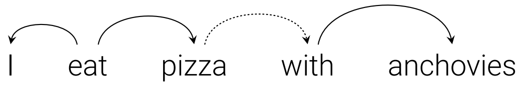
On the other hand, there are times when a single interpretation is possible. In these occasions, the multiple options of attachment observed in (6) are reduced to one. Consequently, the parser has to take the right decision on nesting the PP on the right syntactic head. Even two almost identical sentences formed by the same sequence of phrases may have different structures which the parser has to capture, such as the examples proposed by McLauchlan (2001) and exposed in (8).

(8) a. I ate pizza with anchovies
b. I ate pizza with friends

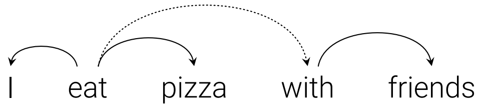
Both sentences (8-a) and (8-b) are formed by the same sequence *VP NP*$_1$ *P NP*$_2$. The only difference is the noun inside of the PP ('anchovies' and 'friends', respectively). The use of a different lexical unit involves a change of the meaning of the sentence because the semantic properties of both lexical units are also different. While in (8-a) the PP introduces one of the ingredients which the pizza that I ate is made of ('anchovies'), in (8-b) the PP expresses the fact that my friends and I ate pizza together.

As a consequence of the semantic differences, the PP modifies or complements a different syntactic head in each sentence, so the attachment happens in a different level of the syntactic tree in every sentence (9).

(9) a. I eat pizza with anchovies



b. I eat pizza with friends



In (9-a), since 'pizza' restricts the semantic interpretation of the PP, which refers to the ingredient which the pizza is made of. Consequently, the PP 'with anchovies' modifies 'pizza' in the previous NP and, hence, it is attached to the previous NP (**n-attached solution**). On the other hand, in (9-b), the noun 'friends' describes a particular aspect of the action expressed in the verb 'to eat', so the PP 'with friends' goes attached to the verb of the predicate (**v-attached solution**).

By default, a parser interprets both examples of (9) as identical cases because they are formed by the same sequence of phrases. For this reason, it retrieves the same solution for both sentences. In the deterministic parsing approach, the parser chooses the same solution for both sentences between a tree with the n-attached solution (9-a) or a tree with the v-attached solution (9-b). Whereas in the non-deterministic approach, the parser retrieves both trees with the n-attached solution (9-a) and the v-attached solution (9-b) for both sentences without taking a decision about the right solution in every sentence.

Therefore, the parser has limited resources to take the right decision by itself in (9), unless some kind of knowledge describing the PP-attachment options is integrated. Specifically, this is the strategy followed in parsing to improve the PP-attachment disambiguation. In the following subsection (§3.2.2), an overview about how this issue has been handled in NLP will be presented, describing the main proposals from the several approaches implemented in the area and observing the main problems pendent to be solved.

3.2.2 Learning the attachment of the PP

The problem of the PP-attachment disambiguation has been considered by many researchers and from several points of view since more than two decades (from late 1980s (Altmann and Steedman, 1988) to nowadays (Belinkov et al., 2014)). This contributions have made it possible to define in detail the scope of the problem (§3.2.1). Despite the abundant literature on this topic, nowadays parsers have still difficulties to reach an acceptable accuracy on assigning the right attachment to the PPs (Kummerfeld et al., 2012). For this reason, it persists the interest in exploring new methods and exploding the big amount of new large databases that are appearing.

The PP-attachment disambiguation has been addressed as an automatic task to classify the PP (formed by P and N_2) in the quadruple *VP* NP_1 *P* NP_2 as a child of the NP_1 (n-attached solution) or as a child of the *VP* (v-attached solution). Specifically, this task has been developed over the combination of four methodological axis, which refer to the disambiguation purpose, the scope of the classification, the amount of supervision and the level of linguistic representation.

Most of the studies agree that finding the right attachment of the PP is a crucial step in order to improve the parsers accuracy. However, they differ on the **disambiguation purpose**. There are proposals focused on the disambiguation process itself (Ratnaparkhi et al., 1994; Stetina and Nagao, 1997; Pantel and Lin, 2000; Olteanu and Moldovan, 2005; Merlo and Ferrer, 2006; Belinkov et al., 2014), which can be named as **isolated approach** in terms of Šuster (2012). Whereas, other proposals carry out the disambiguation task being aware that the classification results are integrated directly into the parsing task (Foth and Menzel, 2006; Agirre et al., 2008; Henestroza and Candito, 2011). These proposals, then, take a **parsing-aware approach** (Šuster, 2012).

Both approaches ensure a satisfactory accuracy on PP-attachment disambiguation: the best score is set at 92.85% in the isolated approach (Olteanu and Moldovan, 2005) and at 87.7% in the parsing-aware approach (Belinkov et al., 2014). However, Foth and Menzel (2006) argue that the isolated approach is not a real method to measure the improvement of the PP-attachment disambiguation for parsing purposes since the classification results may not be relevant for parsing.

From the point of view of the **classification scope**, the classification of the PP-attachment can be performed according to a binary classification or a non-binary classification. The PP-attachment can be seen as a 2-way classification problem (to the main verb or to the previous noun), so the disambiguation is performed over a **binary classification**.

In particular, two tendencies are present among the binary classification proposals. While Hindle and Rooth (1993) consider that the binary classification is performed over the triplet *VP NP PP*, the most accepted tendency establishes a disambiguation based on the quadruple *VP* NP_1 *P* NP_2 (Brill and Resnik, 1994; Ratnaparkhi et al., 1994; Collins and Brooks, 1995; Stetina and Nagao, 1997; Ratnaparkhi, 1998; Pantel and Lin, 2000; Olteanu and Moldovan, 2005; Agirre et al., 2008; Gala and Lafourcade, 2006; Šuster, 2012; Belinkov et al., 2014). The success of the classification by quadruples instead of triplets shows that the majority of authors assume that the properties of the N_2 are relevant for the disambiguation.

On the other hand, Merlo and Ferrer (2006) opt for a **non-binary classification** based on 4-way classification problem. These authors argue that the disambiguation of the PP depends on the notions of attachment (n-attached or v-attached) and argumenthood (argument or adjunct).

When reviewing the several methods for the knowledge representation about the PP-attachment, statistical approaches are used almost exclusively and the rule-based approach is almost non-existent (Brill and Resnik, 1994). Consequently, the works on this topic focus on learning statistically the patterns of PP-attachment.

The learning task is restricted to the **amount of supervision**, which can be defined as supervised or unsupervised. Broadly speaking, the **supervised approach** makes use of automatically or manually annotated data containing information about the PP-attachment decisions to train a classifier (Hindle and Rooth, 1993; Ratnaparkhi et al., 1994; Collins and Brooks, 1995; Stetina and Nagao, 1997; Olteanu and Moldovan, 2005; Merlo and Ferrer, 2006; Agirre et al., 2008). Otherwishe, the classifier uses an **unsupervised approach** in which the training data (raw or annotated) do not provide any information about the PP-attachment disambiguation (Ratnaparkhi, 1998; Pantel and Lin, 2000; Šuster, 2012; Belinkov et al., 2014). Furthermore, there are proposals, such as (Gala and Lafourcade, 2006), that mix both approaches, so the learning is dealt with by a **semi-supervised approach**.

Among the supervised works, the state-of-the-art algorithm is stated at 92.85% by the system of Olteanu and Moldovan (2005). Whereas, the best unsupervised classifier is the system developed by Belinkov et al. (2014) which scores 88.7%. In the table 3.1, several works on PP-attachment disambiguation are ranked according to the amount of supervision or to the rule-based approach.

Classifier	**Approach**	**Precision (%)**
Gala and Lafourcade (2006)	semi-supervised	n/a
Merlo and Ferrer (2006)	supervised	72.0
Hindle and Rooth (1993)	supervised	79.9
Ratnaparkhi et al. (1994)	supervised	81.6
Brill and Resnik (1994)	rule-based	81.8
Ratnaparkhi (1998)	unsupervised	81.9
Agirre et al. (2008)	supervised	83.6
Pantel and Lin (2000)	unsupervised	84.3
Collins and Brooks (1995)	supervised	84.5
Šuster (2012)	unsupervised	87.26
Stetina and Nagao (1997)	supervised	88.1
Belinkov et al. (2014)	**unsupervised**	**88.7**
Olteanu and Moldovan (2005)	**supervised**	**92.85**

TABLE 3.1: Precision scores of PP-attachment classifiers (from different data sets)

In both approaches, the **learning methods** applied are very diverse and can be used independently from the supervision approach. Some examples of these methods are the maximum-entropy model of Ratnaparkhi et al. (1994), the backed-off model that Collins and Brooks (1995) propose, the induction based-on decision trees (Stetina and Nagao, 1997), the classifier based on support vector machines (Olteanu and Moldovan, 2005), and the work on word sense dismabiguation of Agirre et al. (2008).

The majority of authors make use of the information about the context. In particular, they develop the classifiers algorithm with cooccurrence frequencies of the linguistic pieces related to the PP-attachment. The work of Hindle and Rooth (1993) proposes a distributional approach for resolving the PP-attachment by computing the selectional preferences with the association measure of likelihood. Ratnaparkhi (1998) proposes a disambiguation system using a set of heuristics on cooccurrence frequencies of right PP-attachments.

Actually, Pantel and Lin (2000) are the authors who introduce the method of distributional semantics (presented in §7) in the PP-attachment disambiguation. Their algorithm looks for contextually similar words by intersecting similar words and the words that occur in the same context given a dependency relation. Similarly, Gala and Lafourcade (2006) assign cooccurrence probabilities over the training data and they establish similarity between words by calculating the cosine similarity of the contextual word vectors. Compared to this previous work, Šuster (2012) applies the same kind of distributional algorithm but integrated in parsing. Recently, because of the success of the word embedding approach, Belinkov et al. (2014) implements a recurrent neural network to process vector representations for every word of the training corpus.

Concerning the **level of linguistic representation**, several classifiers supervised as well as unsupervised work only with **lexical information** to create transformation rules (Brill and Resnik, 1994), to be applied in a backed-off model (Collins and Brooks, 1995), to establish the selectional preferences (Hindle and Rooth, 1993) or the distributional word similarities (Gala and Lafourcade, 2006; Šuster, 2012). Some other authors combine the lexical information with lexical classes based on mutual information of the word forms (Ratnaparkhi et al., 1994), or with a thesaurus which provide sets of similar words given a word (Pantel and Lin, 2000).

On the other hand, **syntactic information** has been used to learn the PP-attachment patterns, such as the use of PennTreebank (Marcus et al., 1993) in the work of Merlo and Ferrer (2006), or the use of a chunker (Ratnaparkhi, 1998). Some authors like Stetina and Nagao (1997) and Agirre et al. (2008) show that **semantic representations** are the source for successful PP-attachment disambiguation by using algorithms for word sense disambiguation over the synsets of WordNet (Fellbaum, 1998). Finally, classifiers combining both **syntactic and semantic levels** are also present in the literature, such as the proposal of Olteanu and Moldovan (2005) based on the PennTreebank and FrameNet (Baker et al., 1998), and the work of Belinkov et al. (2014) which proposes to add subcategorization information from VerbNet (Kipper, 2005) and semantic information about the hypernyms from WordNet.

In the case of the language languages of interest of our proposal, Ratnaparkhi (1998) proposes a disambiguation system for the Spanish language based on heuristics. On the other hand, Calvo and Gelbukh (2006) develops a classifier for the PP-attachment disambiguation in Spanish from word co-occurrence frequencies following with lexical attraction measures applied (Yuret, 1998). For the Catalan language, there is no literature about the PP-attachment disambiguation.

3.3 Argument Recognition

Nowadays parsers become sophisticated software able to perform high-accuracy syntactic analysis. However, they still have difficulties in assigning the right syntactic relations to the tree arcs because they still show poor-performance in recognizing arguments of the verb predicate. In order to overcome this limitation, parsing turns to the acquisition of verb subcategorization frames because the information about the semantic and syntactic properties of the verb predicate may help the accuracy of the parser (Manning, 1993; Carroll et al., 1998; Zeman, 2002).

In this section, the aspects around the issues on argument recognition will be exposed. Firstly, a description of the main problems that parsers have in recognizing the arguments of a verb will be exposed (§3.3.1). The second part of the section is dedicated to the task about acquiring information about verb subcategorization by means of explaining the main methods, resources and theoretical proposals that have been developed (§3.3.2).

3.3.1 Definition of the problem

In order to explain the problems related to the argument recognition in parsing, it is necessary to go back to the linguistic theory and examine the notion of argument.

In general terms, predicate logic supports the idea that the sentence is formed over a **predicate** (*P*) and one ore more **arguments** (*x*). The predicate is seen as a function which relates the arguments mentioned in the sentence. On the other side, the arguments are the set of entities mentioned in order to help completing the meaning of the predicate. This can be formally formulated like a propositional function *P(x)* which is evaluated in terms of truth. *P(x)* is true iff *x* belongs to the set denoted by *P*. Given an example like (10-a), a prepositional function ('read') and two arguments ('I' and 'paper') can be identified and when combined they produce the proposition (10-b).

(10) a. I read a new paper about parsing
b. read(I,paper)

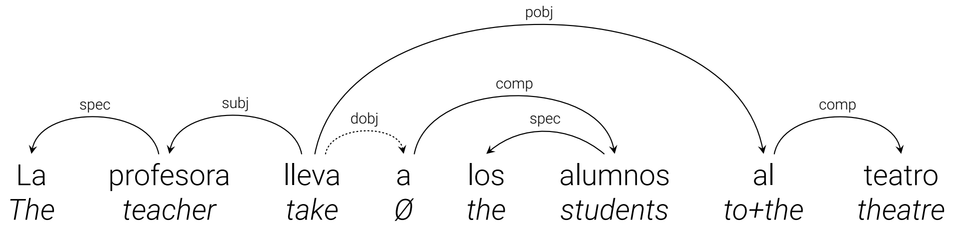
According to the properties of the predicate, every predicate controls a specific number and types of arguments. There are also some other elements named adjuncts that add an optional value to the predicate, so their occurrence is optional.

The syntactic realization of the predicate and the set of syntactic arguments that the predicate controls is known as **subcategorization** in the generative tradition (Chomsky, 1965) or valence in the structural tradition (Tesnière, 1959). Both concepts are referring almost to the same notion. However, while the subcategorization notion involves considering the subject out of subcategorization frame domain, the conceptualization of the predicate according to Tesnière (1959) requires the subject to be an actant (i.e. an argument in Tesnière terms) in the same way that the rest of arguments of the predicate. Despite this, currently subcategorization and valency are used as almost synonyms. In this study, when talking about subcategorization, it is concerned to this modern conceptualization, so the subject is considered part of the subcategorization frame.

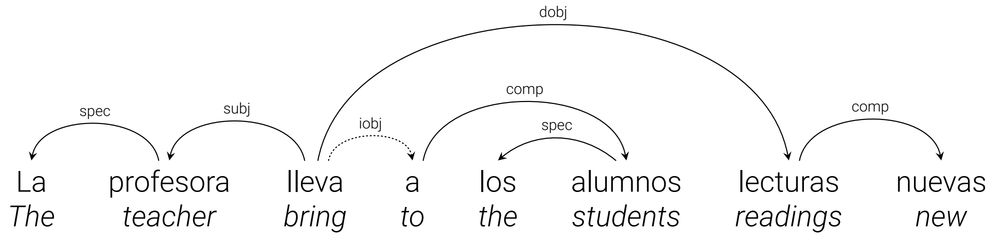
Returning to the example above (10-a), the predicate associated to the verb 'to read' is biargumental because two arguments are expected in its argument structure (i.e. 'I' in the argument 1 and 'paper' in the argument 2). Besides this subcategorization frame, there also are monoargumental (i.e. one argument), biargumental (i.e. two arguments), triargumental (i.e. three arguments) and quadriargumental frames (i.e. four arguments). In addition, there are verbs which accept alternations without any argument such as the verb 'ploure' ('to rain') in the sentence 'Plourà' ('It will rain').

Subcategorization is essentially linked to meaning. The meaning is conveyed in the predicate structure by its arguments and their properties. For this reason, different meanings are conveyed by different subcategorization frames (Levin, 1993). This explains the different predicate structures that polysemous lexical units can take such as in the examples (11) and (12).

(11) La profesora lleva a los alumnos al teatro



(12) La profesora lleva a los alumnos lecturas nuevas



Both examples show different meanings of the verb 'llevar', 'to take' in (11) and 'to bring' (12). Consequently, the subcategorization frame is different in each sentence. The meaning in (11) is conveyed by a triargumental frame formed by a subject ('La profesora'), a direct object ('a los alumnos') and a prepositional object ('al teatro'), whereas the meaning in (12) is expressed by a triargumental frame containing a subject ('La profesora'), a direct object ('lecturas nuevas') and indirect object ('a los alumnos').

Generally, predication is related to the verb argument structure as seen in the previous examples. Furthermore, predicate structure is present in nouns (i.e. 'The America discovery by

Columbus') and adjectives ('I am happy for you').

In terms of parsing, a parser capable of solving the parsing problem is the program that labels the arcs of the tree representation of an input sentence with the right syntactic relation. The task of assigning the right syntactic relation can be performed successfully if the arguments of the verb predicate are recognized and they are labelled with the right syntactic function.

Despite this, a parser per se has big difficulties in identifying the argument structure (Carroll et al., 1998; Zeman, 2002). For example, any parser by itself is not able to distinguish that the same verb, like 'llevar' of the examples (11) and (12), has different subcategorization frames according to both different meanings. Consequently, the analyser cannot identify that the same phrase 'a los alumnos' is a different argument with a different syntactic function in each sentence, a direct object in (11) and a indirect object in (12). However, if information about subcategorization frames is added in the parser, then it will help to raise the accuracy of the syntactic trees performed by the program (Carroll et al., 1998; Zeman, 2002).

The **integration of subcategorization for parsing** has the base on the subcategorization acquisition task, which involves to work with several disciplines such as lexicography, corpus linguistics, or machine learning, as it will be explained in the following subsection (§3.3.2).

3.3.2 Subcategorization frames acquisition

The acquisition of verb subcategorization frames is a complex task, which can be observed because of the huge variety of proposals. Several resources (machine-readable dictionaries or digitalized textual corpora), linguistic information encoded (raw, tagged and/or parsed linguistic data), theoretical principles for verb classification (e.g. lexical-semantically based, ontologically based, etc.), and methods (manual, supervised or unsupervised), can be combined to deal with this task. Due to this multiple combination of strategies, the following explanation is an overview which distinguishes the two main process in the acquisition of subcategorization frames, detection and classification.

The techniques used for the **detection** of verbs and their subcategorization frames automatically from digitalized textual corpora are very diverse and they vary depending on the kind of linguistic information encoded in the text. Since the amount of raw text available is currently unrestricted, several authors propose to add automatically layers of linguistic information in the detection.

Brent (1993) develops a system able to detect English verbs and the phrases of the arguments of these verbs only by applying a set of heuristics over raw text. Other authors (Manning, 1993; Briscoe and Carroll, 1997; Carroll et al., 1998) observe the limitations of the heuristic approach and they propose to extract frames by automatically adding layers of morphological information, first, and syntactic information over the morphological layer. Both techniques generate a noisy bunch of data that needs to be filtered to identify the arguments that belong to the predicate from the other elements of the sentence.

On the other hand, there are proposals that re-use the corpora already annotated in order to ensure the accuracy of the detection. The work of Zeman (2002) directly extracts the data in the form of syntactic subtrees from the syntactic trees of the PennTreebank (Marcus et al., 1993). Since the extraction is immediate, this approach allows to concentrate on the strategy of filtering the good candidates of subcategorization frames.

Because these proposals are based on large textual corpora, they implement statistical techniques for filtering the data by weights or probabilities. According to this filter, the data with highest weights or probabilites is more likely to be a real subcategorization frame rather than the data with lower punctuations, which can be errors or marginal subcategorization frames.

Finally, several corpora semantically and syntactically annotated populate the repertoire of available linguistic resources suitable to extract subcategorization frames. This is the case of PropBank (Palmer et al., 2005), a semi-automatically annotated corpus that assigns predicate argument structure and semantic roles to the syntactic trees nodes of the PennTreebank. Since the information for generating the subcategorization frame is in the corpus already, a framing task is performed for identifying the argument structure and the semantic roles of the verb predicate in parallel to the annotation task. The resulting resource of the framing is a computational lexicon of the subcategorization frames of the corpus (Palmer et al., 2005).

Once the subcategorization frames are successfully captured, they need to be organized in classes. The **classification** of subcategorization frames, which basically relies on the classification of verbs, is dealt with by grouping verbs according to the types of their arguments and the syntactic and semantic properties of the verb predicate.

Several linguistic theoretical works proposed different verb classifications to be applied in Linguistics or in NLP (Dowty, 1979; Verkuyl, 1989; Pustejovsky, 1991; Levin, 1993; Croft, 2008), which show the small consensus among the authors and the complexity of the topic. However, the classification developed from the theoretical linguistics perspective that had a big influence in the area is the proposal of Levin (1993) by showing that the verb itself determine its behaviour concerning the arguments that subcategorize. A clear example of the implementation of Levin (1993) in computational linguistic resources work is VerbNet (Kipper, 2005), one of the largest English verbal lexicons with semantic roles, selectional preferences of the arguments and frames which has been widely implemented in NLP tools.

There also have been verb classifications proposals created in the applied perspective such as WordNet (Fellbaum, 1998) and FrameNet (Baker et al., 1998). WordNet is a network of lexical concepts linked by semantic relations which was developed for other purposes than NLP, but researchers sitting on the area realized very soon about the power of WordNet for representing the meaning for computational purposes. This network goes far beyond the verb classification, but it also includes verbs ontologically classified. On the other hand, FrameNet is a network of semantic frames that describe the event, the relations, the objects, the participants present in the frames, and their realization by means of subcategorization.

Machine-readable dictionaries typically manually developed seem an appropriate resource to populate subcategorization frames lexicons because they contain rich and structured linguistic information about lexical entries, such as ANLT (Boguraev and Briscoe, 1987) and COMLEX (Grishman et al., 1994) for English. However, because they are developed manually, some authors (Manning, 1993; Levin, 1991) argue that these resources tend to be inconsistent and incomplete, as well as they lack of homogeneity because similar words may not be treated similarly.

Concerning automatic verb classifications, supervised (Joanis et al., 2008; Li and Brew, 2008; Ó Séaghdha and Copestake, 2008) and unsupervised (Schulte im Walde, 2006; Sun and Korhonen, 2009; Lenci, 2014) methods have been applied. Both approaches aim to capture the properties of the verbs that will make possible to classify the set of verbs detected in different categories. The difference in the unsupervised approach is the fact that the verbal class which a verb belongs to is unknown and the system has to guess it.

The final classification in a lexicon of verbs and their subcategorization frames is a computational resource for several NLP tools. In the framework which this research focuses on, the **integration** of the acquired subcategorization is orientated to the contribution towards building the syntactic tree when the parser has incomplete information to make a decision (Carroll et al., 1998).

Depending on the characteristics of the parser, subcategorization assists in this task in a different way. Subcategorization information can be used to assign a probability to every possible syntactic tree of a particular sentence, then the parser ranks the trees from most probable to less probable according to the probability assigned (Carroll et al., 1998; Zeman, 2002). In contrast, subcategorization may help to restrict the application of certain rules. Then, when the parser detects the subcategorization frame in the input sentence, it labels the syntactic tree according to the frame discarding any other possible analyses (Lin, 1998b).

Concerning the languages of this proposal, the grammars of HISPAL (Bick, 2006) in the framework of the Constraint Grammar and the SRG in the context of the Head-Driven Phrase Structure Grammar (Marimón, 2010) make use of a lexicon of verb subcategorization frames to disambiguate the possible analyses given a sentence.

Recapitulation

This chapter was focused to the phenomenon of the ambiguity since it is the major problem that prevents parsers to have a better accuracy. A general description from the linguistic point of view has been introduced in order to understand the problem from the parsing point of view next (§3.1). The rest of the chapter has been dedicated to the explanation of two important ambiguous linguistic phenomena in which parsers show low-performance. Firstly, the problems related with assigning the right head to the prepositional phrase have been explained and the main contributions about the disambiguation of the prepositional attachment have been reviewed (§3.2). In the last section of the chapter, the issues involving the recognition of predicate arguments have been described by means of defining the problem from the linguistic point of view as well as the parsing point of view, and reviewing the literature that contributed to the improvement of the argument recognition scenario (§3.3).

CHAPTER 4

METHODOLOGY

The present study pursues the aim of shedding some light on the improvement of dependency grammars accuracy by the acquisition of knowledge and its integration in parsing. In order to conduct this research, an empirical approach has been followed. In other words, the research has mainly relied on real data and experiments in order to formulate and confirm this thesis hypotheses by deducing a set of conclusions observables in the results of the experiments.

For this reason, the research of this thesis is based on three major methodological aspects that follow the principles of the method stated above.

First, **research object** used to develop the research of this thesis and to study the initial hypothesis is a Spanish and a Catalan dependency grammars developed specifically for this research (§5). The initial version of the grammars in §5 is intended to prove the hypothesis 1 (§1.3).

Second, two **experiments** about knowledge acquisition and integration in parsing are performed in order to analyse the initial hypothesis (§7 and §8). Each experiment explores a problem observed in the preliminary definition of the research basis (§1.1) which applies to different kinds of knowledge acquisition strategy and corresponds to one of the hypothesis formulated. On the one hand, statistical knowledge learning is explored §7 and is aimed to provide an answer to the hypothesis 2 (§1.3). On the other hand, linguistic knowledge acquisition is studied in §8 to validate the hypothesis 3 (§1.3).

Third, a **testing method** of the research object is designed (§6) in order to empirically evaluate and analyse the results of the object of study (§6.4) and the two experiments (§7.4 and §8.2).

These three methodological aspects determine the steps followed to conduct this research.

In order to carry out the experiments to test this research hypothesis, firstly, Spanish and Catalan dependency grammars are developed. These grammars work as a deep dependency parsing module of FreeLing NLP pipeline (Padró et al., 2010). FreeLing provides linguistic pre-processing analysis necessary to perform a dependency parsing analysis. Furthermore, this pipeline has a flexible and configurable architecture to develop a module focused on rule-based dependency parsing and to carry out the experiments planned in this research.

Both grammars are developed following a set of linguistically motivated syntactic criteria

(§5.3) as a result of a revision of the Generative Grammar (Chomsky, 1981) and Dependency Grammar (Tesnière, 1959; Mel'čuk, 1988) theoretical basis, as detailed in §5. In addition, a tagset of dependency relations is set (§5.4) to define the syntactic realization of predicate arguments following the Lexical Functional Grammar assumption about functional structure by which syntactic structure is conceived as connected structure by meaningful grammatical relations (Kaplan and Bresnan, 1995).

The development process consists in the representation of syntactic phenomena taking the non-ambiguous syntactic trees generated by the FreeLing chunker and grouping them starting with the simplest structures (e.g. attachments between phrases) and progressively dealing with more complex structures (e.g. from verb predicate structure to clause and sentence level). This process is performed using a development corpus for each language which includes 809 Spanish examples and 50 Catalan examples selected from Spanish and Catalan newspapers or added intuitively, as presented in §5.5.

The development strategy followed is empirical, iterative and incremental, and validates the performance of the new rules by a test right-wrong answers. This strategy summarizes as follows:

1. Observation of a sentence in the corpus.
2. Isolation of a linguistic phenomenon of the observed sentence.
3. Development of a rule or a set of rules to handle the observed phenomenon.
4. Automatic analysis of the corpus with the grammar containing the new rule or rules.
5. Test of observed sentence represented with the analysis build with the changes made in the grammar.

In the case that the result is correct, then, the observed phenomenon is considered that it has been completed. Consequently, the current iteration of development is concluded and shifted to a new one. On the other hand, if the result is wrong, the iteration is complemented with new sentences for the corpus containing the phenomenon observed and rules are modified or added.

Apart from development tests of the grammar, a global evaluation task is designed in order to empirically test every version of the dependency grammar. The quality of the grammar rules is measured by comparing a resource which contains the correct answers of the problem aimed to solve (i.e. dependency parse trees analysed with the correct syntactic representation) and the same resource analysed with the dependency grammars developed in this research.

In this research, the comparison of dependency trees analyses is performed by the evaluation script of the CoNLL competition 2006[1]. This script evaluates the dependency grammars using statistical metrics of accuracy (such as Labeled Attachment Score, Unlabeled Attachment Score and Label Accuracy), precision and recall. This metrics are explained in detail in §6.2.

[1] `http://ilk.uvt.nl/conll/software.html`

With regard to the evaluation data, two kind of resources are used in order to perform a global evaluation task. Evaluation usually focuses on quantitative analysis, i.e., the statistical measurement of data, and it usually puts the analysis of errors aside (McEnery and Wilson, 1996). Both evaluation approaches provide a complementary observation from different perspectives (McEnery and Wilson, 1996). Since this research aims to determine the exact performance of the dependency grammars, both approaches are used, as described in §6.3.

More precisely, a specific resource has been used for each approach. On the one hand, in the quantitative approach, a pre-existent large annotated corpus with syntactic dependencies has been chosen. In particular, as explained in §6.3.1, quantitative evaluation is carried out with the Tibidabo Treebank (Marimón, 2010), which contains 41,620 tokens. Because this corpus and the linguistic criteria of dependency grammars developed does not coincide, a process of criteria harmonization is carried out using heuristic rules that adapt the treebank criteria to the grammars criteria in the majority of cases (§6.3.1.2).

On the other hand, qualitative evaluation is performed with a test suite, which is another kind of linguistic evaluation database and is used specifically in the analysis of errors (§6.3.2). More precisely, due to the lack of test suites in Spanish and Catalan for evaluating dependency parsing, one of the tasks of this research is to develop a test suite for each language, ParTes (§6.3.2.1). In order to collect data about syntactic structure and dependency relations, syntactic phenomena that populate this resource is gathered from the most relevant constructions included in descriptive grammars of the language (Bosque and Demonte, 1999; Solà et al., 2002) and the several word order configurations of verb predicates found in SenSem Corpus (Vázquez and Fernández-Montraveta, 2015).

The two hypothesis that are aimed to be answered in the end of the research are tested by two experiments as explained above. Both experiments are based on the treatment of real data linguistically processed.

First experiment explores two machine learning methods to disambiguate PP-attachment, as described in §7. Firstly, a preliminary exploration of PP-attachment disambiguation is carried out with a supervised learning approach (§7.1). Language model is learned from annotated data of AnCora Corpus (Taulé et al., 2008) which is classified by the algorithms J48, JRip, Naive Bayes and Bayes Net implemented in Weka enviroment (Witten and Frank, 2005)[2]. Secondly, an unsupervised method based on word embeddings (Mikolov et al., 2013a,b) is explored to learn word distributions related to PP-attachment. In this experiment, skip-gram algorithm implemented in word2vec (Mikolov et al., 2013a,b)[3] is used to learn context-predicting vectors of Wikicorpus Spanish and Catalan sentences (Reese et al., 2010). AnCora Corpus is also used as test and gold data. Furthermore, a set of naive supervised classifiers are developed in order to compare with unsupervised models learned. These classifiers disambiguate PP-attachment by analysing training data represented with weighted vectors. Vectors are built by calculating the association measure of Pointwise Mutual Information (Church and Hanks, 1990) with the formula implemented

[2] http://www.cs.waikato.ac.nz/ml/weka/

[3] https://code.google.com/archive/p/word2vec/

by Evert (2005)[4].

The experiment of chapter §8 about improving argument recognition follows a linguistic knowledge strategy. Linguistic knowledge is acquired from robust linguistically annotated data about subcategorization. In particular, this information is extracted from SenSem Corpus (Vázquez and Fernández-Montraveta, 2015) and organized by different granularities according to the proposal of Alonso et al. (2007) and of the CompLex-VS lexicon created in this research (Lloberes et al., 2015b). In chapter §8, the details of the corpus and the classification proposals are explained.

Recapitulation

In this chapter, the main methodology followed in this thesis research has been described briefly. The main methodological principles have been presented and have been related to this research basis and hypothesis. The experiments and the tools built specifically for carrying out this research have been justified methodologically. Next, according to the methodological aspects set for this research, the steps followed to conduct this research have been explained and, specifically, the data and the tools used in every step have been defined from a general point of view. A detailed description of every development process, the experiments, the data and the tools used is provided in every chapter concerning the development and the experiments of this research.

[4] `http://collocations.de/`

Chapter 5

FreeLing Dependency Grammars

This chapter focuses on the resource developed in this proposal, a rule-based dependency grammar manually defined. This grammar has been built in two language versions, Spanish and Catalan. In particular, both grammar versions are integrated into the dependency parsing module of the library FreeLing (Padró et al., 2010).

FreeLing is an open-source library of multilingual Natural Language Processing (NLP) tools that provide linguistic analyses for written texts. It is a complete NLP pipeline built on a chain of modules that provide a general and robust linguistic analysis. Among the available tools, FreeLing offers sentence recognition, tokenization, named entity recognition, tagging, chunking, dependency parsing, word sense disambiguation, and coreference resolution. The dependency parsing module in FreeLing is one of the most important steps in the pipeline and the grammars are the core of the FreeLing rule-based dependency parser, the Txala Parser (Atserias et al., 2005).

In the first part of this chapter, the characteristics of the Txala Parser (§5.1) and the architecture of the grammars (§5) are described. The essential characteristics and a detailed explanation of the features available in the grammars will be detailed, focusing on those aspects that have major consequences in the performance of the parser such as the order in which the rules are applied.

The following section (§5.3) is dedicated to the syntactic criteria designed particularly for the grammars. The representation of controversial syntactic constructions (auxiliary, prepositional phrase, subordinate clauses and coordination) is discussed in detail explaining the positions held by the generative approach and the dependency approach that are largely present in parsing. By the examination of these two frameworks, a proposal of syntactic criteria for the representation of these constructions is offered.

Furthermore, the dependency relations labels that are created are described in section §5.4 and compared to the Universal Dependencies relations tagset. The last section (§5.5) explains the methodology followed in the development of the grammars in order to ensure good performance.

5.1 Txala Parser

Txala Parser (Atserias et al., 2005) is one of the dependency parsing tools available in FreeLing (Padró et al., 2010) and it is responsible for the automatic syntactic analysis based on rules. Txala generates robust, projective and deterministic dependency trees by always making a decision for any input sentence or any sequence contained in the input sentence. Furthermore, the parser algorithm is language independent, which makes it possible to develop grammars for more than one language.

Dependency parsing in FreeLing is one of the most important steps of processing in the automatic language analysis pipeline. This means that, once the linguistic data reaches the parsing module, it has already been interpreted by the language analysis tools previous to the parsing processing, and it contains layers of rich linguistic information automatically annotated such as in figure 5.1. Specifically, at this point, the sentences have been segmented, and each of the input sentences has been transformed into a string of tokens annotated with their correspondent lemma and part-of-speech tags and grouped into chunks in which morphologically ambiguous tokens have been disambiguated, named-entities have been recognized, and dates and numbers have been detected.

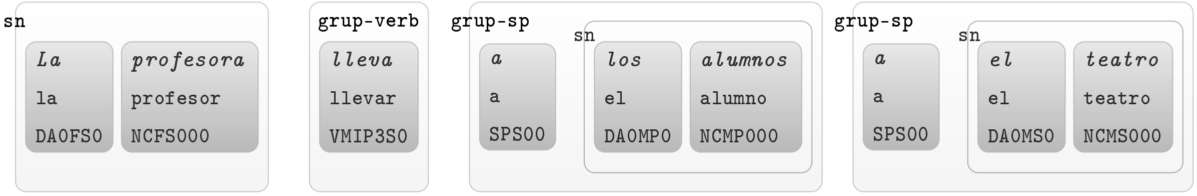


Figure 5.1: Parsing pre-processing analysis by FreeLing of the sentence ‘La profesora lleva a los alumnos al teatro’ (‘The teacher takes the students to the theater’)

The Txala Parser receives a partial syntactic constituency tree produced by the chunker Civit (2003) as input (figure 5.1) and it has to map them to a full robust dependency tree (figure 5.2). In order to map both structures successfully, three operations are performed by the parser algorithm: attachment disambiguation, dependency conversion and dependency relation assignment .

Firstly, the head-child relations are identified using a dependency grammar (i.e. **attachment disambiguation**) and following the criteria that define the syntactic structure of a dependency relation (§5.3). By this operation, the partial tree is progressively completed, providing a deeper syntactic analysis. Secondly, the structurally disambiguated tree is converted into syntactic dependencies according to the dependency principles stated by Mel’čuk (1988) presented in §2.1.2 (i.e. **dependency conversion**). Finally, each dependency arc of the tree is labelled with a tag that expresses the nature of the dependency relation between the two nodes of the arc (i.e. **dependency relation assignment**). In order to assign a label, a set of dependency relations has been defined in §5.4.

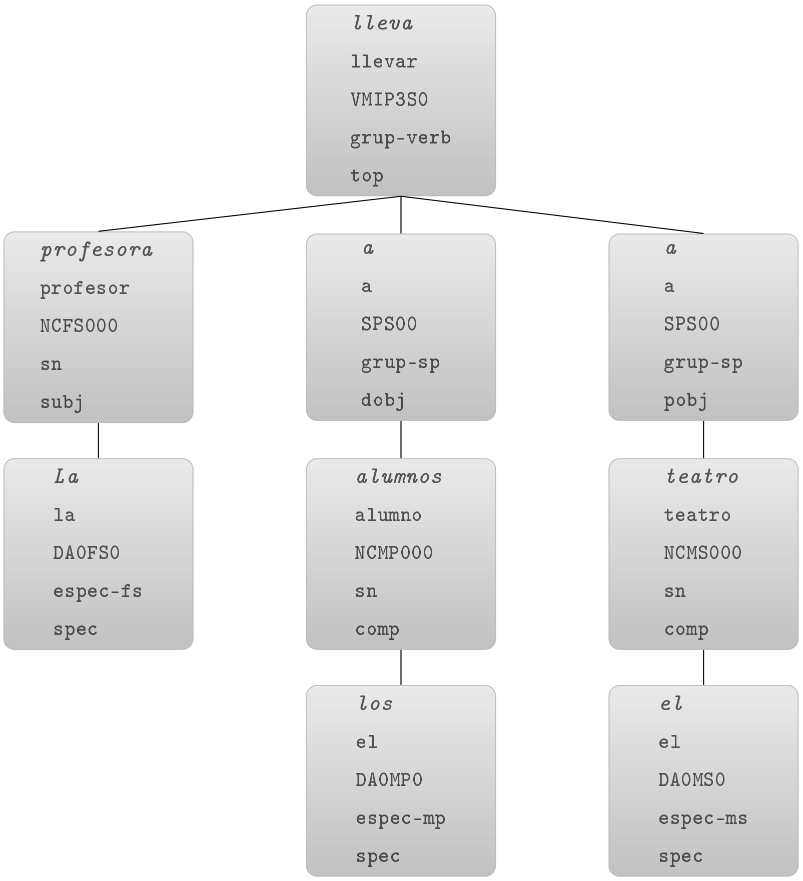


FIGURE 5.2: Dependency parse tree generated by Txala of the sentence ‘La profesora lleva a los alumnos al teatro’

5.2 DEPENDENCY GRAMMARS ARCHITECTURE

The current version of FreeLing includes rule-based dependency grammars for English, Spanish, Catalan, Galician and Asturian (see Table 5.1 for a brief overview of their sizes). However, only the Spanish and Catalan dependency grammars (Carrera et al., 2008; Lloberes et al., 2010) are described because they are the center of our proposal.

The FreeLing Dependency Grammars (FDGs) for these two languages are grammars with hand crafted rules restricted by **constraints** and **context**, and weighted with manually defined **priorities**. The grammars are structured as a set of rules of two kinds in order to deal with the attachment disambiguation (attachment rules) and the dependency relation assignment (labelling rules). Both types of rules are applied based on this principle: *at every step, two adjacent partial trees are attached or are labelled with a dependency relation tag if their rule is the highest ranked for which all the conditions are met*.

The FDGs follow the linguistic basis of syntactic dependencies (Mel’čuk, 1988). However, these grammars propose a different analysis for specific structures such as auxiliary verbs (non-

	Rules		
Language	**Total**	**Attachment**	**Relation**
English	2961	2239	722
Spanish	**2939**	**2547**	**392**
Catalan	**2608**	**2099**	**509**
Galician	178	87	91
Asturian	4438	3842	596

TABLE 5.1: Sizes of the FreeLing Dependency Grammars

finite verb headed), prepositional phrases (preposition headed), subordinate clauses (conjunction headed), relative clauses (verb headed) and coordinating structures (conjunction headed), as argued by Lloberes and Castellón (2011) and extended in the next section (§5.3).

5.2.1 Attachment Rules

The attachment rules that are held in the *<GRPAR>* section of the grammar are focused on the first operation of the Txala parser and are concerned with the disambiguation of syntactic structure. For this reason, this set of rules are placed and applied first. Basically, the operation that these rules carry out corresponds to **linking** two adjacent syntactic partial trees marking the ancestor and the descendant of the dependency. The final result of the application of the attachment rules is a structured dependency parse tree where all the nodes have been linked (i.e. any node remains out of the dependency tree) because the syntactic ambiguities of the input sentence have been resolved.

The rules are structured in four major parts: direction of the dependency, constraints, priority and flag. For instance, the rule shown in the figure 5.3 has priority *5071*, and states that a subtree marked as a verb phrase (*grup-verb*) attaches as a child to the noun phrase (*sn*) to its left (*top_left*) and whose head is an interrogative pronoun (*PT*) when these two consecutive sub-trees are located to the right of a verb phrase (*grup-verb_$$*). Once the conditions are met the parent, i.e., the noun phrase is relabelled as *subord* (i.e. subordinate clause). The application of the rule in figure 5.3 makes it possible to build the structure of the indirect questions such as the analysis in the figure 5.4.

```
5071- grup-verb_$$ (sn{^PT},grup-verb) - top_left RELABEL subord:-
```

FIGURE 5.3: Attachment rule for indirect questions

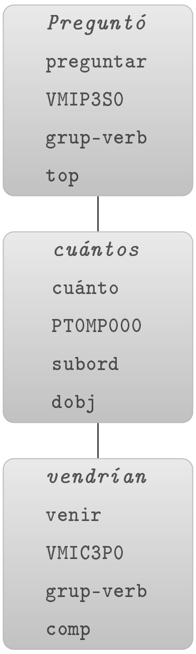


FIGURE 5.4: Dependency parse tree generated by Txala of the sentence 'Preguntó cuántos vendrían' ('$Ø_{3sg}$ asked how many $Ø_{3pl}$ would come')

The attachment is performed according to the **direction** of the dependency and four types of attachment are possible in the FDGs architecture.

Top Left
Given two adjacent partial trees, the rightmost tree it is attached to the leftmost tree. This operation is associated with the *RELABEL* operation, which assigns a new label to the parent node, to the child node or to both.

Top Right
Given two adjacent partial trees, the leftmost tree is attached to the rightmost tree. The *RELABEL* operation is also available.

Last Left
The rightmost sub-tree is attached as a child of the last node within the leftmost sub-tree. This attachment operation is combined with the *MATCHING* operation, which is used to specify the last node of the sub-tree which the sub-tree attaches to. '

Last right
The leftmost sub-tree is attached as a child of the last node within the rightmost sub-tree. This attachment option also works with the *MATCHING* operation.

The attachment rules contain four kind of **constraints** optionally activated regarding morphological (part-of-speech tag), lexical (word form, lemma), syntactic (syntactic context, syntactic properties) and semantic features (semantic properties).

Part-of-speech

It is possible to access the part-of-speech tag of the syntactic head of every adjacent sub-tree by the use of curly brackets ('{ }'). The part-of-speech tags used in FreeLing as well as in Txala are the EAGLES tags for Spanish ([1]) and for Catalan ([2]). For example, a rule defining a sub-tree defined as a noun phrase (*sn*) whose head is a common noun feminine singular (*NCFS000*) is expressed as *sn{NCFS000}*.

Word form

The word form of the syntactic head of a sub-tree is specified with parenthesis ('( )'). For example, a partial syntactic tree defined as *grup-verb(lleva)* means that the syntactic head of the verb phrase read (*grup-verb*) corresponds to the word form *lleva*, i.e., the third person singular of the present indicative of the verb *llevar*.

Lemma

It is defined by pointy brackets ('< >'). In the case of a prepositional phrase headed by the preposition *a*, the rule uses *grup-sp<a>* to express it.

Syntactic context

The left context and the right context of the two adjacent partial trees can be formally described. Some operators can be used:

Affirmation (by default)

Negation, global ('!') or local ('∼')

Matching exactly one chunk, with any label ('?')

Matching zero or more chunks with any label ('*')

Matching at least one chunk, with any label ('?_*)

Matching a sentence boundary

For example, to handle the PP-attachment to the noun such as *(sn,grup-sp)* in contexts other than post-verbal, all the attachments of the PP are allowed, except when this group is preceded by verb phrase (*!grup-verb*).

Syntactic and semantic features

Sometimes syntactic or semantic properties of the syntactic heads are useful to determine their attachment. It is possible to handle this by integrating external modules in the grammar hold in the section *<CLASS>*. The modules are configured as lists of lemmas grouped by syntactic or semantic criteria. This information can be accessed in the rules using square brackets ('[ ]'). For example, this option can be useful to detect apposition declaring that given two adjacent noun phrases the rightmost attaches to the leftmost when the lemma of the rightmost lemma is present in a list of person proper nouns previously defined in

[1] `https://talp-upc.gitbooks.io/freeling-user-manual/content/tagsets/tagset-es.html`

[2] `https://talp-upc.gitbooks.io/freeling-user-manual/content/tagsets/tagset-ca.html`

<CLASS> (*sn,sn[person]*). In addition to access to information about single lemmas, the grammars also allow to access to information about paired lemmas grouped by their syntactic or semantic features. This information is integrated in the grammar section *<PAIRS>* and declared in the rules. For example, this method can be used to handle PP-attachment with semantic preferences by assigning a nominal head (*sn*) to the PP (*sp-de*) when the head and the child inside of the PP are paired in a list of lemmas (*noun-madeof*) defined in *<PAIRS>* (figure 5.5).

```
101 - - (sn,sp-de) noun-madeof::(L.lemma,R:sn.lemma) top_left RELABEL -:-
```

FIGURE 5.5: Attachment rule using information about paired lemmas

The **priorities** are manually defined and follow a system of priorities linguistically motivated and developed specifically for the FDGs (table 5.2). This system establishes the order in which the rules are applied according to the depth of the dependency tree: 0 being the most restrictive value (i.e. high priority) and 9999 the least restrictive value (i.e. low priority).

Specifically, it determines that the phrases are solved first (i.e. they are the most restrictive rules), then the non-finite clauses, the finite clauses and, finally, the structure of the verb predicate is built which is the less restrictive structure. Coordinating structures priorities are assigned transversally. Since coordination is a very complex linguistic phenomenon which can appear at any level of the syntactic structure, it is dealt with after every block of priorities (e.g. coordinated phrases priorities immediately follow phrase priorities, coordinated non-finite clauses priorities are set after non-finite clauses, etc.).

The **flags** allow the activation of rules in certain moments of the processing of the grammar. They are defined as a list of strings separated by vertical bars ('|') in which every string is a flag name. If any flag is defined, the rule is always active. Otherwise, it is operative when the rule is toggled on.

```
5071 INIT|PH1 grup-verb_$$ (sn{^PT},grup-verb) -
     top_left RELABEL subord:- +PH2 -INIT -PH1
```

FIGURE 5.6: Flags in attachment rule for indirect questions

For example, in the previous rule for the formation of indirect questions preceded by a verb phrase (now reproduced in figure 5.6), some flags can be defined, such as *INIT*, *PH1* and *PH2*. This rule will be applied if the conditions are met, as explained at the beginning of this section, and the flags *INIT* or *PH1* are on. Once the rule is applied, the flag *PH2* is switched on and the flags *INIT* and *PH1* are switched off.

Phrase		
Head	**Priority**	**Coordination Priority**
Attachments before attachments at phrase level	0001-0099	1001-1099
Prepositional phrase with preposition *de*	0101-0199	1101-1199
Prepositional phrase with other preposition	0201-0299	1201-1299
Prepositional phrase containing infinitive clause	0301-0399	1301-1399
Attachments before adjective phrase attachments	0401-0499	1401-1499
Adjective phrase	0501-0599	1501-1599
Attachments before adverb attachments	0601-0699	1501-1699
Adverb	0701-0799	1701-1799
Attachments before noun phrase attachments	0801-0899	1801-1899
Noun phrase	0901-0999	1901-1999
Non-Finite Clause		
Head	**Priority**	**Coordination Priority**
Attachments before non-finite clause attachments	2001-2099	3001-3099
Infinitive clause	2101-2199	3101-3199
Attachments before gerund clause attachments	2201-2299	3201-3299
Gerund clause	2301-2399	3301-3399
Attachments before participle clause attachments	2401-2499	3401-3499
Participle clause	2501-2599	3501-3599
Finite Clause		
Head	**Priority**	**Coordination Priority**
Verb arguments and adjuncts within finite clause	4001-4999	5001-5999
Finite clause embedded in finite clause	5000	6000
Finite clause	6001	7001
Main Clause		
Head	**Priority**	**Coordination Priority**
Verb arguments and adjuncts	7001-7999	8001-8999
Punctuation		
Head	**Priority**	**Coordination Priority**
Coma and other marks	9001-9499	——
Full stop	9501-9999	——

TABLE 5.2: System of priorities in FreeLing Dependency Grammars

5.2.2 Labelling Rules

The labelling rules (marked as *<GRLAB>* in the grammar) take care of the third operation of the Txala parser concerned with the disambiguation of the dependency relations. This set of rules operates once the attachment rules are applied and the dependency structure tree is built. The goal of these rules is to label every arc of the dependency tree with a dependency relation according to certain **constraints** and **priority**.

The rules are formed by four types of information: the dependency link, the dependency relation tag to be assigned (the tagset criteria is presented in the section § 5.3), a set of constraints and a priority. For example, in the rule of figure 5.7, the direct object label (*dobj*) is assigned to the link between a verbal head (*grup-verb*) and a prepositional phrase (*grup-sp*) child when the verbal head lemma belongs to the transitive class (*trans*) and the child is post-verbal (*right*), the preposition is *a* (or the contraction *al*), and the nominal head inside of the prepositional phrase has the Top Concept Ontology feature *Human* but not (*!=*) the features *Building* or *Place*. The application of the rule represented in the figure 5.7 produces an analysis like that in the figure 5.8 in which the prepositional object has been interpreted as a direct object, avoiding it being analysed as an indirect object, prepositional object or adjunct.

```
grup-verb   dobj
            d.label=grup-sp
            p.class=trans
            d.side=right
            d.lemma=a|al
            d:sn.tonto=Human
            d:sn.tonto!=Building|Place
```

Figure 5.7: Labelling rule for human direct objects

Regarding the set of **constraints**, they may be concerned with properties of the parent (*p*), the child (*d*) or both nodes of the dependency arc. Then, the rule applies if the parent, the child or both meet the conditions expressed in the constraints. They may refer to morphological (part-of-speech tag), lexical (lemma), syntactic (nesting position, levels of the tree structure, syntactic features) and semantic properties (EuroWordNet Top Concept Ontology features, WordNet Semantic File, Wordnet Synonyms and Hypernyms and other semantic features predefined by the user). In addition, the constraints are affirmative by default, but it is possible to negate them by using the symbol '!=' as in the figure 5.7 (*d:sn.tonto!=Building|Place*).

Part-of-speech

This constraint defines the part-of-speech tag of the syntactic head of the parent or the child nodes. For example, to restrict a rule to dependent nodestagged as a proper noun (*NP00000*), a constraint like *d.pos=NP00000* is used.

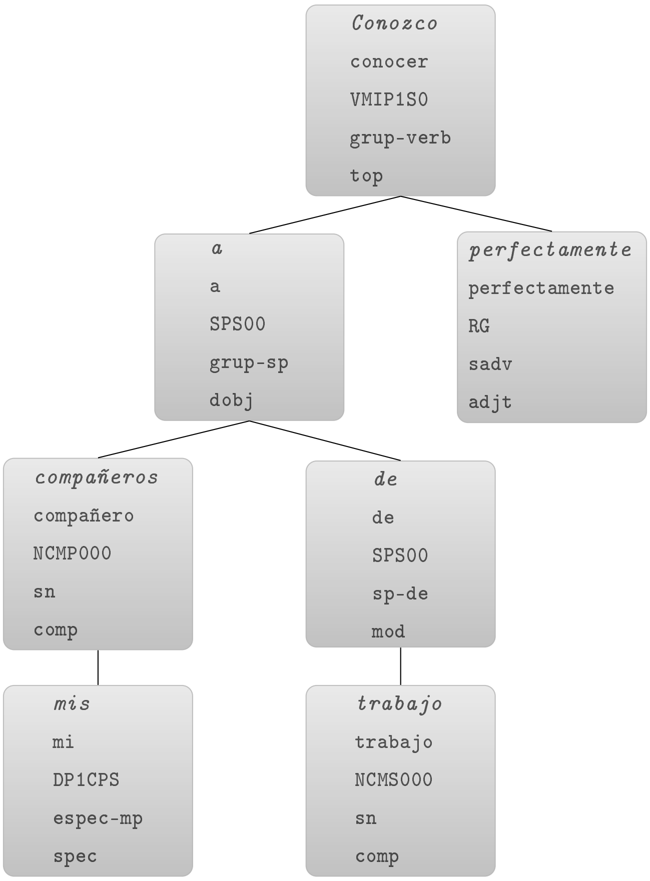


FIGURE 5.8: Dependency parse tree generated by Txala of the sentence 'Conozco a mis compañeros de trabajo perfectamente' ('Ø_{1sg} know my work colleagues perfectly')

Lemma

It is possible to access the lemma of the parent head or the child nodes. In the case of a constraint such as *p.lemma=escribir*, the rule is a candidate to be applied if the lemma of the parent matches with the lemma *escribir*.

Nesting position

This constraint refers to the position of a child node with respect to its parent node, which can be on the left (*d.side=left*) or the right (*d.side=right*).

Navigation through the tree structure

It is possible to navigate through the levels of the tree from the parent or the child nodes. For example, to describe the behaviour of pronominal verbs (e.g. 'hundirse', 'transformarse'), the rule can be restricted by requiring the verbal head to contain the pronominal particle *se* as a child, which is formally expressed like *p:morf-pron.lemma=se*.

Lists of syntactic and semantic properties
Likewise the attachment rules, lists of lemmas grouped by their syntactic or semantic properties in the section *<CLASS>* of the grammar can be also used in the labelling rules for the parent or the child nodes. For example, a rule with a constraint like *p.class=transitive* is a candidate to match if the parent node lemma is in the list of lemmas *transitive* for transitive verbs.

WordNet features
Apart from the semantic properties defined in *<CLASS>*, there is the option to use WordNet features in the constraints of the labelling rules. Part of the structure of the Spanish and Catalan WordNet 1.6 (Atserias et al., 2004) has been integrated in the FDGs, in the section *SEMDB* declared before the labelling rules. Specifically, from the imported WordNet structure, the grammar can access the set of hypernyms lemmas of the parent or child nodes, the WordNet semantic file of the parent or child nodes, and the EuroWordNet Top Concept Ontology features of the parent or child nodes (Álvez et al., 2008).

UNIQUE
This is a special feature that restricts the dependency relations tagged with this feature to be assigned once.

In the labelling rules, the **priority** of rules application is defined by the order in which rules are organized, placing the most restrictive ones first (i.e. rules with the highest priority) and the less restrictive ones last (i.e. rules with the lowest priority). Unlike the attachment rules, priorities have not been systematized since the complexity of the dependency relations' ambiguities is lower than on the attachment ambiguities.

Despite this, some principles have been established for certain high productive structures, such as the prepositional prepositional phrase (PP). The PP depending from a verb phrase can be a prepositional object, direct object, indirect object, predicative complement and adjunct. Rules to label the verbal phrase head and the prepositional head can be prioritized as follows: (1) direct object ('La profesora lleva a los alumnos al teatro' – *The teacher takes the students to the theater*) > (2) indirect object ('Ha dicho a su jefe que no iría a la reunión' – 'He said to the boss that he would not go to the meeting') > (3) time expression with preposition 'a' and infinitive ('Al llegar a casa siguió trabajando' – *When she arrived at home she kept working*) > (4) prepositional object ('La profesora lleva a los alumnos al teatro' – *The teacher takes the students to the theater*) > (5) predicative complement ('Trabaja como una loca' – *She works like crazy*) > adjunct ('Descansa después de trabajar' – *She rests after working*).

5.3 Criteria for Syntactic Dependencies Representation

The framework of Dependency Grammar (DG) originates in the postulates of *Éléments de syntaxe structurale* (Tesnière, 1959) and Mel'čuk (1988) formalized it for Natural Language Processing

(NLP) purposes. As explained in section § 2.1.2, this framework argues that the sentence is a hierarchic tree of connections between lexical units which express the relations established between them. Furthermore, it shows that the dependency relations occur between terminal symbols, i.e., between lexical units directly.

Chomsky (1981) in his work on Generative Grammar (GG) agrees with the DG in principle that the lexical units are structured hierarchically in a tree of connections. Despite this, his arguments show that connections are established between terminal symbols and non-terminal symbols which are more complex syntactic structures formed by the projection of the lexical units (§ 2.1.1).

In particular, the relations established between lexical units form a structure shaped as a hierarchic tree because some nodes govern other ones, as well as these governor nodes are also governed by other nodes. In other words, lexical units are distributed in the hierarchic tree according to the concept of syntactic head, which is going to be detailed in §5.3.1.

Nodes establish connections with other nodes in a binary way, meaning that a head can govern several dependent nodes but a dependent node connects to a single head (Tesnière, 1959; Chomsky, 1981; Mel'čuk, 1988). However, some NLP applications propose nodes to be governed by multiple syntactic heads for particular linguistic phenomena like relative clauses due to their complexity (De Marneffe et al., 2006).

In the development of a rule-based dependency grammar, besides the rule writing task, the definition of criteria that determine the behaviour syntactic heads also is a fundamental task in order to ensure high-accuracy in the syntactic dependency tree representation. The syntactic criteria proposed in this work (Lloberes and Castellón, 2011) assumes the fundamental principles of the dependency theory (Tesnière, 1959; Mel'čuk, 1988). However, in specific syntactic phenomena, this proposal is critical to the arguments of Tesnière (1959) and Mel'čuk (1988). It aims to determine syntactic representations closer to the semantic representation for the phenomena in which the DG lacks proximity to semantics.

As it will be argued in the rest of the section, this proposal is not intended to demonstrate the suitability of the dependency approach nor the generative approach, but to provide guidelines that allow parsers to map from the morphological representation to the semantic representation by providing an accurate syntactic representation.

In particular, the syntactic structures discussed in this section are auxiliaries (§5.3.2), the prepositional phrase (§5.3.3), the main subordinate clauses (§5.3.4), and the coordinating structures (§5.3.5). Although this section almost focuses on establishing criteria about syntactic structure due to its complexity, the list of syntactic relations defined for the development of the FreeLing Dependency Grammars is presented at the end of the section (§5.4).

5.3.1 The Nature of Syntactic Heads

The DG and the GG support the idea that lexical units contain syntactic and semantic features that determine the sentence configuration. However, they differ strongly with regard to the concept of lexical category and they propose different classifications of lexical categories.

Lexical units are not structured in the sentence ad hoc. Actually, the nature of lexical units (in the dependency approach) or constituents representing these lexical units (in the generative approach) determines their place in the structure (Tesnière, 1959; Chomsky, 1981). The lexical units (in DG) or the constituents (in GG) are classified into different **syntactic categories** depending on the the nature of the **syntactic heads** (Hernanz, 2002). For example, the phrase 'revista cultural' ('cultural magazine') is a noun phrase as the head of the phrase is a noun. Similarly, the phrase 'leen muchos libros' ('Ø$_{3sg}$ read a lot of books') is categorized as verb phrase whose syntactic head corresponds to a verb.

Simultaneously, syntactic categories can belong to the class of **lexical categories** or to the class of **grammatical categories** and their features match with the features described for the class that they are grouped in. The lexical categories can be inflected, they are an open list of lexical units, their meaning is complete, and they are considered major categories as they can select complements and are morphologically independent (Hernanz, 2002). Grammatical categories form a closed list of lexical units, are not usually morphologically nor phonologically independent, work as relating elements between two lexical units, and do not contribute to semantic interpretation of the sentence in the same way as the lexical categories (Hernanz, 2002).

Intuitively, classes of lexical units can be identified. However, from the theoretical point of view, there is no consensus on the distribution of the syntactic categories in lexical categories and grammatical categories (Hernanz and Brucart, 1987). Several classifications have been proposed which agree with the nature of specific syntactic categories like nouns and verbs, but have a different opinion about the rest of the categories (adjective, preposition, adverb and conjunction).

According to the generative approach (Chomsky, 1981; Haegeman, 1991), the nature of the lexical unit determines the syntactic category and, at the same time, the nature of the syntactic category conditions the syntactic head. The categories behaving this way are known as **endocentric categories** and contain the bulk of the semantic content, which determines them to work as a syntactic head (Hernanz and Brucart, 1987).

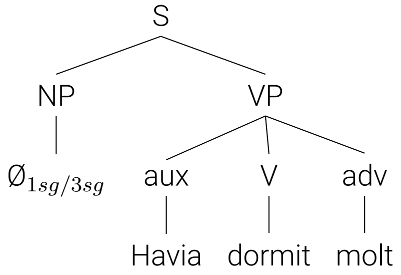
The dependency approach (Tesnière, 1959; Mel'čuk, 1988) addresses the discussion about the different syntactic behaviours observed in the lexical units from another point of view. This framework identifies a set of lexical units with complete meaning which have a semantic function because they represent an idea by themselves. On the other hand, the lexical units which do not have complete meaning are considered semantically empty and they work as functional units in the structure of the sentence (Tesnière, 1959). Their function in the structure corresponds to relating lexical units with complete meaning (Tesnière, 1959). According to this classification of the lexical units, the true lexical units are able to work as syntactic heads, whereas functional units work around the heads and are no able to appear in syntactic head positions.

5.3.2 Auxiliary is Auxiliary

Spanish and Catalan keep a single auxiliary verb, 'haber' in Spanish and 'haver' in Catalan ('to have'), and it occurs in perfect tenses with a non-finite form that takes the participle form. In both languages the auxiliary has a conflictive representation and the linguistic frameworks analysed here understand the auxiliary in different ways. The verb phrase can be auxiliary headed when the auxiliary is the head of the verb phrase (*aux* ⟶ *V*), or it can be verb headed if it is a non-finite verb form which is the head of the verb phrase and the auxiliary its child (*aux* ⟵ *V*).

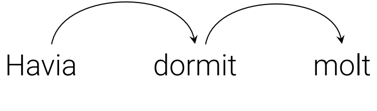
The GG considers (Chomsky, 1957; Bonet and Solà, 1986; Haegeman, 1991) the auxiliary as a lexical unit that provides grammatical properties (e.g. number, person, tense, mode) to the main verb of the sentence in the same way as the determiners contribute to in the noun phrase, for example. For this reason, the analysis from the generative point of view sets the non-finite form in the head position because it is the lexical unit that sub-specifies the rest of the structure of the sentence, while the auxiliary is a lexical unit that transfers the grammatical features to the main verb from a child position (1).

(1) *Havia* dormit molt
'Ø$_{1st/3sg}$ had slept a lot'



On the other hand, the dependency framework (Tesnière, 1959; Mel'čuk, 1988) argues that agreement rules are carried out by lexical units agreeing directly. For this reason, these authors reject that the head of the verb phrase is the non-finite form transferring the grammatical properties from the auxiliary. Therefore, the verb phrase head necessarily is the auxiliary (2).

(2) *Havia* dormit molt
'Ø$_{1st/3sg}$ had slept a lot'



FDGs criterion

In the FDGs, the generative approach is implemented because the representation that it proposes is closer to the semantic representation than the dependency approach representation which is syntactically motivated. The main verb is the lexical unit which has underlying the properties of the predicate argument structure. For this reason, this proposal considers that the auxiliary depends on the main verb as represented in the figure 5.9.

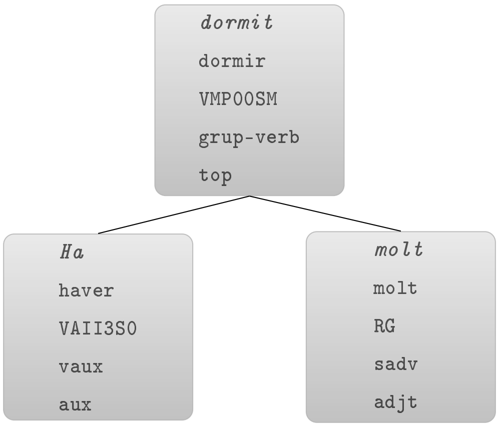


FIGURE 5.9: Dependency parse tree generated by Txala of the sentence 'Ha dormit molt'

Passive voice and raising verbs are analysed following the same pattern as the auxiliary verb in the FDGs because both structures behave the same way. The non-finite form contains the semantic content and the syntactic and semantic properties, and the auxiliary provides the grammatical properties to the non-finite form.

5.3.3 Meaningful Preposition

The syntactic nature of the preposition has been the center of several theoretical works. These works have focused on describing the complexity of the preposition's behaviour in detail (Fabra, 1918; Badia i Margarit, 1962; Alarcos, 1994; Bonet and Solà, 1986; Hernanz, 2002). They argue that, while traditionally the noun, the adjective and the verb are considered lexical categories, the preposition is controversial.

The notion of preposition is vague in the traditional grammar (Fabra, 1918; Badia i Margarit, 1962; Alarcos, 1994). An example of this asystematization is the confusion between some prepositional and adverbial uses (Tesnière, 1959; Bonet and Solà, 1986; Moreno Cabrera, 2000). The preposition has been traditionally classified into transitive uses (3-a), in which the preposition necessarily occurs with a complement, and intransitive uses (3-b), where the complement is optional. In particular, the intransitive uses are the contexts described as adverbs used as prepositions by the traditional grammar (Bonet and Solà, 1986).

(3) a. El pasillo conduce *al comedor*
'The aisle leads to the dinning room'
b. Encontrarás las llaves {*encima de la mesa / encima*}
'$\emptyset_{2sg}$ will find the keys {on the table / on}'

On the other hand, more recent theoretical and descriptive linguistic studies point out that the preposition has similar uses to the lexical categories which are able to have a syntactic head, although its behaviour distinguishes it from the major lexical categories like the noun or the verb (Bonet and Solà, 1986; Hernanz and Brucart, 1987; Hernanz, 2002). The major lexical categories that are inflected, belong to open lists of words, their content is descriptive, select complements, and they are morphologically independent (§5.3.1). However, the preposition is invariable, its meaning is more abstract than the major lexical categories, and it cannot occur without complements (4).

(4) a. Laia escucha *música* $_{NP}$
'Laia listens music'
b. Laia parece *triste* $_{AP}$
'Laia seems sad'dreaming
c. Laia *ríe* $_{VP}$
'Laia laughs'
d. Laia nada *bien* $_{Adv}$
Laia swims well
e. *Laia confía *en* $_{PP}$
'Laia trusts in'

According to the examples in (4), noun phrase (4-a), adjective phrase (4-b), verb phrase (4-c) and adverb (4-d) are endocentric phrases, i.e., phrases that are built over the base of a syntactic head. Tesnière (1959) excludes the prepositional phrase from the endocentric categories. Actually, the prepositional phrase can only be considered as an exocentric category, i.e., phrases that do not contain any head. It always requires a complement in comparison to the other types of phrases (otherwise, the sentence is ungrammatical like in (4-e)). This behaviour shows that it is not able to occur in the syntactic head position.

The GG explains the nature of the preposition from the opposite side (Chomsky, 1981; Bonet and Solà, 1986; Hernanz and Brucart, 1987). It argues that all the syntactic projections are endocentric, including the preposition, as discussed in § 5.3.1. However, unlike other lexical categories that can be transitive and intransitive, the preposition necessarily subcategorizes a complement. For this reason, the sentence in (4-e) is ungrammatical.

This hypothesis is confirmed by the parallel behaviour of some transitive verbs which only accept transitive diathesis (5). The ungrammaticality of the sentence (5-b) involve considering the verb an exocentric category. If it was exocentric, it could not be in the class of lexical categories (Hernanz and Brucart, 1987). Verb is a lexical category because the verb phrase necessarily has a syntactic head. Then, the factor that determines the ungrammaticality of the sentence (5-b) is another one. The set of syntactic and semantic properties of the verb itself are responsible for defining the argument structure of the verb. For example, the verb 'saber' ('to know') of the sentence (5-a) predicts in its argument structure an argument realized syntactically as a direct object. Therefore, the absence of this argument causes the ungrammaticality of the construction.

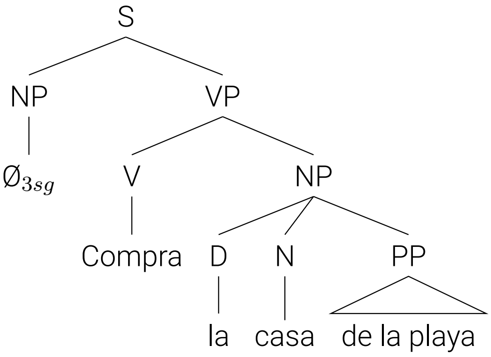
(5) a. Supe *que se había ido*

‘[I] knew that he had left’

b. *Supe
‘[I] knew’

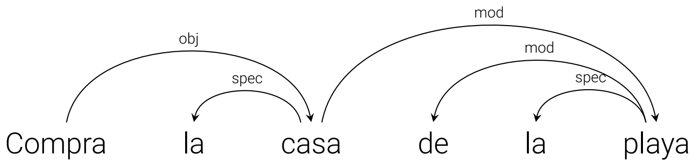
Since Generative Grammar categorizes the preposition as an endocentric category, this syntactic category accepts syntactic projections like the rest of lexical categories, so that the preposition is capable of behaving as a syntactic head, as shown in the example (6).

(6) Compra la casa *de la playa*
‘Ø$_{3sg}$ buys the beach house’



Similarly to traditional grammar, Generative Grammar (Chomsky, 1981) and Dependency Grammar (Tesnière, 1959) consider that the preposition is a relational lexical element. However, the Dependency Grammar argues that the preposition behaves very differently from the noun, the adjective and the verb behaviour (Tesnière, 1959). According to this framework, the preposition does not contain any semantic function and, for this reason, it is classified as a functional unit, i.e., a linguistic unit used in the discourse exclusively to indicate and to transform the category of the lexical units, and to determine the relations existing between the lexical units. As a consequence, the preposition cannot work as a syntactic head (7).

(7) Compra la casa *de la playa*
‘Ø$_{3sg}$ buys the beach house’



FDGs criterion

Freeling Dependency Grammars (FDGs) follow the postulates of the dependency approach. However, they consider that the preposition is a grammatical unit with a semantic function, although it is obvious that it behaves differently from major syntactic categories because it is a relational element. The evidence for this can be observed in the distinction between prepositions meaningfully full and prepositions meaningfully empty (Hernanz and Brucart, 1987). The former type refers to prepositions which introduce adjuncts like locatives (8-a) and the second type refers to the prepositions subcategorized by the verb (8-b).

(8) a. Los excursionistas andan por el campo
'The hikers walk through the field'
b. La empresa apuesta por inversiones en el extranjero
'The company supports investments in foreign countries'

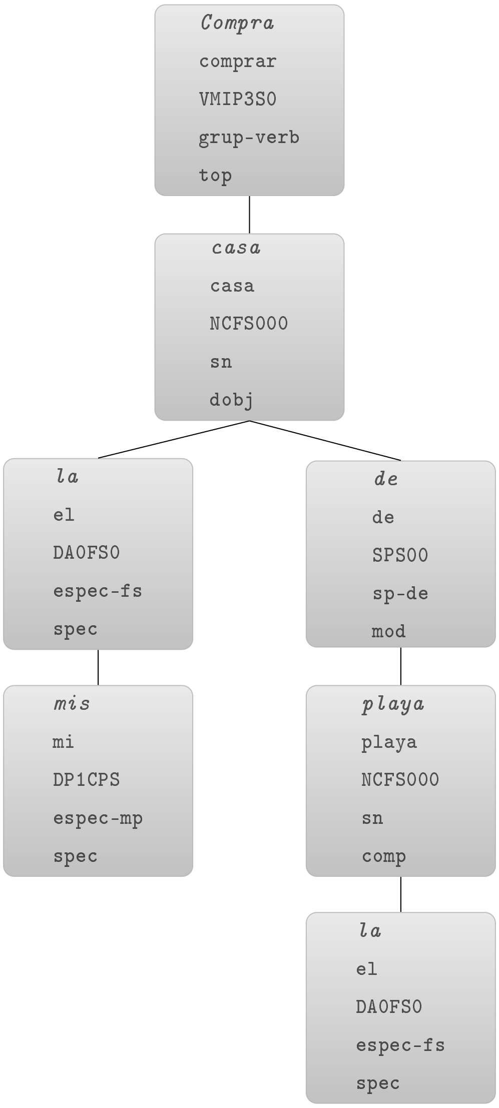


FIGURE 5.10: Dependency parse tree generated by Txala of the sentence 'Compra la casa de la playa'

The possibility that the verb of the predicate subcategorizes an argument introduced by a specific preposition, as shown in (8-b), is further proof of the preposition as a lexical category. Therefore, in FDGs, the preposition is a lexical category that works as a lexical unit and that is able to appear in the syntactic head position (figure 5.10).

5.3.4 Structural Diversity in Subordinate Clauses

Under the term 'subordinate clause' there are several constructions which their own properties and meaning. In this section, the finite clauses are only considered because of the complexity of their representation and the different solutions proposed by the generative and the dependency frameworks. They occur as an argument or an adjunct of the verb of the predicate (Villalba, 2002). Unlike phrases, the relation between the main clause and the subordinate clause is realized by a grammatical operator or marker. Depending on the nature of the subordinate clause its structure varies. In the following sections, several subordinate clauses are examined from the generative and dependency point of view and a criterion is proposed for implementing it in the FDGs. In particular, the clauses discussed here are substantive and adverbial clauses (§5.3.4.1), relatives (§5.3.4.2) and free relatives (§5.3.4.3).

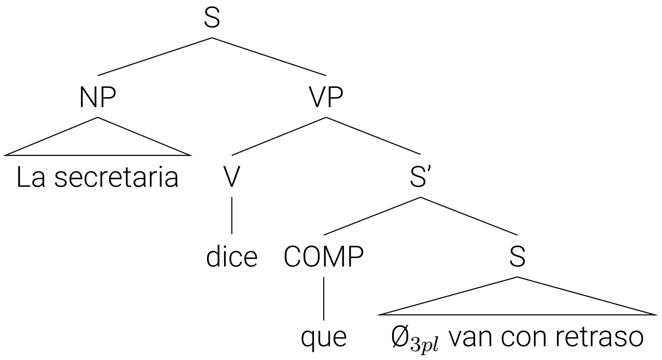
5.3.4.1 Substantive and Adverbial Clauses

The GG argues that the complementiser (COMP) is the lexical category that defines the subordinate construction (Chomsky, 1957; Bonet and Solà, 1986; Haegeman, 1991) and that the traditional grammar identifies it as a conjunction (also as a relative pronoun as it will be described in §5.3.4.2). The complementiser is held to be the syntactic head of the subordinate clause (Bonet and Solà, 1986; Haegeman, 1991) following the schema of the rewriting rules for embedded subordinate clauses (9). However, as argued in the next section (§5.3.4.2), the schema of the clause is different depending on the nature of the complementiser.

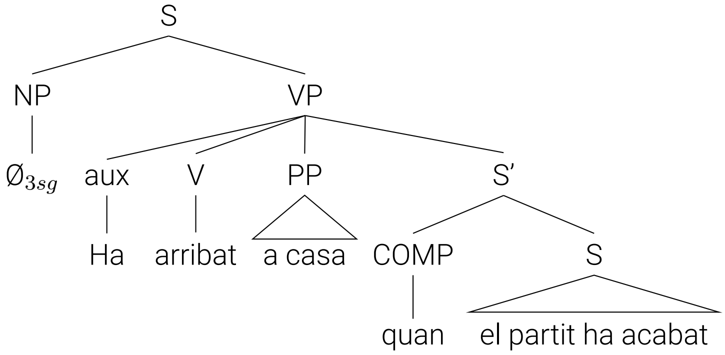
(9) a. O → SN SV
b. O' → COMP O

The schema (9) explains the formation of the substantive clauses and the adverbial clauses (Bonet and Solà, 1986; Haegeman, 1991). The complementiser is exclusively a grammatical operator in both types of clauses which embed the subordinate clause in the main clause. Therefore, the conjunction is a marker that introduces the subordinate clause and that appears in the beginning of the clause, as observed in the following examples of a substantive clause (10) and an adverbial clause (11).

(10) El secretario dice *que van con retraso*
'The secretary says that Ø$_{3pl}$ are late'

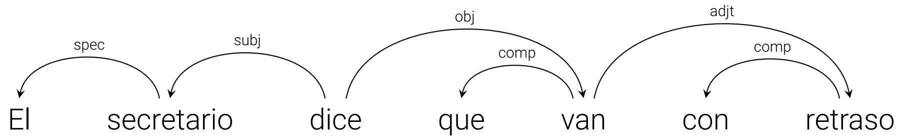


(11) Ha arribat a casa *quan el partit ha acabat*
Ø_{3sg} has arrived at home when the match has finished

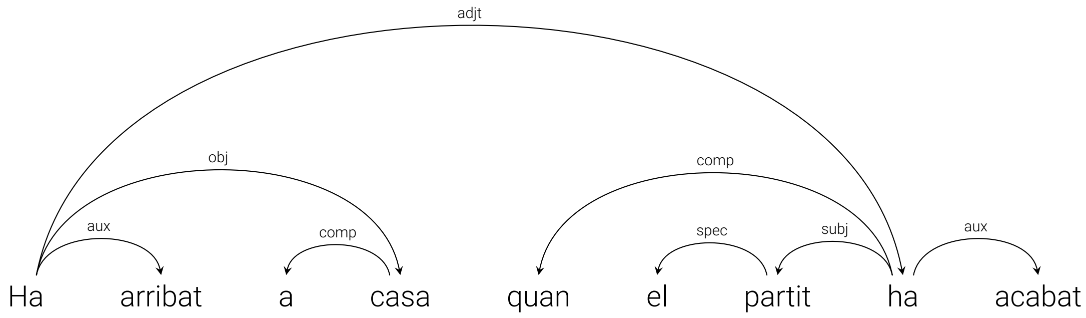


On the other hand, the DG considers that the conjunction belongs to the group of functional units (Tesnière, 1959), i.e., to the group of syntactic categories without explicit content that behave as a grammatical element. Consequently, this category is not capable of working as a syntactic head of the subordinate clause nor the adverbial clause. Actually, the category that introduces this clause is the verb of the subordinate clause as it is a lexical category with semantic content able to work as a syntactic head.

(12) El secretario dice *que van con retraso*
'The secretary says that Ø_{3pl} are late'



(13) Ha arribat a casa *quan el partit ha acabat*
Ø_{3sg} has arrived at home when the match has finished



FDGs criterion

When considering the representation of subordinate clauses in the FDGs, the idea that the conjunction is a grammatical operator is supported according to the GG (Chomsky, 1957) and the DG (Tesnière, 1959). However, this proposal differs from the dependency approach because the conjunction is a syntactic category that introduces the substantive clause (figure 5.11) and the adverbial clause (figure 5.12).

Therefore, the criteria proposed for the subordinate clauses supports the generative argument in which the conjunction is the category that introduces the subordinate clause and it occurs in the head position (Moreno Cabrera, 2000; Bonet and Solà, 1986). Despite this, the behaviour of this category is distant from the lexical categories like noun, adjective, verb and adverb, since it is the marker that embeds the subordinate clause in the main clause.

5.3.4.2 Relative Clause

As mentioned in the previous section (§5.3.4.1), the grammatical operator of the relative clause is different from the subordinate clauses (14-a). Broadly speaking, the marker in the relative clause works the same way as in the subordinate clauses, relating lexical categories. However, simultaneously, it establishes an anaphoric relation with its antecedent (Brucart, 1997; Solà et al., 2002) and, specifically, behaves as a relative pronoun (14-b). Due to its pronominal nature, the relative pronoun is inserted in the argument structure of the subordinate clause and establishes a syntactic function with the verb of the subordinate clause (14-c).

(14) a. El actor *que sale en esa película* le trajo recuerdos de su pasado
'The actor who plays a role in that film brought to him memories from his past'

b. actor ⟵ que

c. *El actor*$_{subj}$ sale en esa película

From the point of view of the generative framework, the relative pronoun introduces the relative clause (Brucart, 1997; Solà et al., 2002). For this reason, this complementiser is the syntactic head of this class of clauses and behaves similarly to the complementiser of the substantive and

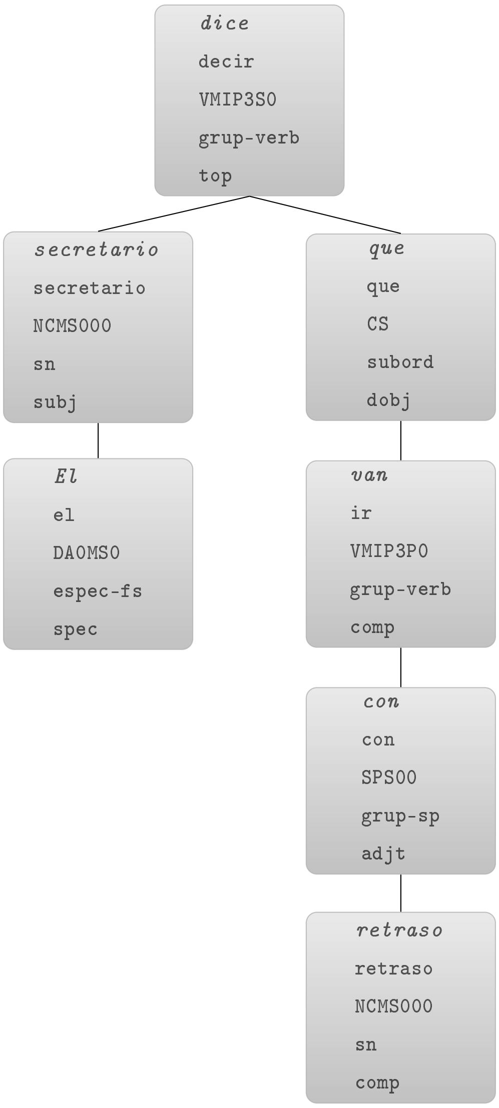


FIGURE 5.11: Dependency parse tree generated by Txala of the sentence 'El secretario dice que van con retraso'

adverbial clauses. However, at the same time they should also be represented expressing their pronominal nature.

Consequently, the GG considers the relative pronouns as a category that appears in the COMP position and is marked as a positive wh-word. Originally, the wh-constituents are predicted in the verb predicate argument structure. However, they are placed in the COMP position like in example (15) because of a rule of movement of wh-word (Bonet and Solà, 1986; Haegeman, 1991).

(15) El actor *que sale en esa película* le trajo recuerdos de su pasado
'The actor who plays a role in that film brought to him memories from his past'

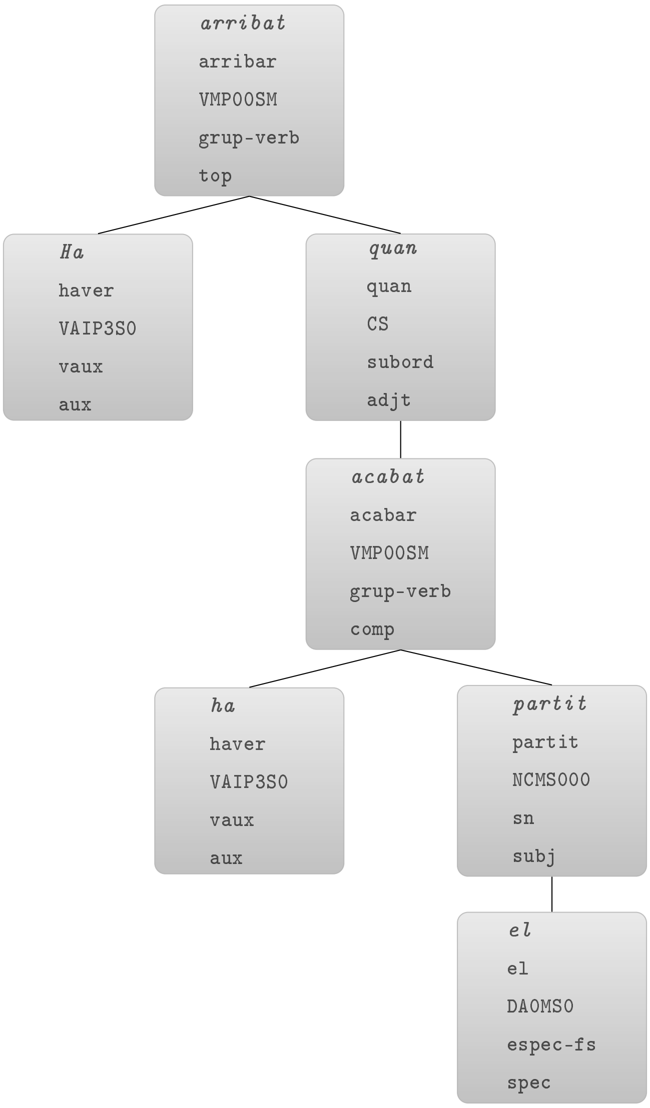
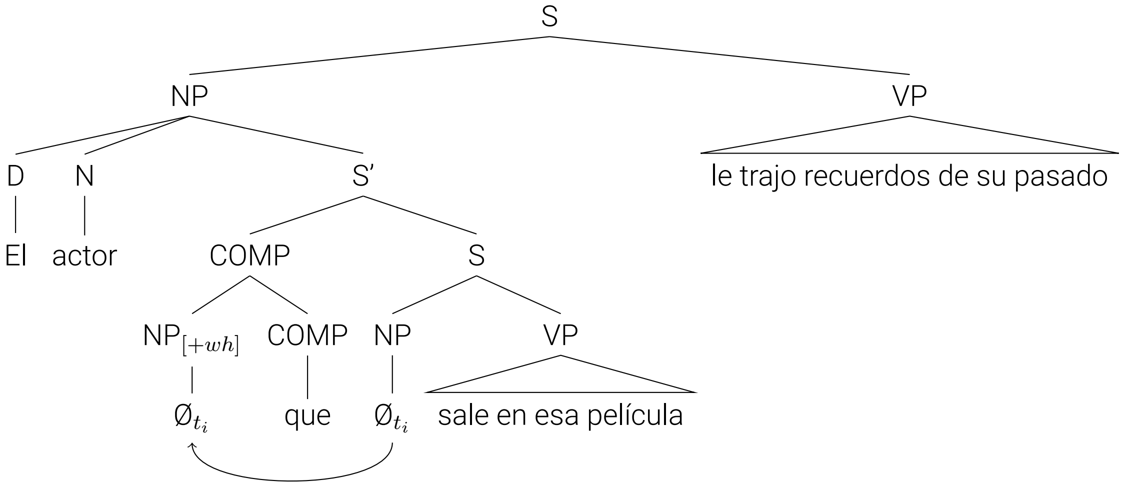
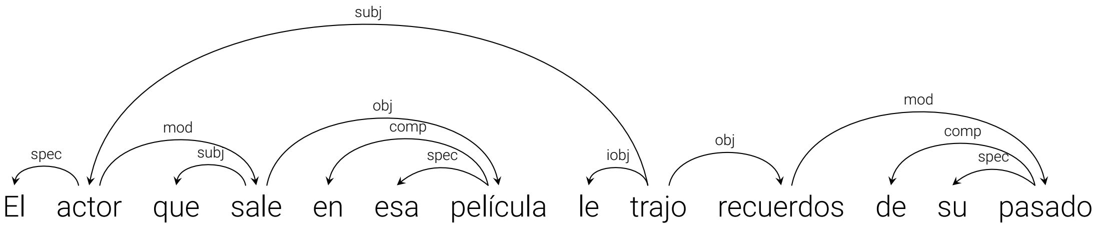


FIGURE 5.12: Dependency parse tree generated by Txala of the sentence 'Ha arribat quan el partit ha acabat'



The DG proposes a completely different description of the relative clause (Tesnière, 1959). Similar to the conjunctions, the relative pronoun is included in the set of functional units. In particular, according to Tesnière (1959), it is a grammatical operator whose function is to transform the category of the semantically complete lexical units and that operates over lexical units to relate them. Therefore, the relative pronoun is not able to behave as a syntactic head (16).

(16) El actor *que sale en esa película* le trajo recuerdos de su pasado
‘The actor who plays a role in that film brought to him memories from his past’



FDGs criterion

In the FDGs, the relative pronoun is understood as a grammatical category, i.e., a category that relates clauses and that marks the boundary of the clause, following the hypothesis of the GG (Bonet and Solà, 1986; Haegeman, 1991). However, the syntactic representation of this category performed in the FDGs supports the postulates of the DG by means of stating that the syntactic head of the relative clause is the verb of this clause (so the relative pronoun depends on this verb). This proposal supports the idea that the relative pronoun is predicted in the predicate argument structure and it is a linguistic unit which has its semantically complete meaning (figure 5.13).

5.3.4.3 Free Relative Clause and Indirect Question

A free relative clause (17-a) is a construction that is syntactically ambiguous with the indirect question (17-b) because both clauses use the same repertoire of markers (Solà et al., 2002). Furthermore, the free relative typically appears in the same position as the substantive and adverbial subordinate clauses (Moreno Cabrera, 2000; Delbecque and Lamiroy, 1999; Bonet, 2002; Villalba, 2002).

(17) a. Ignoraban *de quien hablabas*
‘Ø$_{3pl}$ ignored who Ø$_{2sg}$ were talking about’
b. Ha preguntado *quién viene*
‘Ø$_{3sg}$ asked who Ø$_{3sg}$ is coming’

Despite the coincidences among several types of clauses, the free relative clause is clearly marked by a wh-word, i.e., by the feature [+qu], according to the generative postulates (Bonet and Solà, 1986; Haegeman, 1991). Furthermore, they differ from the indirect questions which are marked with the feature [+QU] (Bonet and Solà, 1986) and, for this reason, the marker can only occur in

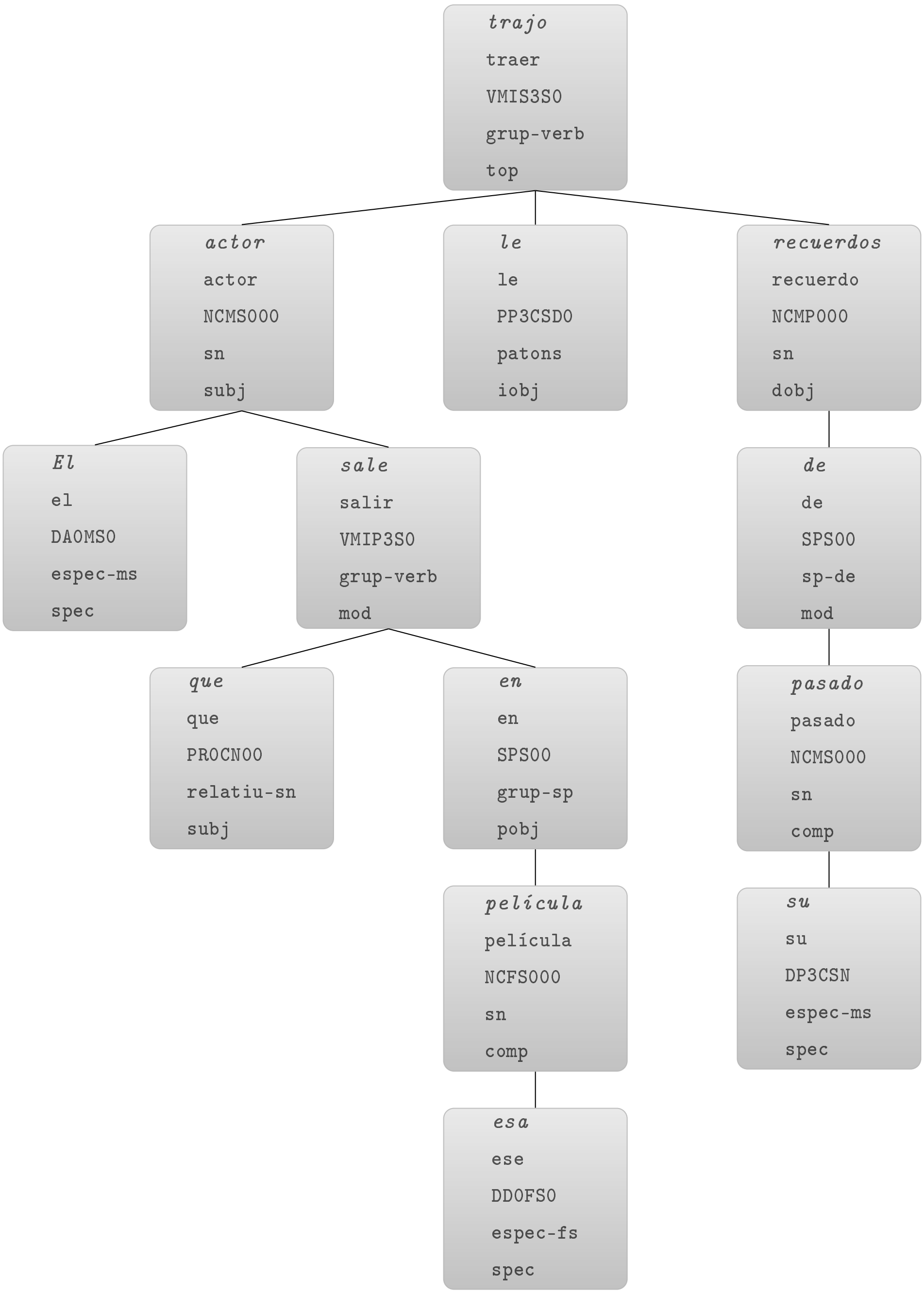
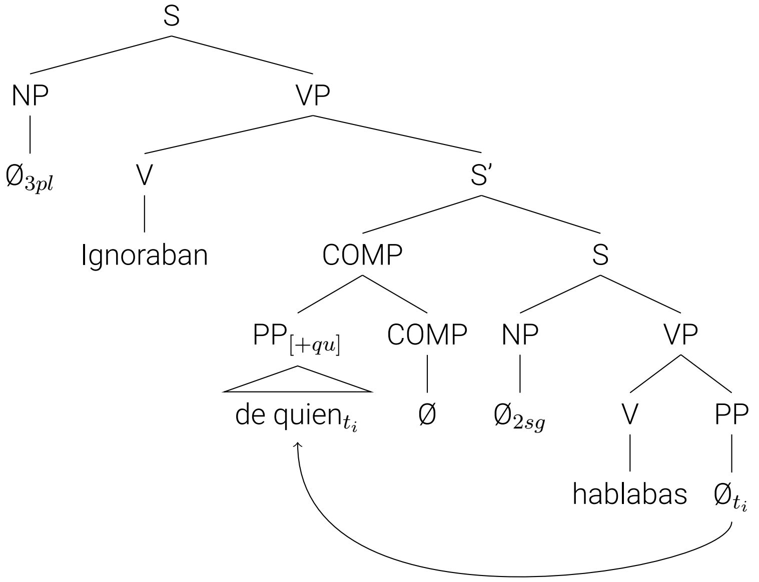


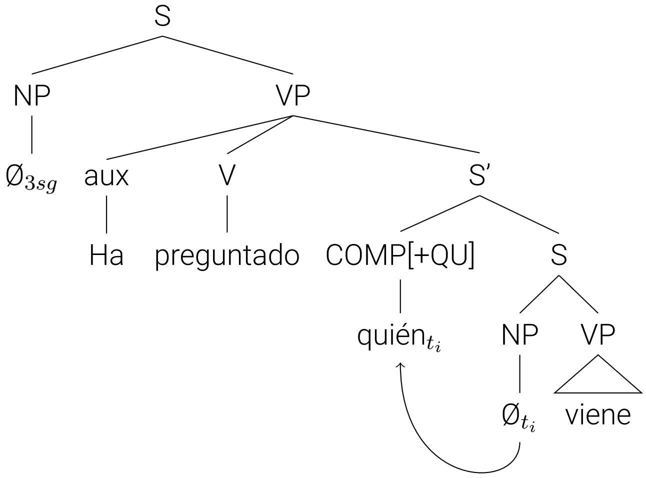
FIGURE 5.13: Dependency parse tree generated by Txala of the sentence ‘El actor que sale en esa película le trajo recuerdos de su pasado’

the complementiser position. As discussed in §5.3.4.2, the generative approach supports that the wh-word is the syntactic head of the clause, so the marker is the head of the free relative. Therefore, from this point of view, the relative clause and the free relative clause have an identical structure (18-a). Meanwhile, the indirect question structure is parallel to the substantive and adverbial clauses (18-b).

(18) a. Ignoraban *de quien hablabas*
‘$Ø_{3pl}$ ignored who $Ø_{2sg}$ were talking about’

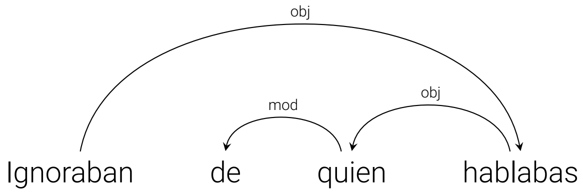


b. Ha preguntado *quién viene*
‘$Ø_{3sg}$ asked who $Ø_{3sg}$ is coming’

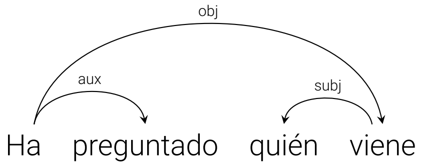


On the other hand, the DG (Tesnière, 1959) considers that the structure of the free relative is parallel to the structure of the relative clauses (§5.3.4.2). In other words, the grammatical operator of the free relative is the relative pronoun that is a functional unit not able to appear in the head position of the clause. Therefore, the syntactic head of the free relative clause is the verb of the subordinate clause (19-a). Concerning the indirect question, this construction is considered parallel to the free relative, then, the marker is a child of the subordinate verb and this verb is the head of the clause (19-b).

(19) a. Ignoraban *de quien hablabas*
‘$Ø_{3pl}$ ignored who $Ø_{2sg}$ were talking about’



b. Ha preguntado *quién viene*
'$\emptyset_{3sg}$ asked who $\emptyset_{3sg}$ is coming'



FDGs criterion

In our approach, the free relative has the same behaviour as the relative clause. The relative pronoun expresses one of the arguments specified in the subordinated verb subcactegorization frame. Therefore, in the FDGs it does not appear in the clause head position, although this pronoun also works as a marker of the clause, and the verb of the subordinate clause is the lexical unit that embeds the free relative in the main clause (figure 5.14).

On the other hand, in the FDGs, the indirect question is analysed as a parallel structure of the substantive clause because following the generative postulates the behaviour of the grammatical operator that introduces this kind of clause is different from the operator of the relative and the free relative clauses. This is why the GG describes it with the feature [+QU] and assumes that is the head of the construction. The FDGs agree with the approach of the generative grammar, so it performs the indirect question construction structure with the marker embedding the interrogative clause in the main clause (figure 5.15).

5.3.5 Encoding Coordination Structures

In the previous sections, syntactic structures based on hierarchical relations and their syntactic heads have been analysed in detail (e.g. auxiliary, prepositional phrase attachment and subordinate clauses). In this section, the discussion is focused on another mechanism of the language to connect words, phrases, clauses or more complex syntactic structures when their relation is non-hierarchical, known as **parataxis**. Paratactic syntactic structures are difficult to represent in the Dependency Grammar since this formalism is based on the notion of dependency. Among the paratactic structures, coordination is discussed since it is a high-frequent phenomenon in language and in which parsers show a low-performance (Popel et al., 2013).

Coordination is a mechanism to link two linguistic structures or conjuncts which usually occurs with a marker corresponding to a coordinating conjunction (e.g. 'and', 'or', 'but' in English). Traditionally, it has been stated that coordinating structures are structured at the same level (Fabra, 1918; Badia i Margarit, 1962; Alarcos, 1994; García, 1999). However, despite the fact that

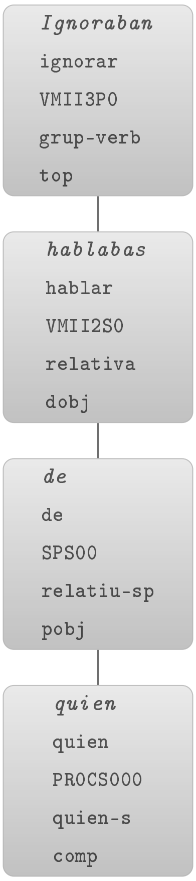


FIGURE 5.14: Dependency parse tree generated by Txala of the sentence 'Ignoraban de quien hablabas'

the construction has been largely studied, the theoretical works focused on the coordination have differ widely.

The discussion mainly centred around whether the phrases are structured at the same level or hierarchically (Bonet and Solà, 1986; García, 1999; Serra i Prunyonosa, 2002). According to the generative representation of coordination, this construction contains at least two equivalent elements regarding their grammatical function in the sentence which are linked at the same level of the structural hierarchy with a linking piece (Dik, 1968). Following Dik (1968) definition, Bonet and Solà (1986) hypothesize that coordination can be explained by the rule (20).

(20) X ⟶ X Coord X

This rule (20) allows empty lexical units and more complex structures. Moreover, it structures all the conjoins at the same level of the structure as showed in the constituency tree representation of (21).

(21) Diuen que acabi de pressa i exigeixen que faci una rebaixa en el preu final

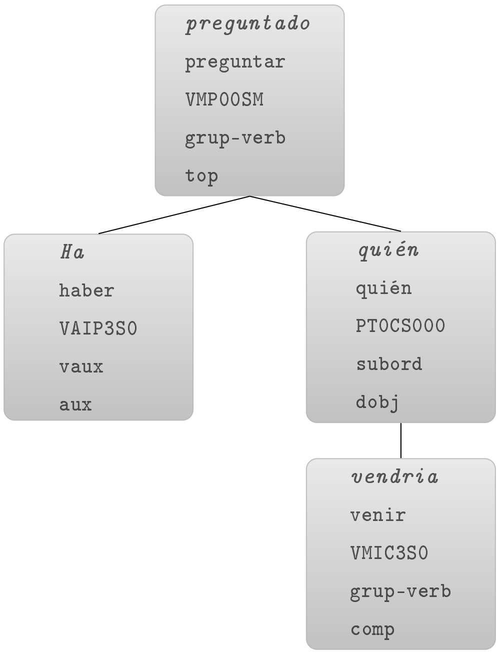
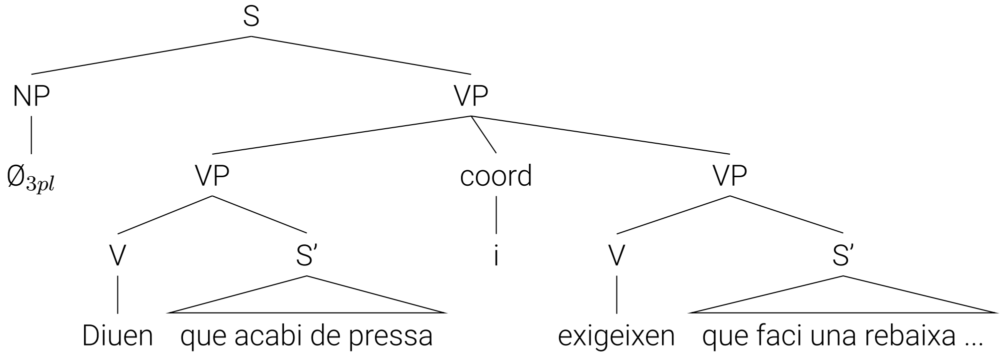


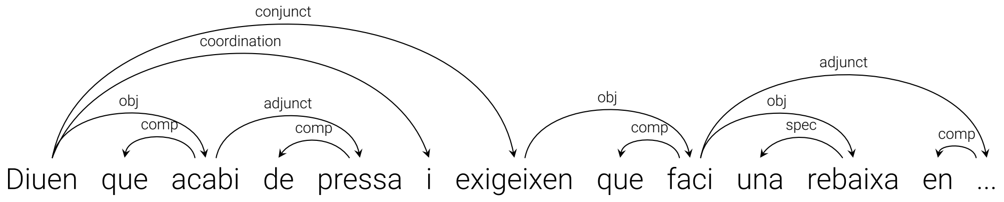
FIGURE 5.15: Dependency parse tree generated by Txala of the sentence ‘Ha preguntado quién vendría’

‘$Ø_{3pl}$ say [to him] to finish fast and $Ø_{3pl}$ demand [to him] to lower the final price’



The DG describes the behaviour of coordinating constructions differently. The coordination beside the conjunction of substantive and adverbial clauses uses a marker which corresponds to a conjunction. Consequently, Tesnière (1959) associates it to the class of functional units. Since the coordinating conjunction does not contain any semantic function, Mel’čuk (2003) states that this category is not able to behave as a syntactic head (22-a).

(22) a. Diuen que acabi de pressa i exigeixen que faci una rebaixa en el preu final
‘$Ø_{3pl}$ say [to him] to finish fast and $Ø_{3pl}$ demand [to him] to lower the final price’

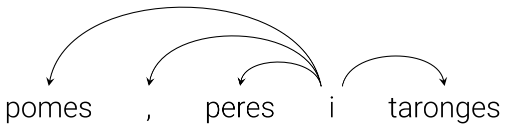

conjunct
coordination
adjunct
obj
comp
adjunct
comp
obj
comp
obj
spec
comp
Diuen que acabi de pressa i exigeixen que faci una rebaixa en ...

Despite Tesnière (1959) postulates, in the dependency approach there is no consensus on the coordination configuration among computational linguistic resources and parsers. Actually, the multiple proposals draw a very sparse scenario which make it difficult to provide a satisfactory solution about coordinating structures for parsing.

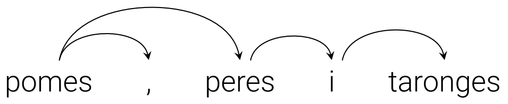
In order to start unifying the several representations of coordination, Popel et al. (2013) study the main treebanks and how they encode this construction. They observe three main tendencies and they name them as the **Prague family** (23) which refers to the resources following the Prague Dependency Treebank (Hajič et al., 2012), the **Moscow family** (24) for the treebank proposals applying Mel'čuk (1988) postulates, and the **Stanford family** (25) that corresponds to the treebanks in which coordination is encoded with the criteria of the Stanford Parser (Klein and Manning, 2003) and which the project of Universal Dependencies about unified syntactic dependencies annotation originates in (Mcdonald et al., 2013).

Specifically, the three treebank families can be classified around four features concerning the choice of head between conjunction or conjunct (leftmost, rightmost, mixed), the attachment of shared modifiers (head, nearest conjunct), the attachment of the coordinating conjunction (part of the chain of conjuncts, previous conjunct, following conjunct), the attachment of punctuation (part of the chain of conjuncts, previous conjunct, following conjunct). Combining the treebank family with several features, Popel et al. (2013) describe a total of 18 configurations of different ways of encoding coordination, which show the necessity of unifying the representation of the coordination structure.

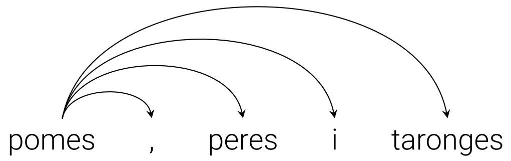
(23) pomes, peres i taronges
'apples, pears and oranges'



(24) pomes, peres i taronges
'apples, pears and oranges'

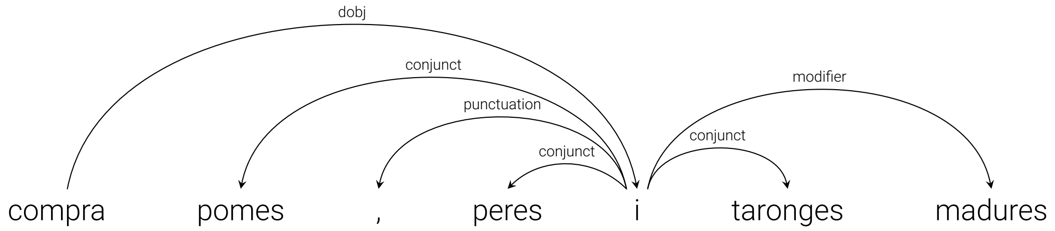


(25) pomes, peres i taronges
'apples, pears and oranges'

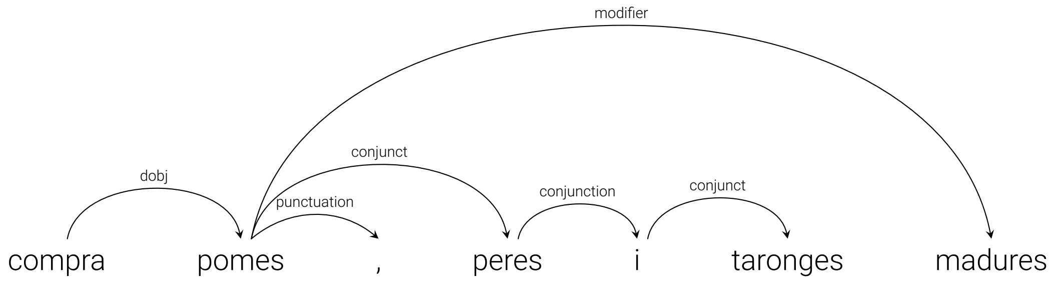


With regarding to dependency relations, in the Prague family the conjuncts are labelled with a dependency relation expressing the relation of coordination (26), whereas in the Moscow and Stanford families the head of the coordination which is one of the conjuncts cannot be labelled as conjunct and needs to be deduced from the structure, as shown in the Moscow configuration representation (27) and the Stanford configuration representation (28).

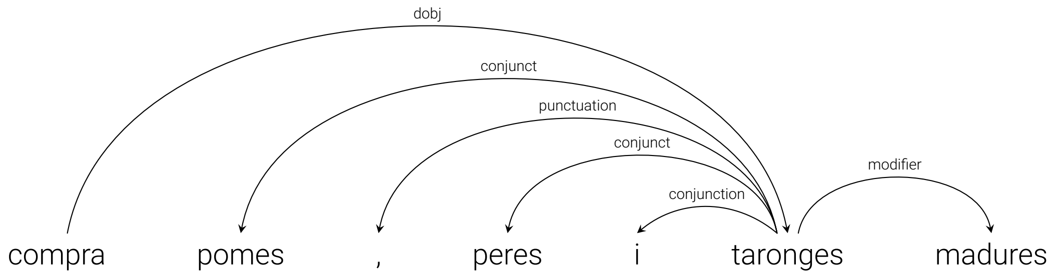
(26) compra pomes, peres i taronges madures
‘$Ø_{3sg}$ ripe apples, pears and oranges’



(27) compra pomes, peres i taronges madures
‘$Ø_{3sg}$ buy ripe apples, pears and oranges’



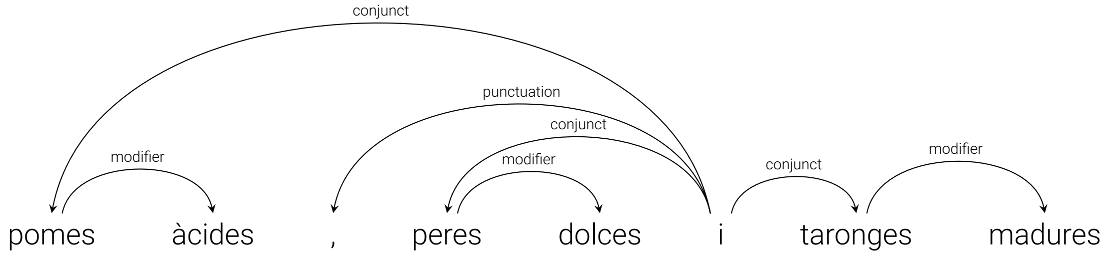
(28) compra pomes, peres i taronges madures
‘$Ø_{3sg}$ ripe apples, pears and oranges’



Furthermore, in most of the Prague family treebanks and, specifically, in the Prague Dependency

Treebank, shared modifiers are attached to the head and local modifiers are attached to the correspondent node. Therefore, the Prague Dependency Treebank avoids any ambiguity in the representation of shared (26) or local modifiers (29).

(29) pomes àcides, peres dolces i taronges madures
‘sour apples, sweet pears and ripe oranges’



On the other hand, in the Moscow (27) and Stanford families (28), it is not possible to interpret a nested modifier as shared among the conjuncts or local (i.e. modifier of the head of the coordination only) from a representational point of view. Therefore, special labels need to be used to distinguish both uses.

As a consequence of these multiple considerations, Popel et al. (2013) conclude that the Prague family is more expressive than the Moscow and Stanford families.

FDGs Criterion

In FDGs, these observations were taken into account in order to propose a robust criterion for the configuration of the coordinating construction. This proposal differ from the traditional dependency framework (Tesnière, 1959) and the Moscow and Stanford families of treebanks.

Following the configuration of the Prague family treebanks, the FDGs agree with Popel et al. (2013) proposal and they state that the head of the coordination construction is the coordinating conjunction (figure 5.16). This proposal agrees with a representation which expresses that the coordination is not a hierarchical structure but a structure where the conjunct are at the same level (Bonet and Solà, 1986).

Furthermore, the hierarchical structure is not supported in this proposal because it causes structural ambiguities frequently (Mel’čuk, 2003). For example, the coordination construction ‘wonderful photographs and paintings’ can be interpreted in two ways. The adjective ‘wonderful’ can modify locally the noun ‘photographs’, or it can modify globally the coordinated nominal structure ‘photographs and paintings’. On the other hand, most of the treebanks of the Prague family attach the shared modifier to the head of the coordination. However, the FDGs attach modifiers to the nearest conjunct due to restrictions of the Txala Parser.

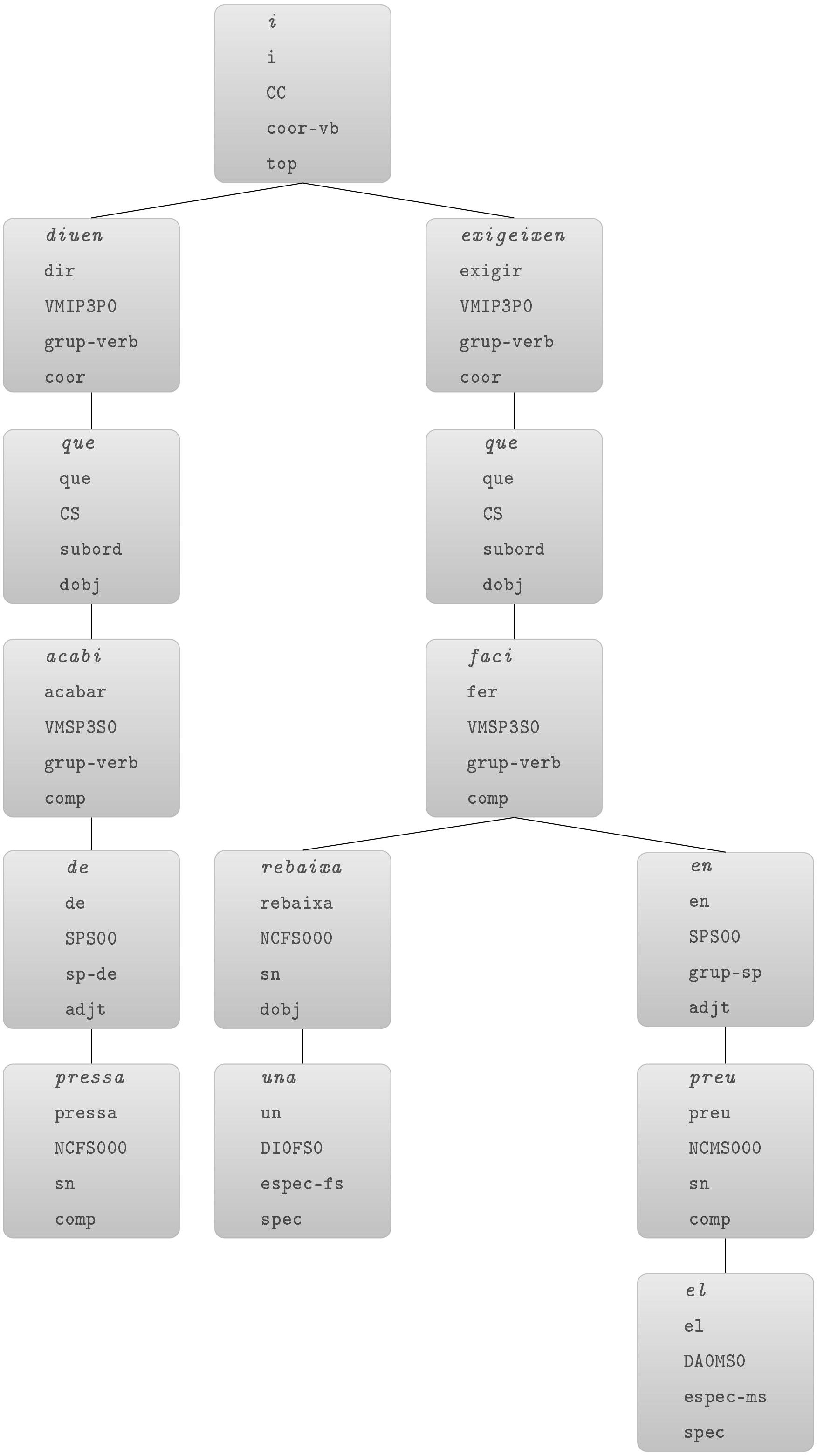


FIGURE 5.16: Dependency parse tree generated by Txala of the sentence 'Diuen que acabi de pressa i exigeixen que faci una rebaixa en el preu final'

Recapitulation

In this section, the main theoretical arguments of the generative and the dependency frameworks have been discussed concerning the representation of the most controversial syntactic structures. In particular, the nature of the syntactic head have been detailed for the auxiliary, the prepositional phrase, the subordinate clauses (substantive and adverbial, relative, free relative and indirect question) and the coordinating construction. This description has made possible to reflect and to determine the syntactic configuration of these constructions for the FDGs. This proposal is summarized in table 5.3.

Construction	**Structure**	**Head**
Auxiliary	$\text{V}_{auxiliary} \longleftarrow \text{V}_{non-finite}$	non-finite verb
Prepositional Phrase	$\text{P}_{reposition} \longrightarrow \text{XP}_{hrase}$	preposition
Substantive Clause	$\text{C}_{conjunction} \longrightarrow \text{V}_{subordinate}$	conjunction
Adverbial Clause	$\text{C}_{conjunction} \longrightarrow \text{V}_{subordinate}$	conjunction
Relative Clause	$\text{R}_{elative\ pronoun} \longleftarrow \text{V}_{subordinate}$	subordinate verb
Free Relative Clause	$\text{R}_{elative\ pronoun} \longleftarrow \text{V}_{subordinate}$	subordinate verb
Indirect Question	$\text{C}_{conjunction} \longrightarrow \text{V}_{subordinate}$	interrogative pronoun
Coordination	$\text{c}_{onjunct} \longleftarrow \text{C}_{conjunction} \longrightarrow \text{c}_{onjunct}$	conjunction

TABLE 5.3: Syntactic structure criteria of the FreeLing Dependency Grammars

As it has been stated at the beginning of this section, this proposal is not a theoretical reformulation of the representation of these syntactic constructions. The criteria proposed here are guidelines to provide closer syntactic representations to the semantics and to facilitate the representation of the phenomena discussed in this section in computational applications and, specifically, in dependency parsers.

5.4 DEPENDENCY RELATIONS

The dependency relations of FDGs aim to represent the same kind of information that Kaplan and Bresnan (1995) define as f-structure in the framework of Lexical Functional Grammar (LFG). Dependency relations are in terms of Kaplan and Bresnan (1995, p.4) 'meaningful grammatical relations' of the syntactic structure that express the predicate argument structure.

According to this, the repertoire of FDGs dependency relations cover the syntactic realization of the verb predicate argument structure, the relations within the rest of the phrases, the relation between a relational nexus like preposition and conjunction, the relation between particles with a grammatical function and their verbal head. Furthermore, specific labels for punctuation and elisions within the sentence.

This main schema of the dependency relations proposed in this research coincides in the majority of cases with the scope of the label set of the Universal Dependencies abbreviated as UD (Mcdonald et al., 2013). The UD labels is a fine-grained dependency relations tagset which encode the basic syntactic relations of the verb predicate including pure grammatical relations for certain relations. For example, four dependency relations are defined for subject: nominal subject (*nsubj*), passive nominal subject (*nsubjpass*), clausal subject (*csubj*), and passive clausal subject (*csubjpass*).

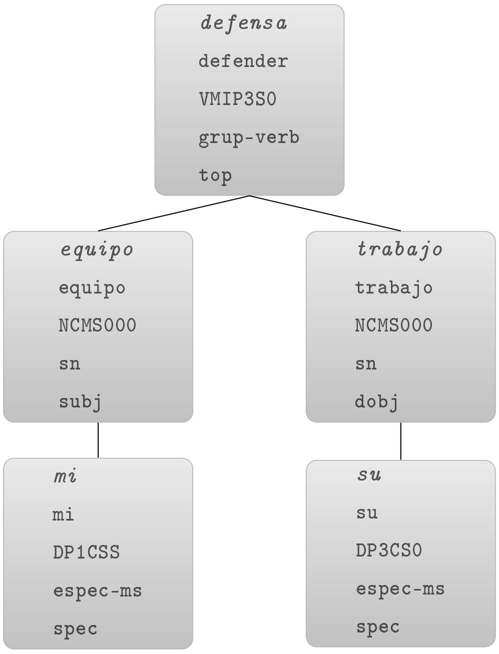


FIGURE 5.17: Dependency parse tree generated by Txala of the sentence 'Mi equipo presenta su trabajo' ('My team defend their work')

On the other hand, the FDGs dependency relations labels do not encode the grammatical distinctions of the UD because they are already present in the syntactic structure. However, the complements of a verb introduced by a preposition or by an adverb are distinguished as *pobj* and *advc* respectively. Since grammatical relations are not encoded, there is only a single label for every dependency relation in FDGs. For example, subject only is defined by a single label *subj*. The specific syntactic construction that expresses the subject is expressed by the syntactic nature of the lexical unit as in the figure 5.17 where the concept of nominal subject is expressed in two different levels, i.e., in two different tags (*sn* expresses the structural level and *subj* expresses the dependency relation itself).

In the table 5.4, the tagset of FDGs dependency relations is defined by the label name and the dependency relation name. Furthermore, the correspondence of FDGs and UD labels is added to show the granularity of both tagsets and the similarity of relations that they cover.

Some 1-to-1 correspondences are observed (e.g. *attr*, *iobj*, *voc*), although most of FDGs labels map to several labels in UD (e.g. *adjt*, *dobj*, *mod*, *subj*). In these cases, the FDGs tagset can be adapted smoothly to the UD tagset implementing heuristics. As explained at the beginning of this

FDGs label	Description	UD label
advc	adverbial complement	advcl, advmod
adjt	adjunct	advcl, advmod, nmod
agnt	agent	— — —
attr	attribute	cop
aux	auxiliary	aux, auxpass
comp	complement	case, mark
— — —	coordinating conjunction	cc
coor	coordination	conj, list, parataxis
— — —	dislocated elements	dislocation
— — —	unbounded dependencies	goeswith
dobj	direct object	ccomp, dobj, nummod, parataxis, xcomp
— — —	foreign words	foreign
gap	elision	remnant
iobj	indirect object	iobj
mod	modifier	advmod, amod, appos, compound, name, nmod
modnomatch	no rule matching	dep
modnorule	no rule in the grammar	dep
mphes	pronominal particle *es*	— — —
— — —	multiword expression	mwe
pobj	prepositional object	nmod
pred	predicative	xcomp
prt	particle	compound
punc	punctuation	discourse, punct
— — —	reparation of overridden text	reparandum
spec	specifier	neg, nummod, det
subj	subject	csubj, csubjpass, nsubj, nsubjpass
top	root of the sentence	root
voc	vocative	vocative

TABLE 5.4: Map of correspondences of FDGs labels and UD labels

section, while UD labels merge structure and relations information in a single tag, FDGs keep both levels separate, but they are always present in the dependency tree by means of distinct labels. Therefore, most FDGs labels can be translated to the UD format by merging syntactic structure and dependency relation label. In the case of a subject (labelled in the FDGs as *subj*), it can be translated to the UD as a nominal subject (*nsubj*) by checking that the syntactic structure tag corresponds to a noun phrase (*sn*).

In addition to the formal translation of labels, labels of both tagsets refer to different criteria. The harmonization of criteria can be also handled by heuristics to capture the uses of labels that should match. This is the case of the UD relation *nmod*, for example. This label is used to mark oblique nominal argument or adjunct when it depends from a verb predicate, or nominal modifier when its head is a noun phrase, adjective phrase or an adverb. These five contexts correspond to three contexts in the FDGs tagset: *pobj* (argument) and *adjt* (adjunct) in the verb predicate, and *mod* in the rest of uses. Consequently, the map from FDGs labels to UD labels needs to capture these three contexts by stating: (1) transform *pobj* label to *nmod*, (2) transform *adjt* label to *nmod*, and (3) transform *mod* label to *nmod* if the head of the relation is a noun phrase, an adjective phrase or an adverb.

On the other hand, other labels' correspondences cannot be performed directly or they do not have a correspondence because of a different syntactic structure criterion. The former case refers to UD relations like *nmod* or *cop* in which the head of the relation is a lexical unit (noun phrase of oblique arguments, adjuncts or modifiers, and complement of a copulative construction, respectively) and the child is a functional unit (preposition and copulative verb, respectively). The latter concerns relations like *cc* which is assigned to coordinating conjunction embedded in the first conjunct. In FDGs, this label is not needed because the coordinating conjunction is the head of the coordinating construction. Furthermore, the relations of multiword *mwe* and *foreign*, which are analyzed in the morphological module of FreeLing, and *dislocation* and *goeswith*, which receive the same label as the other discontinuous node which it is linked to, also belong to this group.

Finally, there is a label that is not defined in the FDGs tagset (*reparandum* which is used to express speech disfluencies), but it can be added. On the other hand, two FDGs labels are not in the UD tagset: the agent expressed by a prepositional complement in the passive voice (*agent*) and the pronominal particle (*mphes*) of the pronominal, impersonal and passive constructions.

Concerning the FDGs, the dependency relations are classified according to the tree structure level in which the dependency relation occurs, and to the head or child properties. Following this principle, the labels refer to dependency relations at sentence level, at verb predicate level, and at phrase level distinguishing between the relations that occur in the verb phrase, on the one hand, and that are concerned with the noun and adjective phrases. Furthermore, some functional relations are also encoded and a special label is used in coordinating construction. Finally, unlabelled dependencies also have a label assigned because all the connections in the dependency tree need a label in the Txala Parser which help in the process of debugging.

The following is a summary of the main characteristics of the dependency relations defined in FDGs.

Sentence

top

Top. Head of the sentence.

(30) Plou $_{top}$
'[It] rains'

punc

Punctuation. Relation between a node of the tree structure and a punctuation mark.

(31) Plou **–** punc → .

gap

Elision. Label to mark the elision of a phrase, a clause or part of a sentence.

(32) Los niños comen pollo y **–** gap → *los adultos paella de marisco* 'The kids eat pollo and the adults seafood paella'

Verb Predicate

subj

Subject. Label to express the relation between a verb and a noun, infinitive clause or finite clause which express the semantic roles agent, experiencer or theme. In the passive sentence, the subject expressing an agent is a patient in the active voice.

(33) *Mi amigos* ← subj **–** salen
'My friends go out'

dobj

Direct object. Syntactic representation of an internal argument that expresses the theme or the patient of the verb predicate.

(34) Ya acabé **–** dobj → *los estudios*
'Ø$_{1st}$ finished the studies already'

pobj

Prepositional object. Syntactic representation of a prepositional argument whose preposition is determined by the verb sense.

(35) Crec **–** pobj → *en miracles*
'Ø$_{1sg}$ believe in miracles'

advc

Adverb complement. Syntactic representation of an adverbial argument.

(36) Están – advc → *aquí*
‘Ø_{3pl} are here’

pred

Predicative. Label to represent syntactically an adjectival argument.

(37) Es veu – pred → *bonica*
‘Ø_{3sg} sees herself beautiful’

attr

Attribute. Relation between a copular verb and its complement.

(38) Es – attr → *grande*
‘Ø_{3sg} is big’

iobj

Indirect object. Syntactic realization of the semantic role benefective.

(39) *M’* ← iobj – agrada el llibre
‘I like the book’

agnt

Agent. Complement of a passive verb expressing the agent and introduced by a preposition (‘por’ in Spanish and ‘per’ in Catalan).

(40) Es cuestionado – agnt → *por sus compañeros*
‘Ø_{3s} is questioned by his colleagues

adjt

Adjunct. Syntactic realization of an optional circumstance of the verb (e.g. manner and time), or realization of discursive aspects like interjections and discourse markers.

(41) Ve – adjt → *amb els seus amics*
‘Ø_{3sg} comes with [his/her] friends’

voc

Vocative. Syntactic realization of the discourse addresser and the main verb.

(42) No lo sé **–** voc → , *Eva*
'Ø$_{1sg}$ do not know, Eva'

Verb Phrase

aux

Auxiliary. Auxiliary verb.

(43) *Ha* ← aux **–** llegado
'Ø$_{3sg}$ has arrived'

prt

Particle. Label for modal and raising verbs.

(44) *Cal* ← prt **–** anar ràpid
'Ø$_{impersonal}$ need to go fast'

spec

Specifier. Label for specific adverbs that specify the meaning of the verb and also for negation. *No* ← spec **–** tiene el certificado
'Ø$_{3s}$ does not have the certificate'

mphes

Morpheme 'es'. Label for particle 'es' (e.g. pronominal, impersonal and passive).

(45) *Es* ← mphes **–** desespera
'Ø$_{3sg}$ gets desperated'

Noun Phrase and Adjective Phrase

mod

Modifier. Modifier of noun phrase and adjective phrase which can be a noun phrase, an adjective phrase, a prepositional phrase, a relative clause, a substantive clause, and a participle clause. In this version of the FDGs, the distinction between complements and modifiers is not handled, hence, both dependency relations are grouped under the same tag.

(46) Es capaz **–** mod → *de trabajar hasta la madrugada*
'Ø$_{3s}$ is able to work until late night'

spec

Specifier. Relation between a nominal or adjectival head and a child that specifies its meaning. Prepositional phrases and adverbs are also allowed to contain a specifier.

(47) Es *un* ← spec – dia per recordar
'Ø$_{3sg}$ is a day to remember'

Functional Relations

comp
Complement. Label to express the functional relation of a prepositional head or a conjunction and its child.

(48) Es de – comp → *mentira*
'Ø$_{3sg}$ is fake'

(49) Diu que – comp → *vindrà*
'Ø$_{3sg}$ says that Ø$_{3sg}$ is coming'

Coordination

coor
Coordination. Label to express the relation between the head of a coordination construction (conjunction headed) and its conjuncts.

(50) Llegeix *novela de ficció* ← coord – i – coord → *biografies*
'Ø$_{3sg}$ reads fiction novels and biographies'

Unlabelled Dependencies

modnomatch
This label is assigned when the parser recognizes some candidate rules, but none of them are assigned because they violate some of the constraints.

modnorule
This label is assigned when the grammar does not contain any rule to express the dependency relation expressed in the dependency arc.

5.5 Grammar Development

The Catalan and Spanish versions of the FDGs have been developed manually by a bilingual linguist in Catalan and Spanish. The progression of both versions has been handled in parallel always working first in Spanish. Once the addition of rules for a particular syntactic phenomenon in Spanish is tested and the good performance of the grammar is guaranteed, the changes are applied next in the Catalan grammar. This method was possible to implement in the development of the FDGs because the languages of the FDGs are very close languages among the Romance languages and, specifically, both of them have a wide range of syntactic phenomena in common.

Therefore, a set of rules for describing a particular syntactic construction in Spanish is likely to be implemented identically for describing the same syntactic construction in Catalan by adapting particular minor language specific features in some cases.

The FDGs are built iteratively starting with covering the basic syntactic constructions and progressively incrementing the complexity of the syntactic construction. As described at the beginning of the chapter (§5.1), the input of the grammar corresponds to partially analysed syntactic trees resulting from the FreeLing chunker Civit (2003). Specifically, these syntactic analyses are unambiguous syntactic chunks (e.g. basic attachments within a chunk like a determiner or adjective and a noun) that the FDGs need to complete by solving syntactic ambiguities.

According to this principle and the syntactic criteria designed in this proposal (§5.3), the attachments within the phrase that the chunker could not solve have been treated. Then, the attachments between non-verbal phrases have been covered (noun phrase, adjective phrase, prepositional phrase and adverb). Once the phrases have been solved, the basic verb predicate structures are handled taking into account the active and the passive voice. After setting the rules of the verb predicate, rules for dealing with the non-finite and finite clauses have been added. Finally, the coordinating construction has been considered by covering structures where the conjuncts are of the same nature. Additionally, a small set of interrogative construction rules have been introduced to deal with the most basic interrogative sentences.

Alongside the attachment rules development, the part of the grammar responsible for assigning dependency relations labels has also been developed. The most frequent word order configurations in Spanish and Catalan have been encoded in the rules. Furthermore, some order alterations have been handled (e.g. clitics, specific subject inversions, shifts between direct object and indirect object positions, interrogative constructions handled by the attachment rules). Specifically, a great effort has been done in order to recognize the dependency relations of the prepositional attachment in the verb predicate (adjunct, prepositional object, direct object –only in Spanish–, indirect object, and predicative). In addition, a great deal of work has been also done in the recognition of finite and non-finite clauses (e.g. subject, direct object and adjunct).

Every iteration is performed using a development corpus. This corpus is a collection of linguistic examples (809 Spanish examples and 50 Catalan examples) selected from electronic newspapers in Spanish and Catalan available online (El Periodico[3] and La Vanguardia[4]), or intuitively added. For every construction covered by the grammar, there is at least one example collected. Furthermore, the corpus contains several configurations of the same construction in order to ensure the coverage of every phenomenon handled.

Once the grammar contains the set of rules that cover a scope of the target construction, they are tested by analysing the development corpus with the Txala Parser. If the grammar provides an acceptable analysis for the construction developed, the iteration is finished and the development process shifts to the next iteration. In those cases where the parser does not perform successfully the target construction, parallel examples are added in the development corpus and the rules

[3] http://www.elperiodico.com/

[4] http://www.lavanguardia.com/

are modified and tested until the solution is acceptable.

Recapitulation

This chapter was dedicated to the development of the FreeLing Dependency Grammars in Spanish and in Catalan. The NLP pipeline and the parser in which they are implemented have been described (§5.1). Also, the architecture of the grammar and its available features have been presented (§5.2). The second part of the chapter has focused on the syntactic criteria designed for these grammars in order to represent certain controversial syntactic phenomena (§5.3), and the labels proposed for these grammars to represent the dependency relations (§5.4). In the last section, the development method has been described (§5.5).

CHAPTER 6

DEPENDENCY GRAMMARS EVALUATION

In software development, the evaluation consists of measuring the data processed in order to determine the performance of the software and to assess the suitability of the software in solving a given problem. The NLP community is aware of this necessity, so the evaluation of NLP tools is a common practice in this area. Without an empirical validation of the linguistic data processed by these tools, their performance cannot be measured, neither the progress in the area can be proved.

In particular, evaluation has been a relevant task in the area of parsing for two decades (Lin, 1998b; Tapanainen and Järvinen, 1998; Collins, 2000; Yamada and Matsumoto, 2003; Bick, 2006). Despite this, the editions of the SIGNLL Conference on Computational Natural Language Learning (CoNLL) in 2006 and 2007 have a special importance. They promoted a campaign for dependency parsing and consolidated the methods for parsers evaluation (Buchholz and Marsi, 2006; Nivre et al., 2007). Later on, a parser has rarely been released without being empirically validated and the evaluation method proposed in CoNLL has been used extensively by the parsing community.

Our proposal for two dependency grammars for Spanish and for Catalan supports the necessity of an empirical evaluation task. Actually, the claim in this chapter is that evaluation task is essential to grammar development and it has to be performed in order to assess the progress of the grammar. Furthermore, this proposal aims to prove the importance of a evaluation task, which is demonstrated in the chapters 7 and 8.

In order to observe the importance of evaluation, this chapter focuses on a design of an empirical method to validate the FDGs. Firstly, the most frequent evaluation methods used in NLP are described in order to argue for the method proposed here (§6.1). Next, the evaluation task of the FDGs is presented (§6.1), defining the metrics used to compute the performance of the Txala Parser running with the grammars (§6.2) and the data used to carry out this task (§6.3). More specifically, the section §6.3 is concerned with explaining the data used to carry out an evaluation task from a quantitative point of view (§ 6.3.1) and a qualitative point of view (§ 6.3.2).

Finally, an evaluation task of the Spanish and Catalan FDG (§5) is performed (§6.4) using the established methodology, the metrics and the resources developed to perform a quantitative and qualitative evaluation.

6.1 Evaluation Methods

An evaluation task makes use of a particular approach according to the set-up of the evaluated system and goal that is to be achieved by the evaluation. Consequently, several approaches have been used in the evaluation of NLP tools with regard to these two aspects.

One of the main initiatives on evaluation in Europe was the project EAGLES (King et al., 1996), which was intended to establish the basis for evaluation of Natural Language Processing tools. It distinguishes three types of evaluation: progress evaluation, adequacy evaluation and diagnostic evaluation. The **progress evaluation** validates a NLP tool's performance comparing it to a desired target. The **adequacy evaluation** measures the adequacy of a NLP tool according to its use. The **diagnostic evaluation** assesses the output errors of a NLP tool and the factors that caused these errors.

Paroubek et al. (2007) review the main works on NLP which focused on evaluation to some degree such as EAGLES (King et al., 1996), ELRA (Choukri and Nilsson, 1998), SENSEVAL (Edmonds and Kilgarriff, 2002), and SEMEVAL (Agirre et al., 2007). Furthermore, Paroubek et al. (2007) update the framework on evaluation approaches, which are summarized as follows:

Intrinsic versus Extrinsic
An intrinsic approach performs the evaluation by assessing an isolated NLP tool and compares its performance to a gold standard (Sparck Jones and Galliers, 1996). On the other hand, an extrinsic approach validates the performance of a system when it is embedded in another system.

Black-box versus Glass-box
The black-box approach focuses on such aspects as the accuracy of a system, the quality of the results and the speed. The glass-box approach provides information about the design of a system, the algorithm implemented and the resources integrated.

Objective versus Subjective
The evaluation is considered objective when the task is performed by measuring statistically the data processed by an evaluated tool. Otherwise, it is a subjective evaluation which is based on the human perception of the results of such a tool.

Quantitative versus Qualitative
A quantitative evaluation measures some aspect of the performance of a tool, whereas a qualitative approach looks at the cases where an evaluated tool fails and explains the causes of the failures.

Technology-orientated versus User-orientated
When an evaluation task assesses the performance of a tool on a generic task, the approach is technology-orientated. On the other hand, in cases where the use of a tool by real users is evaluated, the approach is user-orientated.

The evaluation method followed in evaluation contests like CoNLL 2006 (Buchholz and Marsi, 2006) and 2007 (Nivre et al., 2007) is based on the comparison of a NLP tool output (i.e., dependency trees generated by a parser from a test corpus in the CoNLL campaigns) commonly called **hypothesis**, and the representations of a gold standard (i.e., dependency trees manually annotated of a corpus in the CoNLL campaigns) which are used as a **reference** (Mitkov, 2003).

The main goal of the evaluation of FDGs is to measure and describe the failures of the grammars, so this task is oriented towards a diagnosis of their progress. In order to carry out a task according to these goals, the evaluation has been designed as follows:

Intrinsic. An isolated evaluation of the Txala Parser working with the FDGs.

Black-box. Evaluation of the performance focusing on accuracy and error analysis.

Objective. The task is carried out by using statistical measures that inform about the accuracy and point to the nature of the errors.

Quantitative and Qualitative. Evaluation campaigns have been frequently only focused on quantitative analysis. In order to provide a global interpretation of the results, a quantitative and qualitative analysis are required. The implementation of both approaches in the FDGs evaluation task is described in detail in sections 6.3.1 and 6.3.2.

Technology-orientated. The evaluation focuses on validating the results of the parser itself.

In order to fulfil the goals, the methodology established in the CoNLL 2006 and 2007 campaigns (Buchholz and Marsi, 2006; Nivre et al., 2007) has been followed in the quantitative and qualitative approaches by measuring using the metrics proposed in these contests (§ 6.2). Therefore, the dependency trees generated by the Txala Parser (hypothesis) are compared to a gold standard (reference). In particular, two gold standards have been used as a reference, one is orientated to carry out a quantitative analysis (§6.3.1) and another is for handling a qualitative analysis (§6.3.2).

6.2 Evaluation Metrics

Accuracy metrics used in the FDGs evaluation are statistical measure to assess the performance of the Txala Parser when running with the grammars, and they provide an objective measure to compare the hypothesis (i.e. the system output by means of automatically generated dependency trees) and the reference (i.e. gold standard containing a desired analysis by means of manually annotated dependency trees). They have been computed using the CoNLL evaluation script Nivre et al. (2006) which compares an hypothesis with regard to a reference.

Specifically, the metrics used to evaluate the performance of the FDGs are accuracy, precision, recall and F1, which are defined by the following formulas:

Accuracy

$$\begin{aligned} \text{LAS} &= \frac{\text{correct attachments and labellings}}{\text{total tokens}} \\ \text{UAS} &= \frac{\text{correct attachments}}{\text{total tokens}} \\ \text{LAS2} &= \frac{\text{correct labellings}}{\text{total tokens}} \end{aligned}$$

LAS looks at the performance of both syntactic structure and dependency relations, and calculates the proportion of tokens in a sentence for which a system predicts the right head and dependency relation. *UAS* focuses exclusively on the syntactic structure, and measures the proportion of tokens in a sentence for which a system predicts the right head. *LAS2* takes care of the performance on dependency relations assignment, and computes the proportion of tokens in a sentence for which a system predicts the right dependency relation.

Precision

$$\text{P} = \frac{\text{system correct tokens}}{\text{system tokens}}$$

Recall

$$\text{R} = \frac{\text{system correct tokens}}{\text{gold tokens}}$$

F-measure

$$\text{F1} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

P is the ratio between correct tokens of a sentence retrieved by a system and the total number of tokens in a sentence retrieved by the system (i.e. predicted and unpredicted tokens). On the other hand, *R* corresponds to the ratio between correct tokens of a sentence retrieved by a system and the total number of tokens of a gold standard. *F1* is commonly interpreted as an average of the precision and recall measuring in a scale from 0 to 1 in which 0 is the worst score and 1 is the best score.

6.3 Evaluation Data

Traditionally, two methods of analysis have been defined for evaluation purposes: quantitative analysis (§6.3.1) and qualitative analysis (§6.3.2). Both approaches are complementary and they can contribute to a global interpretation of a NLP tool's predictions as shown in this section.

The main difference is that quantitative analysis relies on statistically informative data, while qualitative analysis talks about the richness and precision of the data (McEnery and Wilson, 1996).

Representativeness by means of frequency is the main feature of quantitative studies. That is, the observed data cover the most frequent phenomena of the data set. Rare phenomena are considered irrelevant for a quantitative explanation. Thus, quantitative descriptions provide a close approximation of the real spectrum.

Qualitative studies offer an in-depth description rather than a quantification of the data (McEnery and Wilson, 1996). Frequent phenomena and marginal phenomena are considered items of the same type because the focus is on providing an exhaustive description of the data.

In terms of methods of analysis and databases, two resources have been widely used: corpora and test suites. Language technologies find these resources a reliable evaluation test because they are coherent and they are built on guidelines.

A corpus contains a finite collection of representative real linguistic utterances that are machine readable and act as a standard reference of the language variety that is represented in the resource itself (McEnery and Wilson, 1996). From this naive conceptualization, Corpus Linguistics takes the notion of representativeness as a presence in a large population of linguistic utterances, where the most frequent utterances are represented as a simulation of the reality and are annotated according to the resource goals. That is why corpora are appropriate test data for quantitative studies.

On the other hand, test suites are structured and robust annotated databases which store an exhaustive collection of linguistic utterances according to a set of linguistic features. They are built over a delimited group of linguistic utterances where every utterance is detailed and classified according to rich linguistic and non-linguistic annotations (Lehmann et al., 1996). Thus, the control over test data and their detailed annotations make test suites a perfect guidance for qualitative studies.

Corpora have also been used in qualitative analysis, but they collect representative linguistic utterances by means of frequency rather than representative linguistic utterances by means of exhaustiveness. As a result, they are not the most appropriate tool for qualitative studies.

6.3.1 Quantitative Analysis

As argued above, corpora and quantitative analysis are closely linked. Because of the explosion of new techniques in language technologies, big collections of linguistic data are needed for development and evaluation purposes, such as the British National Corpus of 100 million words (Consortium, 2007) and Google Books of 155 billion (Davies, 2011). For this reason, initiatives to create large corpora linguistically annotated have been the focus of part of the NLP community, such as PennTreebank (Marcus et al., 1993) and OntoNotes (Weischedel et al., 2011). Simultaneously, recent advances in NLP tools have sped up the creation of large language corpora by automatizing and simplifying some parts of the process of manual linguistic annotation.

As observed when talking about the distribution of NLP tools and resources according to languages in §2.3, Spanish and Catalan have a smaller presence in NLP compared to languages like English. Despite this, there are some open-source treebanks in Spanish available and suitable for carrying out quantitative evaluation tasks of parsers: AnCora Dependencies (Taulé et al., 2008), AnCora-UPF (Mille et al., 2013), SSyntSpa Corpus (Kolz et al., 2014) and Tibidabo Treebank

(Marimon et al., 2014). Concerning Catalan, only AnCora Corpus is available.

AnCora Dependencies

This is a version of AnCora Corpus manually annotated with constituents (Taulé et al., 2008) and automatically transformed to dependency format (Civit et al., 2006). AnCora contains 500,000 words in the Spanish and Catalan versions of the corpus. The collection of the Spanish data comes from several sources: 75,000 words of Lexesp (Sebastián et al., 2000), 225,000 words of the EFE Spanish news agency, 200,000 words of the Spanish version of the newspaper *El Periódico*. The sources of the Catalan data are: 75,000 words of EFE news agency, 225,000 words of the ACN Catalan news agency and 200,000 words of the Catalan version of *El Periódico*. The method followed for the annotation task is semi-automatic. A first layer of morphological annotations is handled automatically using a version of the FreeLing tagger (Padró et al., 2010). To the morphological layer, the constituency-based syntactic layer is added manually by three native annotators per language.

AnCora-UPF

This corpus (Mille et al., 2013) corresponds to 3,513 sentences (100,892 tokens) from the Spanish AnCora Corpus semi-automatically annotated with the Meaning-Text Theory principles (Mel'čuk, 1988) presented in §2.1.2.1. Following the MTT, multiple layers of linguistic information are annotated (three different Part-of-Speech tags in granularity, surface syntactic dependencies and deep syntactic dependencies) and, specifically, a layer of semantic representation is built in in order to be a proper resource for NLP tools using dependencies. The annotation is carried out is semi-automatic. Every layer is the result of a manually defined mapping grammar that establishes the correspondences between linguistic annotated layers. The mapping is performed automatically and manually reviewed.

AnCora Surface Syntax Dependencies

Kolz et al. (2014) automatically adapted the Spanish AnCora Corpus formatted in syntactic constituents. As a result, they obtained the 547,723 tokens of AnCora annotated with surface syntactic dependencies following purely syntactic criteria and labelling the dependency relations with the labels of the Standford tagset (De Marneffe and Manning, 2012), which are the source for the Universal Dependencies (Mcdonald et al., 2013). In order to map from constituents to dependencies, a set of linguistic rules to detect the head of each constituent is created following the strategy initially stated by Magerman (1994). As a consequence of this method, Kolz et al. (2014) need to overcome flat structures where more than one constituent is a head candidate. They propose specific rules that insert the sequence of constituents as dependants of one of the constituents. On the other hand, AnCora dependency relations mix pure dependency relations and constituents tags. For this reason, these authors overcome this problem by mapping automatically these relations using information about Part-of-Speech tags and manually assigned argument structure tags of AnCora Corpus.

Tibidabo Treebank

This Spanish corpus takes a sub-set of sentences of the AnCora Corpus (Taulé et al., 2008) which correspond to 41,620 tokens. The construction of this resource follows a semi-automatic strategy automatically annotating tagged sentences (Padró et al., 2010) with the HPSG grammar SRG (Marimón, 2010). The output contains the first 500 trees ranked by a maximum entropy method converted to the dependency format. Finally, a expert annotator sets the best syntactic dependency analysis for each sentence.

Because there is only one corpus available for the two languages of this proposal, the AnCora Corpus (Taulé et al., 2008) was used as a resource for the quantitative evaluation task, initially. However, prior to the evaluation, a process of linguistic criteria harmonization is needed to overcome the differences between the corpus and the FDGs (§6.3.1.1).

6.3.1.1 Mapping of AnCora Corpus

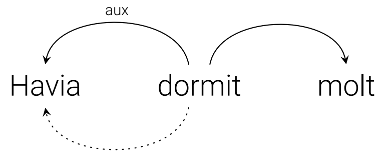
In this proposal, the same strategy for transforming from constituency format to dependency format proposed by Gelbukh et al. (2005) and Civit et al. (2006) was followed. Both works convert the format of the constituents of AnCora Corpus by a set of heuristics that marks the head of the constituent by matching patterns. The task of adapting linguistic criteria is analogous to the format transformation since detecting a head and reassigning a new head is basically based on the same operation. Therefore, this proposal creates and implements heuristic rules in order to deal with corpus conversion to the FDGs linguistic criteria (§5.3).

In this section, the main criteria of both resources are described showing the similarities and the differences. The linguistic criteria established for the Spanish corpus are the same as the criteria for the Catalan version of the corpus. Consequently, the transformations applied in one language can be directly implemented in the other language. The criterion followed in the AnCora Corpus is expressed in the upper arcs of every example, and the criterion in FDGs corresponds to the dotted lower arcs.

Periphrastic Verbs

Both resources place the auxiliary verb and the modal verbs as dependent nodes of the non-finite verb form which is expressing the semantic content.

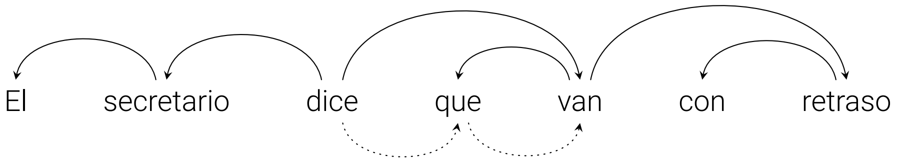
(1) *Havia* dormit molt
‘Ø$_{1st/3sg}$ had slept a lot’



Subordinating Conjunctions

While AnCora considers that the verb of the subordinate clause is the head and the conjunction is the child, the FDGs states that the head of the subordinate clause is the conjunction and the verb of the subordinate clause is a child of the conjunction.

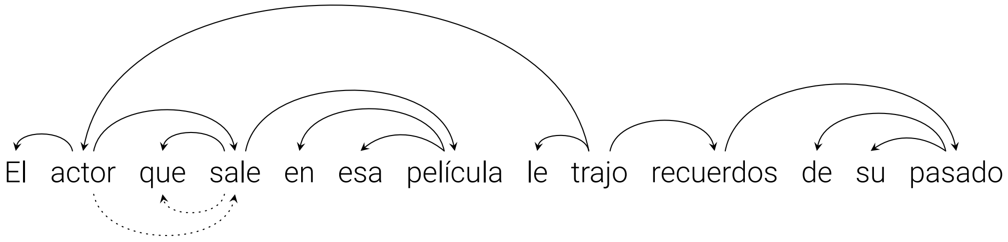
(2) El secretario dice *que van con retraso*
'The secretary says that Ø$_{3pl}$ are late'



Relative Clauses

Both resources analyse this construction in the same way, i.e., the head is the subordinate verb and the relative pronoun a child.

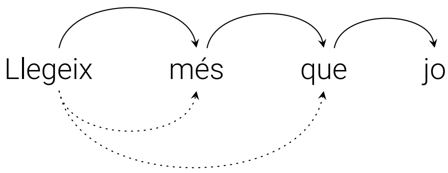
(3) El actor *que sale en esa película* le trajo recuerdos de su pasado
'The actor who plays a role in that film brought to him memories from his past'



Comparative Structures

Some comparative constructions are formed by two parts which correspond to the compared elements. In the AnCora corpus, the second part of the comparison is embedded in the first part, while in FDGs both parts depend at the same level on the head which the comparison is attached to (4).

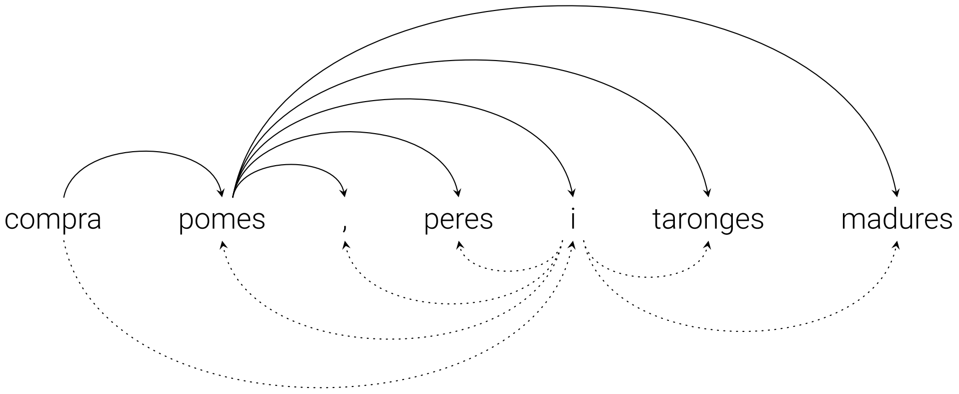
(4) Llegeix *més que jo*
'Ø$_{3pl}$ reads more than me'



Coordinating Constructions

Ancora treats the first conjunct as the head, and the rest of conjuncts, the punctuation and the coordinating conjunction are dependent on the first conjunct. In FDGs, the coordinating conjunction is the head, hence, the conjuncts and the punctuation are dependent on the conjunction.

(5) compra pomes, peres i taronges madures
‘$\emptyset_{3sg}$ ripe apples, pears and oranges’



After establishing the basic mappings for these criteria an experimental evaluation was performed (Lloberes et al., 2010). For that task, 25 sentences of the Spanish AnCora Corpus were selected randomly and they were analysed by the Txala Parser using the Spanish FDG. Then, the 25 sentences of the corpus and the same 25 sentences analysed by the parser were compared with the metrics of the CoNLL 2006 (Buchholz and Marsi, 2006) as shown in the table 6.1.

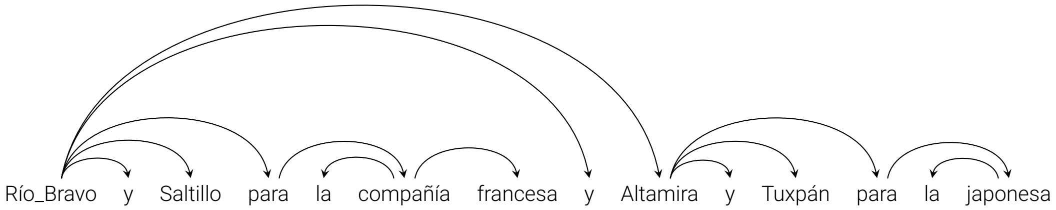
Corpus	**LAS**	**UAS**	**LAS2**
25 sentences	73.88	81.13	78.81

TABLE 6.1: Spanish FDG accuracy scores

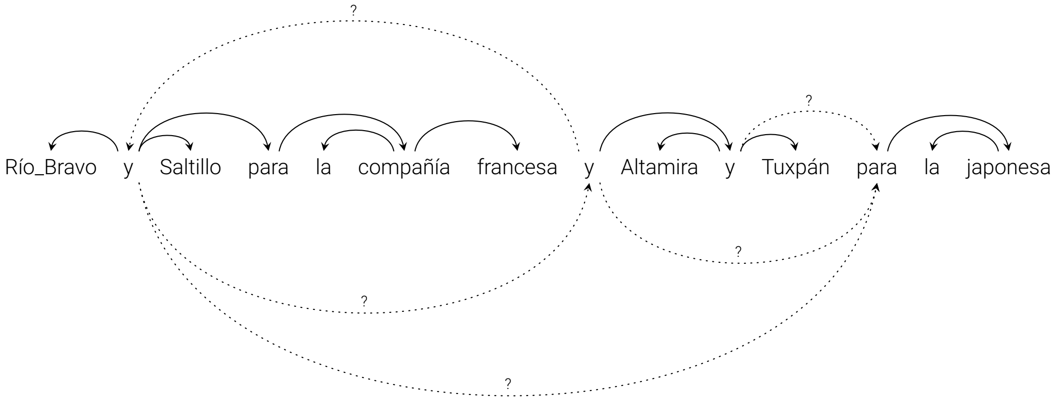
A close look at the data reveals two major problems not resolved in the mapping. On the one hand, coordinating constructions embedded in other coordination structures cannot be converted because there is not enough information to decide which node should be the head of the construction (6) since it could be the first coordinating conjunction or the second one (6). Furthermore, another ambiguity surfaces when mapping both resources. The shared modifier ‘para la japonesa’ admits three possible attachments (i.e. child of the first coordinating conjunction, child of the second one or child of the third one), as shown in the example (6).

(6) Río Bravo y Saltillo para la compañía francesa y Altamira y Tuxpán para la japonesa
‘Río Bravo and Saltillo for the French company and Altamira and Tuxpán for the Japanese one’

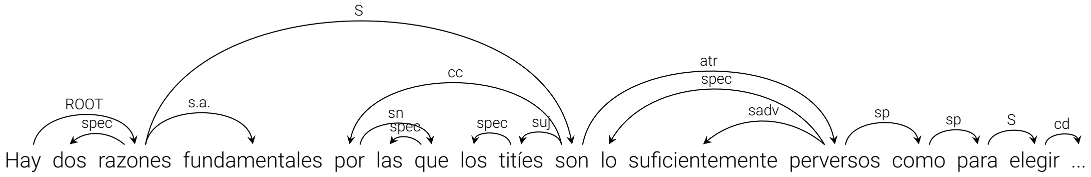
a. Analysis in AnCora


Río_Bravo y Saltillo para la compañía francesa y Altamira y Tuxpán para la japonesa

(6) Possible maps to FDGs


?
?
Río_Bravo y Saltillo para la compañía francesa y Altamira y Tuxpán para la japonesa
?
?
?

(7) Hay dos razones fundamentales por las que los tities son lo suficientemente perversos como para elegir la poliandria
'There are two fundamental reasons why marmosets are wicked enough to choose polyandry'


S
atr
cc
spec
ROOT
s.a.
sadv
sp
spec
sn
spec
spec
suj
sp
S
cd
Hay dos razones fundamentales por las que los tities son lo suficientemente perversos como para elegir ...

On the other hand, the dependency relations mix pure dependency relations (e.g. *suj*, *cd*, *atr*) and constituency tags (e.g. *sn*, *sp*, *s.a.*, etc.) as shown in the example (7). While some of the labels can be translated directly (e.g. *spec,* f), there are labels that have more than one use (e.g. *sp*), which in the end is a blockage for the translation of the dependency labels.

For these reasons, this proposal abandoned the conversion of the AnCora Corpus to the FDGs linguistic criteria, but continued applying the methodology proposed with another resource as will be explained in §6.3.1.2. As a consequence, it has to be assumed that the Catalan FDG will not be evaluated since AnCora is the only corpus available in Catalan for now.

6.3.1.2 Mapping of Tibidabo Treebak

The Tibidabo Treebak (Marimón, 2010) has been used as a resource for quantitative evaluation. Since this corpus is analysed by a Spanish HPSG grammar (Marimón, 2010), the head is already marked, which simplifies the conversion of controversial syntactic construction representations. Furthermore, the syntactic trees of the corpus are labelled with a syntactic motivated dependency relation in comparison with the AnCora dependency labels (Taulé et al., 2008).

On the other hand, the methodology followed to develop this corpus is by selecting manually the best parse tree generated by the HPSG grammar which ensures the quality of the resource, while in an automatic methodology the same quality is not guaranteed.

The whole corpus contains 589,542 tokens, but in this proposal only the tokens from manually reviewed sentences are selected, which corresponds to a sub-corpus of 45,215. This resource has been encoded by different syntactic structure criteria and uses different labels for syntactic relations. For this reason, it is also needed a conversion of the corpus to the FDGs criteria.

For this conversion, a set of rules is created and works in two steps. Firstly, the structures that are differently represented are handled by a set of rules that identify the syntactic head. Later, the label of the syntactic relation is translated. The rules are applied sequentially, i.e., the first match is applied.

The rules to reassign a syntactic head are applied if the pattern described in a rule matches the sub-tree of the parse tree. In total, 18 rules have been created for dealing with the different structure configurations. In general, the criteria of the Tibidabo Treebak coincide, except particular syntactic constructions, as will be explained below. Consequently, these exceptions are the cases that need to be handled by the rules (figure 6.1).

```
^hd-cmp haber_v-vp_part - - 2 2 n
```

FIGURE 6.1: Rule for reassigning head

The rule exemplified in figure 6.1 reassigns the head in periphrastic verbs stating that if a participle verb form is headed by an inflected verb with the lemma ‘haber’ (*ˆhd-cmp haber_v-vp_part*) the participle becomes the head and the auxiliary ‘haber’ becomes the child of the periphrastic verb (*2*).

Concerning the rules for relabelling syntactic relations, two sub-processes are carried out. Firstly, the relations of verb arguments are handled (3124 rules) and, later, more general rules treat the rest of syntactic relations (32 rules). Both kind of rules are formed by four parts: a regular expression for the targeted tree node, conditions of the head, conditions of the child, name of head’s lexical rule, new tags for the arguments (figure 6.2).

```
^ct-hd .   .   _v_acc_dlr dobj
```

FIGURE 6.2: Rule for relabelling syntactic relation

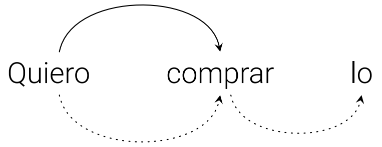
The rule of figure 6.2 is used to relabel the relation of a verb containing a clitic (*ˆct-hd*) and an accusative clitic (*_v_acc_dlr*) with the tag *dobj* (i.e. direct object).

As mentioned at the beginning of this section, the Tibidabo Treebak has different criteria (Marimon and Bel, 2015) from the FDGs that need to be harmonized. Next, the main differences among criteria are described.

Enclitics

They are encoded in the morphology in the Tibidabo Treebak, while in the FDGs are retokenized as an independent token of the verb. In the examples, the uper thick arcs correspond to the analysis of the Tibidabo Treebak and the dotted arcs below of the sentence correspond to the FDGs analysis.

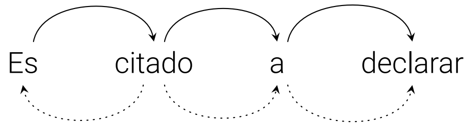
(8) Quiero comprar*lo*
'Ø$_{1sg}$ want to buy it'



Auxiliary Verbs

The Tibidabo Treebak considers the auxiliary verb as the head of the periphrastic verb and only the verb 'haber' ('to have') is accepted as auxiliar. On the other hand, the FDGs places the auxiliar in the child position of the periphrastic verb and the verb 'haber' ('to be') and 'ser' ('to be') are seen as auxiliars.

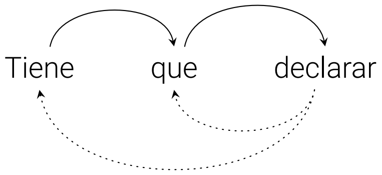
(9) Es citado a declarar
'Ø$_{3sg}$ called to declare'



Raising and Modal Verbs

The Tibidabo Treebak applies the same criteria than the auxiliary verbs, while the FDGs considers the non-finite form the head of the verbal group.

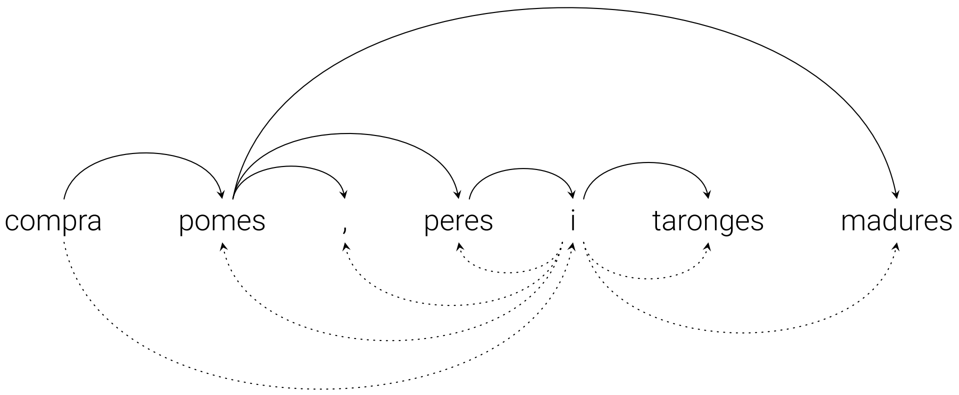
(10) Tiene que declarar
'Ø$_{3sg}$ has to declare'



Coordination

The Tibidabo Treebak strictly follows the Moscow treebanks' family criterion, i.e., analyses this construction in cascade. The FDGs follows the Prague treebanks family, so the head of the construction is the coordinating conjunction, and the conjuncts and the punctuation are dependents.

(11) compra pomes, peres i taronges madures
'$\emptyset_{3sg}$ ripe apples, pears and oranges'



All the criteria have been harmonized to the FDGs syntactic criteria proposal (§5.3). However, this does not apply to coordination, which is the only construction in which the harmonization follows the Tibidabo Treebak criteria. As observed in the example (11), there is a shared modifier of several of the conjuncts of the coordination ('madures'). However, from a representational point of view, it is not possible to know whether it modifies to the set of conjuncts or only the first conjunct. Because of this ambiguity, adapting the corpus to the FDGs criteria would create errors. For this reason, the coordination in FDGs has been changed to the criterion of the Tibidabo Treebak, accepting that the grammar will be evaluated with the coordination criterion of the Tibidabo Treebak.

With regard to the harmonization of dependency relation labels, all the labels have been transformed because the Tibidabo labels are covered by the FDGs labels. Despite of this, the correspondences are not direct in the majority of cases, so the rules have helped in detecting these cases. A summary of the dependency relations correspondences can be checked in the table 6.2.

6.3.2 Qualitative Analysis

The main aim of qualitative studies is to offer empirical evidence about the richness and precision of the data (McEnery and Wilson, 1996). For this reason, qualitative analyses are deep and detail-orientated. In this approach representativeness of the studied phenomena focuses on exhaustiveness rather than frequent data.

Test suites are controlled and exhaustive databases of linguistic utterances classified by lin-

Description	**FDGs**	**Tibidabo**
adjunct	adjt	ADV, MOD
agent	agnt	BYAG
attribute	attr	ATR
auxiliary verb	aux	AUX
complement	comp	COMP
conjunct	coor	CONJ, PUNC, ENUM
direct object	dobj	DO
elision	gap	COMP-GAP, MOD-GAP, SUBJ-GAP
head of a sentence	top	ROOT
indirect object	iobj	IO
modal verb	prt	
modifier	mod	MOD, COMP
particle *es*	mphes	IMPM, PASM, PRNM
predicative	pred	PRD, OPRD
prepositional object	pobj	OBLC,PP-DIR, PP-LOC
punctuation	punc	PUNCT
specifier	spec	SPEC, MOD
subject	subj	SUBJ

TABLE 6.2: Harmonization of dependency relations of FDGs and Tibidabo Treebak

guistic features. These collections of cases are internally organized and richly annotated (Lehmann et al., 1996). Controlledness, exhaustiveness and detailedness properties allow these databases to provide qualitatively analyzed data.

They developed in parallel with the NLP technologies. The more sophisticated the software became, the more complex and systematic the test suites turned to (Lehmann et al., 1996). Therefore, from a collection of interesting examples, they evolved into deeply structured and richly annotated databases as shown in the table 6.3, where the HP test suite (Flickinger et al., 1987), the test suite developed by one of the groups of EAGLES (EAGLES, 1994) and the TSNLP (Lehmann et al., 1996) are summarized.

The HP test suite (Flickinger et al., 1987) is an English and general purpose resource developed to diagnose and monitor the progress of NLP software development. The main goal of this test suite is to evaluate the performance of heuristic-based parsers under development. The suite contains a wide-ranging collection of linguistic examples that refer to syntactic phenomena such as argument structure of verbs and verbal subcategorization among others. It also includes some basic anaphora-related phenomena. Furthermore, these phenomena are represented by a set of

Features	**HP**	**EAGLES**	**TSNLP**
Domain	general	specific	general
Goal	parsing	grammar checkers	NLP software
Languages	English	English	English, German, French
Annotation	minimal	minimal	robust
Content	syntax	taxonomy of errors	(extra-)linguistic

TABLE 6.3: HP, EAGLES & TSNLP features

artificially constructed sentences and the annotations are shallow. This resource has a minimal internal classification since the suite organizes the test data under headings and sub-headings.

In order to take a further step, subsequent test suites have been developed as in-depth resources with rich structure and annotations. One of the groups of the Expert Advisory Group on Language Engineering Standards (EAGLES) proposes a set of guidelines for evaluating grammar checkers based on test suites (EAGLES, 1994). The test suite is a collection of attributes that makes it possible to validate the quality of the functions of the evaluated tool. It is derived from a taxonomy of errors, where each error class is translated into a feature which is collected in the test suite. The final result is a classification of sentences containing an error, the corresponding sentence without the error, the name of the error and the guidelines for the correction process.

The TSNLP (Lehmann et al., 1996) is a multilingual test suite (English, French and German) richly annotated with linguistic and meta-linguistic features. This test suite is a collection of test items with general, categorial and structural information. Every test item is classified according to linguistic and extra-linguistic features (e.g. number and type of arguments, word order, etc.). These test items are also included in test sets by means of positive and negative examples. Furthermore, the TSNLP includes information about frequency or relevance for a particular domain.

Concerning the languages of this study, a test suite for Spanish was developed by Marimon et al. (Marimon et al., 2007). The goal of this test suite is to assess the development of a Spanish HPSG grammar which offers grammatical and agrammatical test cases. On the other hand, there is no test suite for Catalan to the best of our knowledge. Because the background of test suites in Spanish and Catalan is very limited, this proposal supports the idea of creating new resources updated with the new requirements of the NLP tools. On the other hand, the test suite of Marimon et al. (2007) is restricted to the HPSG framework. However, the qualitative evaluation task presented here is a less complex process than the development of a HPSG grammar. For these reasons, in this proposal a new test suite for Spanish and the first test suite for Catalan are created in order to assist in the parsers evaluation tasks, as will be described in the next section (§6.3.2.1).

6.3.2.1 ParTes Test Suite

ParTes (Parsing Test Suite) is a test suite of syntactic phenomena for qualitative parsing evaluation freely distributed under the Creative Commons Attribution-ShareAlike 3.0 Unported License (Lloberes et al., 2014, 2015b).

This resource includes a version in Spanish (ParTesEs) and another one in Catalan (ParTesCa). Every version includes three sets of data, the test suite data, development data (syntactically annotated and unannotated) and test data (syntactically annotated and unnanotated), in which every sentence corresponds to one of the syntactic phenomena of the test suite.

Guidelines

This test suite has been developed according to the main specifications in test suite design (Flickinger et al., 1987; EAGLES, 1994; Lehmann et al., 1996). Simultaneously, a set of guidelines has been designed particularly for its construction in order to develop a coherent and useful resource.

Specific purpose
While some test suites are general purpose like TSNLP, ParTes has been originally built for evaluating the performance of parsers. Specifically, it is oriented to validate the accuracy of the dependency trees generated by the Txala Parser from a qualitative point of view. For this reason, the test cases are related to syntactic phenomena and the test suite has been annotated with several syntactic features.

Particular language representation
Because ParTes is for parsing assessment, the language representation of this resource is restricted to the syntactic level. Despite other test suites which are developed in a particular framework (Marimon et al., 2007), ParTes is independent of the linguistic theory. However, it is based on the notion of hierarchy, so the parent-child relations are expressed.

Characterization by features
ParTes is not a simple collection of linguistic test cases nor a set of linguistic features, actually. This resource contains the syntactic phenomena that configure a language defined by a set of syntactic features (e.g. the syntactic category of the head or the child, the syntactic relation with the node that governs it, etc.).

Hierarchy of syntactic phenomena
Previous test suites were a collection of test sentences, optionally structured (EAGLES and

TSNLP). ParTes proposes a hierarchically-structured set of syntactic phenomena to which tests are associated.

Polyhedral hierarchy

Test suites can define linguistic phenomena from several perspectives (e.g. morphologic features, syntactic structures, semantic information, etc.). Because ParTes is built as a global test suite, it defines syntactic phenomena from two major syntactic concepts: syntactic structure and argument order, as will be presented later.

Exhaustive test suite

In order to evaluate NLP tools qualitatively, test suites list exhaustively a set of linguistic samples that describe in detail the languages of the resource, as discussed in the explanation about resources for qualitative evaluation (§6.3.2). ParTes is not an exception and it contains an exhaustive list of the main syntactic phenomena of the languages under consideration. Despite this, some restrictions are applied to this list. Otherwise, listing the whole set of syntactic phenomena of a language is not feasible, and it is not one of the goals of the test suite's design. Consequently, the several variations of the syntactic phenomena included are not treated in the current version of the test suite.

Representative syntactic phenomena

As mentioned, lists of test cases need to be delimited because test suites are controlled data sets. Similarly to corpora development, the syntactic phenomena to be included in the test suite can be selected according to a certain notion of representativeness. Consequently, representative syntactic phenomena are relevant for testing purposes and they should be added in the test suite, whereas peripheral syntactic phenomena can be excluded. In the specifications explanation, the definition of representativeness in ParTes and how it is implemented are detailed.

Rich annotations

Every syntactic phenomenon of ParTes is annotated with precise information that provides a detailed description and that makes it possible the qualitative interpretation of the data. The annotations refer to several linguistic and extra-linguistic features that determine the syntactic phenomena.

Controlled data

As argued in §6.3.2, there is a direct relation between qualitative evaluation, test suites and controlled test data. Because ParTes is a test suite for qualitative evaluation, there is a strong control over the test data and, specifically, the control is applied in two ways. The number of test cases is limited to human-processing size. The sentences of the test cases are controlled to avoid ambiguities and interactions with other linguistic utterances. For this reason, test cases are artificially created in this version of the test suite.

```
<constituent name="nounphrase">
   <hierarchy name="child">
      <realization id="0037"
         name="prepositionalphrase"
         class="noun" subclass="prepobj"
         link="n-s" freq="0.084357"
         parent_devel="recurso"
         child_devel="para"
         parent_test="libro"
         child_test="para"
         devel="Es un recurso para los
         alumnos"
         test="Los alumnos tienen un
         libro para la lectura"/>
   </hierarchy>
</constituent>
```

FIGURE 6.3: Structure phenomena in ParTes

Semi-automatically generated
Linguistic resources usually have a high cost in terms of human effort and time. For this reason, automatic methods have been implemented whenever it has been possible. Manual linguistic description of syntactic structures has been the main method to annotate structural syntactic phenomena. On the other hand, argument order annotations have been automatically generated and manually reviewed, using the automatization process of the SenSem corpus Vázquez and Fernández-Montraveta (2015).

Multilingual
The architecture of this resource allows it to be developed in any language. The current version of ParTes includes the Spanish version of the test suite (ParTesEs) and the Catalan version (ParTesCa).

Specifications

As a result of applying the guidelines, a test suite has been created for every language. ParTesEs contains a total of 161 syntactic phenomena in Spanish (99 relate to the syntactic structure and 62 to the word order) and ParTesCa is formed by a total of 146 syntactic phenomena in Catalan (100 are concerned with the syntactic structure and 46 with the word order).

The **syntactic structure** phenomena have been manually extracted from descriptive grammars (Bosque and Demonte, 1999; Solà et al., 2002) and represented following the linguistic criteria of the FDGs (§5.3). Their representativity is validated by the relative frequency of head-child relations of the AnCora Corpus (Taulé et al., 2008) computed automatically.

```
<class name="subj#V">
   <schema name="subj#V">
      <realization id="0104"
         func="subj#v"
         cat="pron#v"
         parent="perdre"
         children="tot"
         constr="passive-pron"
         sbjtype="full"
         freq="0.001875"
         idsensem="45074#45239#48770"
         test="Tot s’ha perdut"/>
   </schema>
</class>
```

FIGURE 6.4: Word order in ParTes

At the first level of the hierarchy, phenomena are classified by taking into account if they occur inside a chunk or whether they refer to the connection between a clause marker and the verb of the clause (*level*). As shown in figure 6.3, the phrase or the clause considered in the syntactic phenomenon (*constituent*) and its position (*head* or *child*) in the structure (*hierarchy*) are described. Finally, a set of syntactic features is associated to the phrase or clause observed (*realization*).

Specifically, the syntactic features of the *realization* are concerned with the grammatical category, the phrase or the clause that defines the structure phenomenon (*name*), its syntactic specifications (*class*, *subclass*), the arc between the parent and the child (*link*), and the relative frequency of the link (*freq*) extracted from the AnCora Corpus. Additionally, every phenomenon includes a numeric *id*.

For every syntactic structure phenomenon, two linguistic examples have been associated, one of them to be used for development purposes (*devel*) and the other one for testing purposes (*test*). The lemmas of the parent and the child of the exemplified phenomenon are also provided (*parent_devel*, *parent_test*, *child_devel*, *child_test*).

Word order is built over the most frequent argument structure frames of the SenSem Corpus (Vázquez and Fernández-Montraveta, 2015).

The figure 6.4 shows that the word order’s hierarchy is structured firstly by the number and the type of arguments of the word order schema (*class*). Every class is defined by a set of *schemas* about the argument order and the specific number of arguments. The most concrete level (*realization*) describes the properties of the schema.

This set of properties refers to the syntactic function (*func*) and the grammatical category (*cat*) of every argument of the schema. Furthermore, the type of construction (*constr*) in which the schema occurs in and the type of subject (*sbjtype*) are provided. The relative frequency of the word order schema in the SenSem Corpus is associated (*freq*). In addition, a numeric *id* is assigned to every schema and a link to SenSem Corpus sentences with the same schema is created (*idsensem*).

Every schema recorded is exemplified with a sentence for testing purposes (*test*). For every test sentence, the lemmas of the *parent* and the *children* corresponding to the head of the arguments of the schema are added.

Data Sets

The development data and the test data are built over the linguistic examples of the set of syntactic phenomena of the ParTes. Up to the current version, the sentences referring to the syntactic structure phenomena are distributed as follows: 95 sentences in the development data set of ParTesEs and 99 sentences in the test data set, and 98 sentences in the development data set of ParTesCa and 99 sentences in the test data set.

ParTesEs and ParTesCa have been semi-automatically annotated by Txala Parser using the FDGs and the output has been reviewed manually by two native annotators. Both sets of data are distributed in plain text format and in the CoNLL annotation format (Nivre et al., 2007).

6.4 Evaluation Task

An evaluation task has been performed to validate the performance of the Spanish and Catalan versions of the FDGs (§5) following the methodology (§6.1), the metrics (§6.2) and the quantitative and qualitative resources (§6.3) presented in this chapter. Specifically, the dependency parse trees generated by both versions of the grammar are compared to the two gold standards, a large corpus in the quantitative analysis and a test suite in the qualitative analysis. Therefore, in the quantitative analysis, Spanish FDG dependency parse trees and Tibidabo Treebank are contrasted. In the qualitative analysis, Spanish and Catalan FDGs dependency parse trees are compared to the ParTes test data set. Furthermore, the general accuracy results, accuracy concerning the syntactic structure and the results about labelling dependency relations are presented.

Grammar	LAS	UAS	LAS2
Bare FDG	81.52	89.57	83.95

Table 6.4: Quantitative accuracy results of Spanish Bare FDG

Grammar	LAS	UAS	LAS2
Bare FDG	80.99	90.11	81.94

Table 6.5: Qualitative accuracy results of Spanish Bare FDG

Concerning the Spanish FDG, both quantitative (table 6.4) and qualitative analysis (table 6.5) shows that the grammar gets a medium-accuracy on providing a dependency analysis for the sentences of the Tibidabo Treebank and the ParTes data (*LAS*). In particular, the grammar show high-accuracy results in recognizing the dependency tree structure of a sentence (*UAS*), while it has some limitations on assigning dependency relations to dependency tree arcs (*LAS2*).

Category	Gold	System	%
determiner	7040	6854	97
noun	10475	9955	95
pronoun	2110	2009	95
adjective	2490	2326	93
negation	470	436	93
number	626	560	89
verb	6812	5988	88
conjunction	745	640	86
date	127	109	86
preposition	**5837**	**4506**	**77**
adverb	**1462**	**1102**	**75**
coordination	**1121**	**739**	**66**
interjection	**22**	**10**	**45**

TABLE 6.6: Quantitative UAS results of Spanish Bare FDG

Category	Gold	System	%
interjection	1	1	100
negation	6	6	100
number	1	1	100
determiner	82	81	99
adjective	26	25	96
pronoun	47	45	96
verb	155	145	94
noun	101	94	93
preposition	**49**	**39**	**80**
adverb	**41**	**29**	**70**
conjunction	**17**	**8**	**47**
date	0	0	0
coordination	0	0	0

TABLE 6.7: Qualitative UAS results of Spanish Bare FDG

A closer look at the accuracy of the tree structure reveals that the accuracy is not equal for all the syntactic categories (tables 6.6 and 6.7). There are categories in which the grammar recognizes better their head (e.g. determiners, nouns, pronouns, adjectives, verbs, etc.) than other ones (e.g. prepositions, adverbs, coordinating conjunctions, subordinate conjunctions).

Specifically, among the categories with a higher error ratio, the prepositional phrase attachment (PP-attachment) is not the most frequent error in FDG. However, the qualitative data of ParTes shows that the majority of errors occur in the prepositional phrase (PP) that should be attached to the noun noun phrase and when the PP is headed by a preposition other than 'de' ('of') (e.g. 'Los alumnos tienen un libro *para la lectura*', 'The sudents have a book for reading'), adjective attached (e.g. 'Eres capaz de volver a su casa', '$Ø_{2sg}$ capable *of coming back to his house*') or adverb attached (e.g. 'Estoy cerca *de ti*', '$Ø_{1sg}$ am close to you'). Similarly, adverbs are wrongly attached to verb. They should be attached to adjective (e.g. 'Algunos están *muy* decepcionados', 'Some of them are very disappointed') or to adverb (e.g. 'Es un artículo redactado *muy* rápidamente', 'It is an article written very fast').

On the other hand, coordinating structures get a low score. Although ParTes does not yet treat coordinating constructions, the observation of some examples of the Tibidabo Treebank show that the majority of errors are due to confusions about the detection of coordination. Furthermore, the data of ParTes reveals that the FDG has difficulties in recognizing the head of subordinate clauses. In particular, the majority of errors are because the FDG does not recognize properly the comparative construction.

Label	Gold	Correct	System	Precision	Recall	F1
comp	6465	6231	6336	96.38	98.34	97.35
spec	8186	7907	8050	96.59	98.22	97.40
aux	435	424	440	97.47	96.36	96.91
prt	474	446	455	94.09	98.02	96.01
top	3832	3597	3828	93.87	93.97	93.92
attr	926	811	875	87.58	92.69	90.06
coor	2228	1725	1879	77.42	91.80	84.00
subj	3250	2571	2841	79.11	90.50	84.42
mod	5650	4208	4389	74.48	95.88	83.84
mphes	476	473	718	99.37	65.88	79.23
dobj	3036	2133	2574	70.26	82.87	76.05
agnt	**59**	**37**	**50**	**62.71**	**74.00**	**67.89**
adjt	**3161**	**1688**	**2567**	**53.40**	**65.76**	**58.94**
pred	**166**	**110**	**278**	**66.27**	**39.57**	**49.55**
iobj	**342**	**99**	**123**	**28.95**	**80.49**	**42.58**
pobj	**637**	**562**	**2529**	**88.23**	**22.22**	**35.50**
gap	**14**	**0**	**8**	**0.00**	**0.00**	0.00
modnomatch	0	0	1130	NaN	0.00	0.00
modnorule	0	0	267	NaN	0.00	0.00

TABLE 6.8: Quantitative LAS2 results of Bare FDG

The assignment of dependency relations is performed successfully (tables 6.8 and 6.9). Despite of this, while the FDG recognizes some relations without difficulties (e.g. subject *subj*, attribute *attr*, specifier *spec*), it has serious problems on identifying the dependency relation performed by some verb arguments (e.g. prepositional object *pobj*, indirect object *iobj*, predicative *pred*, agent complement of passive *agnt*) and the adjunct (*adjt*).

The majority of errors are related to arguments realized with a prepositional phrase. Consequently, the FDG has to decide which of the multiple prepositional arguments is given a prepositional phrase (e.g. labelling *pobj* instead of *adjt* in 'Quiero que vengas *con amigos*', '$\emptyset_{1sg}$ want you to come with friends').

Furthermore, in the qualitative analysis, a lot of modifiers (*mod*) are erroneously detected as a consequence of the multiple errors on PP-attachment observed in *UAS* metric. Concerning the predicative complement (*pred*), there is an over-generalization of labelling rules for this dependency relation (e.g. labelling *pred* instead of *adjt* in 'Se fue acabada la fiesta', '$\emptyset_{3sg}$ left once the party finished'), so the rules need to be tightened in order to restrict their application. Finally, the qualitative data reveals that apposition needs to be covered.

Label	Gold	Correct	System	Precision	Recall	F1
top	99	95	99	95.96	95.96	95.96
spec	115	106	106	92.17	100.00	95.93
aux	11	11	12	100.00	91.67	95.65
mphes	9	7	7	77.78	100.00	87.50
attr	28	24	27	85.71	88.89	87.27
subj	51	40	43	78.43	93.02	85.10
prt	6	5	6	83.33	83.33	83.33
comp	66	46	48	69.70	95.83	80.70
dobj	39	33	49	84.62	67.35	75.00
mod	**40**	**19**	**24**	**47.50**	**79.17**	**59.38**
adjt	**39**	**17**	**26**	**43.59**	**65.38**	**52.31**
iobj	**7**	**2**	**2**	**28.57**	**100.00**	**44.44**
pred	**2**	**1**	**4**	**50.00**	**25.00**	**33.33**
pobj	**11**	**9**	**38**	**81.82**	**23.68**	**36.73**
agnt	**1**	**0**	**0**	**0.00**	**NaN**	**0.00**
apos	**2**	**0**	**0**	**0.00**	**NaN**	**0.00**
modnomatch	0	0	35	NaN	0.00	0.00

TABLE 6.9: Qualitative LAS2 results of Spanish Bare FDG

The Catalan FDG performs similarly to the Spanish FDG (table 6.10), although the scores are slightly lower than the Spanish grammar (table 6.5). In parallel to the Spanish, the grammar has a medium-high accuracy (*LAS*), high accuracy on structure rules (*UAS*) and medium-high accuracy on labelling rules (*LAS2*).

Grammar	LAS	UAS	LAS2
Bare FDG	79.41	88.24	80.88

TABLE 6.10: Qualitative accuracy results of Catalan Bare FDG

Here again prepositions, adverbs and subordinate conjunctions tend to be attached to the wrong head. The analysis of ParTes test data shows that the causes are the same as those observed in Spanish: confusions of nominal, adjectival or adverbial head in prepositions, tendency to attach the adverb to the verb (although it modifies adjectives or adverbs), and subordinate conjunctions in comparative sentences are attached wrongly because this construction is not covered enough.

Category	Gold	System	%
negation	9	9	100
number	1	1	100
interjection	1	1	100
determiner	92	88	96
noun	103	95	92
pronoun	36	33	92
verb	181	163	90
adjective	29	25	86
adverb	**34**	**27**	**79**
preposition	**42**	**30**	**71**
conjunction	**16**	**8**	**50**
date	0	0	0
coordination	0	0	0

TABLE 6.11: Qualitative UAS results of Catalan Bare FDG

With regard to labelling dependency relations, the behaviour of the Catalan grammar is parallel to the Spanish one as well, but the drop in the accuracy is concentrated only in the recognition of the prepositional phrase as an argument (*pobj*) or adjunct (*adjt*), and the identification of the adjective phrase as argument (*pred*) or adjunct (*adjt*).

Although Spanish and Catalan FDGs perform successfully, they need to be improved to handle some linguistic phenomena concerning syntactic structure and the dependency relations in order to get high accuracy scores. In particular, among the syntactic structure phenomena, the PP-attachment, adverb attachment, the coordinating constructions and the comparative clauses all need to be improved. On the other hand, the recognition of prepositional arguments and adjectival argument recognition should be handled better in the rules. Specifically, the two following chapters are focused on methods for solving the PP-attachment (§7), on the one hand, and for improving argument recognition on the other (§8).

Recapitulation

This chapter has focused on the evaluation of the performance of dependency parsing. The revision of the main methods applied in the NLP tools assessment allowed us to set the appropriate method to validate the accuracy of FDGs (§6.1). Furthermore, the statistical metrics used in the FDGs evaluation approach have been presented (§6.2). At the same time, the quantitative and qualitative linguistic resources available in Spanish and in Catalan have been described (§6.3).

Label	Gold	Correct	System	Precision	Recall	F1
spec	121	109	109	90.08	100.00	94.78
mphes	9	9	10	100.00	90.00	94.74
aux	37	33	34	89.19	97.06	92.96
top	98	89	99	90.82	89.90	90.36
attr	20	18	20	90.00	90.00	90.00
dobj	42	38	51	90.48	74.51	81.72
comp	62	45	49	72.58	91.84	81.08
iobj	3	2	2	66.67	100.00	80.00
prt	8	5	5	62.50	100.00	76.92
subj	43	29	33	67.44	87.88	76.32
mod	55	35	39	63.64	89.74	74.47
pobj	**13**	**11**	**24**	**84.62**	**45.83**	**59.46**
adjt	**30**	**15**	**25**	**50.00**	**60.00**	**54.55**
pred	**2**	**2**	**9**	**100.00**	**22.22**	**36.36**
cc	**1**	**0**	**0**	**0.00**	**NaN**	**0.00**
modnomatch	0	0	35	NaN	0.00	0.00

TABLE 6.12: Qualitative LAS2 results of Catalan Bare FDG

Specifically, the work around the adaptation of a quantitative corpus for FDGs evaluation have explained (§6.3.1) and the creation of a qualitative resource has been detailed (§6.3.2). In the last section of the chapter (§6.4), the results of a quantitative an qualitative task have been presented showing the suitability of the evaluation method and the emerging issues of the current version of the FDGs. In particular, some of these will be considered in the following chapters.

Chapter 7

Exploring PP-attachment

In the discussion about language ambiguities (§3), the PP-attachment has been noted as a highly ambiguous linguistic phenomenon in parsing (§3.2). A parser without added knowledge has a lot of problems with assigning the correct syntactic structure to a prepositional phrase (PP). Given a sentence like 'I ate pizza *with anchovies*', the parser cannot distinguish by itself if the right attachment of the PP is to the noun like 'pizza with anchovies' (n-attached solution) or to the verb like 'ate with anchovies' (v-attached solution).

The evaluation of the FDGs developed in this proposal (§6.4) shows that the PP-attachment ambiguities are a limitation for the grammar, as 77% of the prepositions in Spanish are attached to the right head and 71% of prepositions in Catalan. In particular, the results of the qualitative evaluation point to the fact that n-attached PP performs well when the prepositional head is 'de' ('of') like in the sentence 'La sopa del día es fantástica' ('The soup of the day is fantastic'). Furthermore, v-attached PP is not problematic if the preposition is other than 'de' ('He comido pizza con amigos', 'I ate pizza with friends').

However, the majority of problems concentrate on the n-attached PP when the preposition value is other than 'de' ('He comido pizza con anchoas', 'I ate pizza with anchovies'). In addition, following this argument, low accuracy should be expected in the v-attached PP with preposition 'de' when it is preceded by a noun ('Baja las cajas del desván', 'Bring the boxes down from the attic'), although this case is not in the evaluation data.

Because of the low performance and limitations on solving particular PP-attachments in FDGs (Lloberes et al., 2010), this problem needs to be solved in order to increase the accuracy of the grammars. In this chapter, two proposals are explored for disambiguating the problematic PP-attachment of the pattern *VP* NP_1 *P* NP_2 in Spanish and Catalan.

Firstly, an automatic learning method based on a supervised classifier for Spanish is described (§7.1). Aguilar et al. (2011) being aware of the limitations of the FDGs propose a supervised approach using a classifier using several layers of linguistic information in order to provide a solution of the ambiguous pattern *VP* NP_1 *P* NP_2 for the dependency grammars developed in this proposal (§5).

On the other hand, automatic learning word embeddings of word vector representations is also studied (§7.2). In particular, this approach is chosen for two reasons as follows: the results of a supervised experiment show limitations on generalizing unseen data (§7.1), and the successful results of a study about PP-attachment disambiguation in Arabic and English applying the word embeddings method (Belinkov et al., 2014) encourages us to extend this method to other languages.

The methodology and the results analysed are a first approximation of the PP-attachment disambiguation problem and the implementation of the word embeddings strategy in Spanish and Catalan (§7.2.1, §7.2.2, §7.2.3 and §7.2.4). Furthermore, the research presented in this chapter is a first attempt at the integration of the knowledge learned in Spanish and Catalan parsing and, specifically, in rule-based parsing of FDGs (§7.3). Finally, in order to empirically measure the contribution of this knowledge in parsing accuracy performance, the output of the FDGs with PP-attachment knowledge added will be evaluated (§7.4).

7.1 A Supervised Approach for PP-attachment

A supervised classifier is developed in this part of the experiment to solve PP-attachment ambiguity of the pattern *VP* NP_1 *P* NP_2 (Aguilar et al., 2011). Then, the goal of the classifier is to decide if the PP attaches to NP_1 or to *VP* (v-attached solution) using a supervised model.

This model relies on the data of the Spanish AnCora Corpus (Taulé et al., 2008). In particular, 4764 examples containing the ambiguous pattern *VP* NP_1 *P* NP_2 are used in the classification. This data contains 3171 instances of PP-attachment with the preposition 'de' ('of') and the rest of the 1593 instances occur with a preposition other than 'de', in which the second most frequent preposition is 'en' ('in') with 390 examples and the third most frequent preposition is 'a' ('to') with 302 examples. With regard to the attachment, 78% (3748 examples) are n-attached and 22% (1015 examples) are v-attached. 90% of the data has been used for training the classifiers and remaining 10% has been used as test data in the evaluation of the classifiers results.

In order to classify the PP-attachment examples, the following five features isolated or combined are used: preposition (lemma of the preposition), morphology (word form and lemma of NP_1 or *VP*, the number of words between NP_1 or *VP* and the preposition), morphosyntactic (lemma of NP_1 or *VP*, and word form, lemma and frequency of NP_2), syntactic (*VP* lemma argument structure frame subcategorizing a preposition) and semantic (Top Concept Ontology and Semantic File features of NP_1 or *VP* as well of NP_2).

The learning algorithms chosen are available in Weka environment (Witten and Frank, 2005) and the classifiers used are based on decision trees (J48), decision rules (JRip), Naive Bayes and Bayes Net. On the other hand, two baselines have been built classifying PP-attachment instances in the most frequent class (n-attached solution) and randomly classifying PP-attachment instances.

The results show that the random class baseline scores 66.3% in accuracy and the most frequent class baseline scores 78.7%. Among the classifiers, the two best results are the JRip classifier with the prepositional and the semantic features combined (87.84%), and the J48 classifier also working with prepositional and semantic features (86.79%). In the third place, the classifiers using J48, Bayes Net and Naive Bayes are tied at 85.74% when classifying only with the preposition feature.

Therefore, the lexical information and, specifically, the lexical value of the preposition is useful for disambiguating the PP-attachment. Furthermore, semantic information can contribute decisively to the disambiguation process. Despite the high-accuracy results obtained, the training corpus is rather too small to capture some relations and to make generalizations over unseen examples. Because of these conclusions, the implementation of an unsupervised approach can minimize the limitation of training with a small annotated corpus, as it will be explained in the second part of the experiment in the following section (§7.2).

7.2 Learning PP-attachment Distributionally

There are linguistically annotated corpora in Spanish and Catalan (such as AnCora, AnCora-UPF, AnCora Surface Syntax Dependencies and Tibidabo Treebank) that can be used in supervised automatic learning. However, as pointed out in the supervised part of the experiment (§7.1), the resources available for these languages are smaller than for supervised learning.

On the other hand, an unsupervised learning approach makes it possible to work with raw text. Consequently, the dependence on learning from human annotated data is not a limitation any more. Furthermore, large amounts of data is freely available in both languages in the Web or other resources like Wikipedia. For these reasons, an unsupervised approach is an appropriate method to follow in order to disambiguate the PP-attachment.

7.2.1 Distributional methods

Among the unsupervised methods currently available, the method applied in this experiment should capture the relations between the semantic properties of the words involved in the PP-attachment. Given a set of sentences like the following examples (1), words can be generalized by their semantic properties such as the features presented in (2), and grouped according to the semantic properties (3).

(1) a. Como pizza con anchoas
‘Ø$_{1sg}$ eat pizza with anchovies’
b. Como espaguetis con tomate
‘Ø$_{1sg}$ eat spaghetti with tomato’
c. Compro pantalones con botones
‘Ø$_{1sg}$ buy pants with buttons’
d. Como pizza con amigos
‘Ø$_{1sg}$ eat pizza with friends’

e. Como espaguetis con cuchara
'$\varnothing_{1sg}$ eat spaghetti with a spoon'

f. Compro pantalones con la paga extra
'$\varnothing_{1sg}$ buy pants with the extra income'

(2) a. *food*: 'pizza' (1-a), 'espaguetis' (1-b)
b. *ingredient*: 'anchovies' (1-a), 'tomato' (1-b)
c. *garment*: 'pants' (1-c)
d. *fastening method*: 'buttons' (1-c)
e. *company*: 'friends' (1-d)
f. *tool*: 'cuchara' (1-e)
g. *payment method*: 'paga extra' (1-f)

(3) a. *food* : *ingredient*
b. *garment* : *fastening method*
c. *eat* : *company*
d. *eat* : *tool*
e. *buy* : *payment method*

The classification of words by their semantic properties determines the direction of the attachment. Consequently, it is possible to say that the PP of the examples (1-a)–(1-c) is n-attached because there is a semantic relation between the properties of the nouns preceding the preposition and the nouns following the preposition. On the other hand, the PP of the examples (1-d)–(1-f) is v-attached because there is a particular relation between the semantic properties of the verb preceding the preposition and the semantic properties of the noun following the preposition.

In particular, distributional methods are able to capture semantic relations between words (i.e. semantic similarities) by applying the **distributional hypothesis** which states that words occurring in similar contexts tend to have similar meanings (Harris, 1954; Turney and Pantel, 2010; Clark, 2015). In addition, Lin and Pantel (2001) observe that patterns that co-occur with similar pairs tend to have similar meanings, which they called the **extended distributional hypothesis**. Similarly, Turney (2008) points that pairs of words co-occurring in similar patterns tend to have similar meanings, which correspond to the extended distributional hypothesis inversed and he names this the **latent relation hypothesis**.

From this point of view, a word can be represented by a vector that expresses the number of times that this word occurs (target word) in a context, which can be expressed as a single word, window of words (Lund and Burgess, 1996), syntactic dependencies (Lin, 1998a; Padó and Lapata, 2007) or syntactic dependencies combined with selectional preferences (Erk and Padó, 2008).

Co-occurrence frequencies are not expressive enough to indicate how two words are associated. As Evert (2005) points out, two frequent words with similar values are not necessarily similar because their co-occurrence may happen to be ad hoc. Furthermore, co-occurrence frequencies only refer to the information contained in the database, so the statistical inferences of co-occurrences does not allow it to make generalizations about the language.

In this context, **association measures** (e.g. Log-Likelihood, Chi^2, Local Mutual Information, Pointwise Mutual Information, etc.) can be used instead. They show the statistical attraction between co-occurring words and reduce the effect of the random co-occurrence factor. In this experiment, the measure of **Pointwise Mutual Information** (PMI) is used as Church and Hanks (1990) have formulated and Evert (2005) has implemented. PMI measures the amount of mutual information or overlap between two co-occurring words taking into account their statistical independence. In particular, the mutual information of two words (*x* and *y*) with probabilities *p(x)* and *p(y)* is calculated by the formula (4), which compares the probability of *x* and *y* co-occurring with the probabilities *x* and *y* occurring independently.

(4) $$\mathrm{PMI(x,y)} \quad = \quad \log_2 \frac{\mathrm{p(x,y)}}{\mathrm{p(x)p(y)}}$$

In the case of big matrices, a smoothing process can be applied in order to reduce the effect of little discriminative vectors. There are several strategies to smooth a matrix such as computing vectors over a certain threshold and applying singular value decomposition.

The representation of words by vectors allows them to be compared. The comparison of vectors is based on computing the similarities between vectors, which is handled by statistical measures or similarity measures (such as, cosine similarity, Euclidean distance and Manhattan distance). Among them, **cosine similarity** is one of the most popular measures to compare vectors in distributional methods. Specifically, it compares the distance of vectors by measuring the cosine of their angle (5).

(5) $$\cos(\mathrm{x,y}) \quad = \quad \frac{\mathrm{x}}{\|\mathrm{x}\|} \cdot \frac{\mathrm{y}}{\|\mathrm{y}\|}$$

Within the distributional methods, two tendencies of learning strategies can be differentiated (Baroni et al., 2014): traditional distributional methods and word embeddings. While the traditional distributional methods rely on context-counting semantic vectors (as described above), **word embeddings** are based on context-predicting semantic vectors (Collobert and Weston, 2008; Mikolov et al., 2013a,b), following the terminology proposed by Baroni et al. (2014). Word embeddings capture latent features of words (Turian et al., 2010) by setting weighted vectors of these words to predict the contexts in which these words usually occur (Baroni et al., 2014). In order to achieve this, latent features are induced using neural networks (Bengio et al., 2003).

The word embeddings method has already been tested successfully in PP-attachment for English and Arabic (Belinkov et al., 2014). The authors of this study develop several models for predicting the right head of ambiguous pp-atachment of the pattern *VP NP*$_1$ *P NP*$_2$ using the algorithm skip-gram implemented in word2vec (Mikolov et al., 2013a,b) and described in §7.2.4.

Once the models are learned, they compare the results of the experiments with several dependency parsers available in the languages being studied: Malt Parser (Nivre, 2003), MST Parser (McDonald et al., 2005), Turbo Parser (Martins et al., 2013) and RGB parser (Lei et al., 2014). Furthermore, two constituency parsers are used for the comparison: a recurrent neural network parser (Socher et al., 2013) and the Charniak self-trained reranking parser (McClosky et al., 2006).

In particular, they build three different models capturing latent features of the following patterns. Firstly, the **Head–Child model** learns the vectors from the head (*VP* or NP_1) and the child (NP_2). Secondly, the **Head–Prep–Child Ternary model** uses the three nodes involved in the PP-attachment (i.e. the head *VP* or NP_1, the preposition *P* and the child NP_2). Thirdly, the **Head–Prep–Child model** learns the relation between the preposition *P* and its possible head (i.e. *VP* or NP_1) and between the preposition *P* and its child NP_2.

In order to learn the models, the skip-gram algorithm with the default parameters proposed by Mikolov et al. (2013a,b) is applied. However, word vectors are relearned to improve the accuracy of the models by optimizing the training algorithm.

Furthermore, since the vectors are learned from raw text, they are enriched with morphosyntactic, syntactic and semantic information. Firstly, Part of Speech tags of the candidate word and its subsequent word are specified. Secondly, the set of verbs of the training data subcategorizing a prepositional argument which are in VerbNet (Kipper, 2005) add the lexical value of the subcategoized preposition specified in VerbNet. Thirdly, given a candidate word, WordNet hypernym (Fellbaum, 1998) of the candidate's head is provided.

In addition, apart from training vectors with their linear context of surrounding words as a default parameter of skip-gram algorithm, this algorithm is also trained with syntactic vectors of words containing the candidate word, its parent node, dependency relation label and its grandparent node which are obtained from automatically parsed text.

The context-predicting vectors learned are tested and evaluated in order to measure their performance in predicting unseen data. The best accuracy results of this study score 82.6% in Arabic and 88.7% in English. These scores are obtained with the Head-Prep-Child model when elements of the pattern *VP* NP_1 *P* NP_2 are in a distance smaller than 5 tokens, and when vectors are relearned, enriched with linguistic information, and complemented with syntactic vectors.

Because of the promising results of the word embeddings in PP-attachment disambiguation in Arabian and English (Belinkov et al., 2014), this method can be extended to other languages such as the ones of this proposal, i.e., Spanish and Catalan.

The remaining sections of this chapter focus on this initial approach to the disambiguation of the PP-attachment ambiguous pattern *VP* NP_1 *P* NP_2 by applying the word embeddings method. In this approach to the problem, latent semantic features of words are learned and represented by context-predicting vectors using the skip-gram algorithm (Mikolov et al., 2013a,b). In particular, the first model of a series of models is developed for this proposal and corresponds to the Head–Child model proposed by Belinkov et al. (2014), i.e., the model corresponding to the pattern head (*VP* or NP_1) and child (NP_2) with a delexicalized preposition.

In the following sections, the development of the experiment is explained. Firstly, the data sets which are used for training and test tasks are described (§7.2.2). In the first part of the experiment, a set of naive supervised classifiers has been developed in order to compare their naive results to the word embeddings results (§7.2.3). After learning the simple vectors by the naive classifiers, the central task of the experiment has been carried out and two models of context-predicting vectors have been learned and compared by calculating the similarity of their cosine (§7.2.4). In the last part of that section (§7.2.4), an evaluation of the accuracy of the naive supervised models and the embeddings models is provided. The last section is dedicated to a description of the integration of the knowledge learned about PP-attachment in FDGs (§7.3).

7.2.2 PP-attachment Training and Test Data

Word embeddings algorithms learn from unannotated text, which makes it possible to work with any kind of digitalized text. Since the size of digitalized text is currently enormous, a big corpus for training a model could be established. However, in order to reduce the amount of noise in the data, a pre-processed linguistic resource can be used. In particular, there is a linguistic resource available in Spanish and Catalan, **Wikicorpus** (Reese et al., 2010), which is suitable for the training task (§7.2.2.1 and §7.2.2.2).

This resource is a trilingual corpus that includes a large part of Wikipedia in English (600 million of tokens), Spanish (120 million of tokens) and Catalan (50 million of tokens). It is distributed in a raw text and a linguistically processed version. The linguistically annotated version of Wikicorpus is processed automatically by FreeLing NLP pipeline (Padró et al., 2010), which provides a lemma and a Part of Speech tag to every token of the corpus. Furthermore, tokens are automatically annotated with senses from WordNet (Fellbaum, 1998) using the Word Sense Disambiguation algorithm UKB (Agirre and Soroa, 2009).

To test and evaluate the models learned, a different set of data is used. In particular, **AnCora Corpus** (Taulé et al., 2008) can be an appropriate resource for these tasks because the syntactic annotations of this corpus contain the right answers to the problem that are trying to be resolved in this experiment (§7.2.2.3).

7.2.2.1 Training Data for Naive Classification

In this part of the experiment, the annotated versions of the Spanish and Catalan Wikicorpus have been used. For every sentence in the corpus, instances of the pattern $VP\ NP_1\ P\ NP_2$ are extracted. Order has not been defined as one of the parameters in the algorithm. Consequently, a sequence like $NP_1\ P\ NP_2\ VP$ is considered a variant of the pattern $VP\ NP_1\ P\ NP_2$. The pattern have been extracted by looking at the Part of Speech tags of tokens and matching them to the PP-attachment pattern. During this process, proper nouns and high-frequency verbs have been excluded due to their lower semantic contribution.

Then, the list of tokens matching the pattern is stored in two different data sets differentiated by the treatment of the preposition: a set with the preposition delexicalized (i.e. tokens represented by the pattern $VP\ NP_1\ NP_2$) and another set with the preposition lexicalized (i.e. tokens represented by the pattern $VP\ NP_1\ P\ NP_2$). Furthermore, each data set is classified according to the two possible PP-attachments (n-attached or v-attached). Two sub-sets of data for the n-attached (i.e. pattern $NP_1\ NP_2$) and the v-attached (i.e. pattern $VP\ NP_2$) solutions are created.

		Delexicalized preposition		Lexicalized preposition	
Dataset	**Pattern**	**Occurrences**	**%**	**Occurrences**	**%**
Train	NN	735,846	52.67	845,155	50.19
	VN	661,276	47.33	838,862	49.81
	total	1,397,122	100.00	1,684,017	100.00
	paris in pattern	1,397,122	2.47	1,684,017	2.98
	pairs in corpus	56,531,719	100.00	56,531,719	100.00

TABLE 7.1: Spanish training data set size

		Delexicalized preposition		Lexicalized preposition	
Dataset	**Pattern**	**Occurrences**	**%**	**Occurrences**	**%**
Train	NN	326,622	52.66	355,994	50.73
	VN	293,574	47.34	345,680	49.27
	total	620,196	100.00	701,674	100.00
	paris in pattern	620,196	2.61	701,674	2.96
	pairs in corpus	23,740,607	100.00	23,740,607	100.00

TABLE 7.2: Catalan training data set size

Therefore, the final training data set for each language (table 7.1 for Spanish and table 7.2 for Catalan) is composed of four sub-data sets of paired lemmas organized by lexicalization of the preposition (i.e. lexicalized or delexicalized) and attachment of the target NP_2 (i.e. NP_1 or VP).

The co-occurrence frequency is computed automatically for every pair NP_1 NP_2 and VP NP_2. Using this information, every pair is weighted by calculating the amount of association between the tokens of the pair NP_1 NP_2 and VP NP_2 with the PMI measure (Church and Hanks, 1990; Evert, 2005). The final set of vectors is not very large, so the risk of obtaining a sparse matrix is low. However, in order to avoid even a small amount of sparseness, the matrix is smoothed by discarding pairs with a co-occurrence frequency lower than a threshold.

7.2.2.2 Training Data for Word Embeddings

For the unsupervised learning task of the experiment, the Wikicorpus is also used and patterns with the delexicalized preposition are considered in this first run of the experiment. In particular two sets of data are created corresponding to corpus sentences and filtered sentences. The set of corpus sentences contains all the sentences of the corpus non-filtered and distributed in windows whose size corresponds to sentence length. On the other hand, set of filtered sentences includes all the corpus sentences filtered by nouns, verbs and adjectives and distributed in windows whose size is equivalent to sentence length. Sentences of one token have been discarded because they cannot be used for comparison. In this step of the research, only lexical information is used and, specifically, lemmas of the tokens are represented.

Proper nouns and high-frequency terms are filtered out because they tend to lack content that can determine the attachment of the PP. Furthermore, sentences formed by one token are discarded because they are useless in the comparison of tokens of the same sentence.

Dataset	**Pattern**	**Occurrences**	**%**
	non-filtered sentences	3,942,874	76.48
Train	filtered sentences	5,153,903	99.97
	Wikicorpus sentences	5,155,273	100.00

TABLE 7.3: Spanish training data set size

Dataset	**Pattern**	**Occurrences**	**%**
Train	non-filtered sentences	1,509,093	72.09
	filtered sentences	2,093,156	99.99
	Wikicorpus sentences	2,093,443	100.00

TABLE 7.4: Catalan training data set size

The final result of the training data extraction process is two sets of data for every language with filtered or non-filtered sentences. Each set is formed by a list of lemmas distributed in windows whose size corresponds to the sentence length which these lemmas belong to (table 7.3 for Spanish and table 7.4 for Catalan).

7.2.2.3 TEST DATA

In order to test the results of the experiment, a data set containing the correct attachments needs to be used. As presented in §6.3 about syntactically annotated corpora, there is only a single resource available in both languages of this proposal, which is the AnCora Corpus in the syntactic dependencies version. Therefore, this treebank is used as a test and the gold standard corpus.

Similar to the creation of the training data set, the pattern *VP* NP_1 *P* NP_2 is extracted from the AnCora Corpus sentences, specifically, from the dependency arcs of the treebank's sentences. Then, two versions of the data are created with the preposition lexicalized (i.e. pattern *VP* NP_1 *P* NP_2) and delexicalized (i.e. pattern *VP* NP_1 NP_2). A process of filtering proper nouns and high-frequency terms is applied to exclude data that does not contribute to attachment. Once again, the two versions are defined by the two possible attachments of the target *NP*$_2$, i.e. *NP*$_2$ n-attached (*NP*$_1$ *NP*$_2$) and *NP*$_2$ v-attached (*VP* *NP*$_2$).

		Delexicalized preposition		**Lexicalized preposition**	
Dataset	**Pattern**	**Occurrences**	**%**	**Occurrences**	**%**
	NN	14,422	51.98	15,112	51.73
	VN	13,322	48.02	14,101	48.27
Test	total	27,744	100.00	29,213	100.00
	paris in pattern	27,744	5.43	29,213	5.72
	pairs in corpus	510,665	100.00	510,665	100.00
	NN	6,044	45.74	6,240	45.16
	VN	7,169	54.26	7,577	54.84
Gold	total	13,213	100.00	13,817	100.00
	paris in pattern	13,213	2.59	13,817	2.71
	pairs in corpus	510,665	100.00	510,665	100.00

TABLE 7.5: Spanish test and gold data sets size

		Delexicalized preposition		**Lexicalized preposition**	
Dataset	**Pattern**	**Occurrences**	**%**	**Occurrences**	**%**
	NN	11,880	53.07	14,209	52.10
	VN	10,505	46.93	13,065	47.90
Test	total	**22,385**	100.00	**27,274**	100.00
	paris in pattern	22,385	4.66	27,274	5.68
	pairs in corpus	480,132	100.00	480,132	100.00
	NN	6,535	50.27	5,936	46.72
	VN	6,465	49.73	6,769	53.28
Gold	total	**13,000**	100.00	**12,705**	100.00
	paris in pattern	13,000	2.71	12,705	2.65
	pairs in corpus	480,132	100.00	480,132	100.00

TABLE 7.6: Catalan test and gold data sets size

Finally, the instances for the gold standard set are reduced to unique occurrences of the pairs *NP*$_1$ *NP*$_2$ and *VP* *NP*$_2$ with the preposition lexicalized or delexicalized (tables 7.5 and 7.6), whereas all occurrences of the pairs are kept in the test data set for classifying purposes (tables 7.5 and 7.6).

The test data that was classified according to the several models learned in this experiment is evaluated. The answers of several of the classifiers are compared to the solution of the gold standard data. In order to measure the results of the experiments, precision (*P*), recall (*R*) and f-measure (*F1*) are used (§6.2). Moreover, three patterns are evaluated: instances of correctly paired *NN* and *VN* in the total amount of instances (row **total** in tables), correctly paired *NN* in the total number of paired *NN* (row **NN** in tables), and correctly paired *VN* in the total number of paired *VN* (row **VN** in tables).

7.2.3 Naive Supervised Classifiers

A model of n-attached and v-attached assignments of PP is learned from the words representation vectors of the lemmas involved in the ambiguous PP-attachment (i.e., the pairs of lemmas NP_1 NP_2 and *VP* NP_2 with preposition lexicalized or delexicalized). The training data created in §7.2.2.1 is used in this learning task.

In particular, the learning algorithm uses the information about weight of the vectors. A right assignment is assumed to be handled by the pair (i.e. NP_1 NP_2 or *VP* NP_2) with the highest weight. Consequently, a target NP_2 is n-attached if the weight of the pair NP_1 NP_2 is higher than the weight of the pair *VP* NP_2. On the other hand, a target NP_2 is v-attached if the weight of the pair *VP* NP_2 is higher than the weight of the pair NP_1 NP_2. Finally, the cases where pairs are assigned the same weight are disambiguated with comparing co-occurrence frequency, so that the pair with the higher score is selected.

Three different classifiers using the model learned are built in order to set three baselines (tables 7.7 and 7.8) to compare the results of the experiment applying word embeddings method. A first classifier (**Weight Classifier**) disambiguates test pairs by matching: if the test pair is in the model, disambiguate according to the model pattern, otherwise the pair remains ambiguous. A second classifier (**Most Frequent Classifier**) makes decisions according to the model and by assigning test pairs to the most frequent pattern (i.e. NP_1 NP_2) if the pair does not match with any pair in the model. A third classifier (**Single Class Classifier**) groups all the test pairs (i.e. NP_1 NP_2 either *VP* NP_2) in the most frequent class.

In addition, only test pairs with at least three occurrences in the corpus are classified. Below this threshold occurrences are marginal cases and can cause some sparseness. Other thresholds have also been tested (i.e. pairs $\geq$ 5 occurrences and pairs $\geq$ 10 occurrences), but lower accuracy results were observed.

The results of the evaluation task about PP-attachment disambiguation show a parallel behaviour of the data in Spanish (tables 7.9 and 7.10) and in Catalan (tables 7.11 and 7.12). Concerning the Spanish PP-attachment disambiguation, the model with delexicalized preposition (*VP* NP_1 NP_2) has the best precision score (table 7.9) in the Weight Classifier (0.6154), although its recall score is very low (0.2245). On the other hand, the best recall score is obtained by Most Frequent Classifier (0.5594), although the precision of this classifier is very low (0.2664). With regard to

		Delexicalized preposition		Lexicalized preposition	
Classifier	**Pattern**	**Occurrences**	**%**	**Occurrences**	**%**
Weight	NN	2,471	51.27	1,669	59.65
	VN	2,349	48.73	1,129	40.35
	total	4,820	100.00	2,798	100.00
	paris in pattern	4,820	17.37	2,798	9.58
	pairs in corpus	27,744	100.00	29,213	100.00
Most Frequent	NN	25,395	91.53	28,084	96.14
	VN	2,349	8.47	1,129	3.86
	total	27,744	100.00	29,213	100.00
	paris in pattern	27,744	100.00	29,213	100.00
	pairs in corpus	27,744	100.00	29,213	100.00
Single Class	NN	27,744	100.00	29,213	100.00
	VN	0	0.00	0	0.00
	total	27,744	100.00	29,213	100.00
	paris in pattern	27,744	100.00	29,213	100.00
	pairs in corpus	27,744	100.00	29,213	100.00

TABLE 7.7: Size of Spanish test data classified by naive supervised models

the F1 measure, the best classifier is Most Frequent Classifier (0.3610), but this result is still low.

The same situation is observed when the model with lexicalized preposition (*VP* NP_1 *P* NP_2) is tested (table 7.10). The classifier with the highest precision score is the Weight Classifier (0.6751), the best recall score as well as f-measure is obtained by the Most Frequent Classifier (0.4900 and 0.3147, respectively), but these scores are not significant either.

Catalan results of the classifiers using the lexicalized and the delexicalized models (tables 7.11 and 7.12) are not an exception to the tendency observed in the Spanish data. The best classifier in precision is the Weight Classifier (0.7163 using the delexicalized model and 0.7400 using the lexicalized model), the best classifier in recall and f-measure is the Most Frequent Classifier (0.5428 and 0.3988 respectively in the delexicalized model, and 0.4815 and 0.3061 respectively in the lexicalized model).

From these observations, it can be concluded that in Spanish and in Catalan delexicalized model obtain better results because precision and recall are slightly more balanced than in the lexicalized model, but the lexicalized model is more precise in the three classifiers. Furthermore, the pattern NP_1 NP_2 tends to get a high precision (in almost all the results of the Weight Classifier

		Delexicalized preposition		Lexicalized preposition	
Classifier	**Pattern**	**Occurrences**	**%**	**Occurrences**	**%**
Weight	NN	1,092	57.26	949	67.98
	VN	815	42.74	447	32.02
	total	1,907	100.00	1,396	100.00
	paris in pattern	1,907	8.52	1,396	5.12
	pairs in corpus	22,385	100.00	27,274	100.00
Most Frequent	NN	21,570	96.36	26,827	98.36
	VN	815	3.64	447	1.64
	total	22,385	100.00	27,274	100.00
	paris in pattern	22,385	100.00	27,274	100.00
	pairs in corpus	22,385	100.00	27,274	100.00
Single Class	NN	22,385	100.00	27,274	100.00
	VN	0	0.00	0	0.00
	total	22,385	100.00	27,274	100.00
	paris in pattern	22,385	100.00	27,274	100.00
	pairs in corpus	22,385	100.00	27,274	100.00

TABLE 7.8: Size of Catalan test data classified by naive supervised models

in both models and in both languages). In particular, the results of this pattern are high in the lexicalized model because the variability of the lexical value of the preposition is more restricted than in the pattern *VP NP*$_2$, so the classifier has a smaller range of cases to solve.

In addition, among the classifiers, the Weight Classifier receives the best precision score, but it is extremely limited in making generalizations for unseen data as the low recall results show. On the other hand, the Most Frequent Classifier and the Single Class Classifier improve on the generalization task as the recall scores are better, but they are not at all satisfactory because they are all around 50%. These two classifiers cannot be reliable tools for disambiguating PP-attachment either because their precision scores drop drastically compared to Weight Classifier. Therefore, a robust automatic learning method needs to be applied in order to provide better results on PP-attachment disambiguation.

7.2.4 Learning Word Embeddings

In the unsupervised learning task of the experiment, word2vec is used to learn a model for PP-attachment based in word embeddings (Mikolov et al., 2013a,b). As previously described (§7.2.1),

		Delexicalized preposition					
Classifier	**Pattern**	**Gold**	**System**	**Correct**	**P**	**R**	**F1**
Weight	total	13,213	4,820	2,966	**0.6154**	0.2245	0.3290
	NN	6,044	2,471	1,618	0.6548	0.2677	0.3800
	VN	7,169	2,349	1,348	0.5739	0.1880	0.2833
Most Frequent	total	13,213	27,744	7,392	0.2664	**0.5594**	**0.3610**
	NN	6,044	25,395	6,044	0.2380	1.0000	0.3845
	VN	7,169	2,349	1,348	0.5739	0.1880	0.2833
Single Class	total	13,213	27,744	6,044	0.2178	0.4574	0.2951
	NN	6,044	27,744	6,044	0.2178	1.0000	0.3578
	VN	7,169	0	0	0	0	0

TABLE 7.9: Evaluation scores of Spanish naive classifiers with delexicalized preposition

		Lexicalized preposition					
Classifier	**Pattern**	**Gold**	**System**	**Correct**	**P**	**R**	**F1**
Weight	total	13,817	2,798	1,889	**0.6751**	0.1367	0.2274
	NN	6,240	1,669	1,358	0.8137	0.2176	0.3434
	VN	7,577	1,129	531	0.4703	0.0701	0.1220
Most Frequent	total	13,817	29,213	6,771	0.2318	**0.4900**	**0.3147**
	NN	6,240	28,084	6,240	0.2222	1.0000	0.3636
	VN	7,577	1,129	531	0.4703	0.0701	0.1220
Single Class	total	13,817	29,213	6,240	0.2136	0.4516	0.2900
	NN	6,240	29,213	6,240	0.2136	1.0000	0.3520
	VN	7,577	0	0	0	0	0

TABLE 7.10: Evaluation scores of Spanish naive classifiers with lexicalized preposition

this is a NLP tool that learns the word distributions by applying a Recurrent Neural Network. Automatic learning tools using a neural network have usually been associated with deep learning techniques. Despite this, the authors of word2vec (Mikolov et al., 2013a,b) explicitly state that the algorithms of this tool cannot be considered deep learning since they make use of two shallow neural networks: continuous skip-gram (skip-gram) and continuous bag of words (cbow).

		Delexicalized preposition					
Classifier	**Pattern**	**Gold**	**System**	**Correct**	**P**	**R**	**F1**
Weight	total	13,000	1,907	1,366	**0.7163**	0.1051	0.1833
	NN	6,535	1,092	845	0.7738	0.1293	0.2216
	VN	6,465	815	521	0.6393	0.0806	0.1431
Most Frequent	total	13,000	22,385	7,056	0.3152	**0.5428**	**0.3988**
	NN	6,535	21,570	6,535	0.3030	1.0000	0.4650
	VN	6,465	815	521	0.6393	0.0806	0.1431
Single Class	total	13,000	22,385	6,535	0.2919	0.5027	0.3694
	NN	6,535	22,385	6,535	0.2919	1.0000	0.4519
	VN	6,465	0	0	0	0	0

TABLE 7.11: Evaluation scores of Catalan naive classifiers with delexicalized preposition

		Lexicalized preposition					
Classifier	**Pattern**	**Gold**	**System**	**Correct**	**P**	**R**	**F1**
Weight	total	12,705	1,396	1,033	**0.7400**	0.0813	0.1465
	NN	5,936	949	851	0.8967	0.1434	0.2472
	VN	6,769	447	182	0.4072	0.0269	0.0504
Most Frequent	total	12,705	26,827	6,118	0.2243	**0.4815**	**0.3061**
	NN	5,936	26,827	5,936	0.2213	1.0000	0.3624
	VN	6,769	447	182	0.4072	0.0269	0.0504
Single Class	total	12,705	27,274	5,936	0.2176	0.4672	0.2970
	NN	5,936	27,274	5,936	0.2176	**1.0000**	0.3575
	VN	6,769	0	0	0	0	0

TABLE 7.12: Evaluation scores of Catalan naive classifiers with lexicalized preposition

Broadly speaking, **skip-gram** is an algorithm which predicts the neighbouring words of a target word (*c*) given a window size of *n* words around the target word *w*, which can be formally expressed as a conditional probability like *p(c|w)*. On the other hand, **cbow** predicts a target word (*w*) given the neighbouring words (*c*) of the window where this target word appears, which is described by a conditional probability like *p(w|c)*.

The evaluation of both algorithm architectures shows that the skip-gram performs better than cbow, although it is slower in performance (Mikolov et al., 2013a). Skip-gram scores 53.3% in accuracy, while cbow stays at 36.1% of accuracy. In particular, the scores of skip-gram are obtained when the parameters are set as follows: vector size at 100, window size at 10 tokens, word frequency at 5 occurrences minimum. Consequently, the models of this experiment are learned using skip-gram with the parameters set with the best configuration.

In this experiment, the task of learning word embeddings relies on building two models of context-predicting vectors from two training data sets for Spanish and Catalan languages containing the whole set of Wikicorpus sentences (i.e. non-filtered sentences as they appear in the corpus) and Wikicorpus filtered sentences (i.e. sentences of the corpus filtered by nouns, verbs and adjectives), as presented in §7.2.2.2.

As a result of the training task, two different models are built: a model of non-filtered sentences and a model of filtered sentences. The former model contains the set of semantic vectors expressing the word distributions of the non-filtered sentences and the latter model is defined by the semantic vectors expressing the word distributions of the filtered sentences.

Once the two models are built, they are tested with the aligned pairs *NP*$_1$ *NP*$_2$ and *VP NP*$_2$ (from the pattern *VP NP*$_1$ *NP*$_2$ with delexicalized prepositions) of the test data extracted from the AnCora Corpus (§7.2.2.3). The cosine distance of the angle formed by a target word vector ($\overrightarrow{NP_2}$) and one of the two possible context vectors ($\overrightarrow{NP_1}$ or $\overrightarrow{VP}$) is computed for predicted pairs in the model of filtered sentences and the model of non-filtered sentences. The result is a scale of integer scores in which the highest score (1) corresponds to the absolute similarity and the lowest score (0) is the absence of similarity. Consequently, the pairs not predicted by the models are disambiguated by assigning them a 0 score.

The previous supervised classifiers (§7.2.3) are discriminative in the way that if *NP*$_2$ is n-attached it cannot be v-attached at the same time. On the other hand, since distributional models measure the similarity of items in a gradation, a solution can have more than one answer. Consequently, a *NP*$_2$ can be n-attached if its semantic properties and semantic properties of *NP*$_1$ are similar. It can be v-attached if its semantic properties and semantic properties of *VP* are similar. However, it can also be both n-attached and v-attached, although both solutions are not identical. One of the two solutions is more prototypical from the point of view of semantic similarity when similarity measure scores higher.

According to this, test pairs are classified into three grades: strongly similar, weakly similar and ambiguous (tables 7.13 and 7.14). Furthermore, they are classified according to the disambiguation solution, n-attached (*NN*) or v-attached (*VN*), as shown in tables 7.13 and 7.14. Therefore, the following solutions are possible in the classification of PP-attachment disambiguation by cosine similarity:

1. If the cosine similarity of $\overrightarrow{NP_2}$ and $\overrightarrow{NP_1}$ is higher than the similarity of $\overrightarrow{NP_2}$ and $\overrightarrow{VP}$ (i.e. similarity $\overrightarrow{NP_1 NP_2}$ > similarity $\overrightarrow{VPNP_2}$), then, *NP*$_2$ is strongly n-attached and weakly v-attached.

2. If the cosine similarity of $\overrightarrow{NP_2}$ and $\overrightarrow{VP}$ is higher than the similarity of $\overrightarrow{NP_2}$ and $\overrightarrow{NP_1}$ (i.e. similarity $\overrightarrow{VPNP_2}$ > similarity $\overrightarrow{NP_1NP_2}$), then, *NP*$_2$ is strongly v-attached and weakly n-attached.

3. If the cosine similarity of $\overrightarrow{NP_2}$ and $\overrightarrow{NP_1}$ is the same than the similarity of $\overrightarrow{NP_2}$ and $\overrightarrow{VP}$ (i.e. similarity $\overrightarrow{NP_1NP_2}$ = similarity $\overrightarrow{VPNP_2}$), then, *NP*$_2$ remains ambiguous. The number of these cases is marginal, so no disambiguation process is applied.

Model	**Similarity**	**Pattern**	**Occurrences**	**%**
Filtered Sentences	Strong	NN	8,431	63.31
		VN	4,886	36.69
		total	13,317	100.00
	Weak	NN	3,790	40.95
		VN	5,465	59.05
		total	9,255	100.00
	Ambiguous	NN	5	50.00
		VN	5	50.00
		total	10	100.00
	Classified pairs		22,582	81.39
	Pairs in test		27,744	100.00
Non-filtered Sentences	Strong	NN	9,090	65.90
		VN	4,704	34.10
		total	13,794	100.00
	Weak	NN	4,128	37.59
		VN	6,855	62.41
		total	10,983	100.00
	Ambiguous	NN	2	50.00
		VN	2	50.00
		total	4	100.00
	Classified pairs		24,781	89.32
	Pairs in test		27,744	100.00

TABLE 7.13: Size of Spanish test data classified by cosine similarity

The accuracy of the test pairs classified by their cosine similarity has been evaluated for both semantic vectors models of filtered and non-filtered sentences, the three-way similarity classification (strong, weak and ambiguous similarity), and the two possible PP-attachment disambiguation solutions (n-attached or v-attached). The answers of the classification have been compared

Model	Similarity	Pattern	Occurrences	%
Filtered Sentences	Strong	NN	5,891	61.03
		VN	3,762	38.97
		total	9,653	100.00
	Weak	NN	2,569	45.30
		VN	3,102	54.70
		total	5,671	100.00
	Ambiguous	NN	1	50.00
		VN	1	50.00
		total	2	100.00
	Classified pairs		15,326	68.47
	Pairs in test		22,385	100.00
Non-filtered Sentences	Strong	NN	7,194	63.75
		VN	4,091	36.25
		total	11,285	100.00
	Weak	NN	3,858	41.47
		VN	5,446	58.53
		total	9,304	100.00
	Ambiguous	NN	5	50.00
		VN	5	50.00
		total	10	100.00
	Classified pairs		20,599	92.02
	Pairs in test		22,385	100.00

TABLE 7.14: Size of Catalan test data classified by cosine similarity

to the answers of the gold standard pairs (tables 7.15 and 7.16), and they have been measured with the metrics of precision (*P*), recall (*R*) and f-measure (*F1*), that were introduced in §6.2.

The total results show that both models of filtered and non-filtered sentences in both languages get a F1 score between 60% and 65%, and 72% in the non-filtered sentences model for Catalan (tables 7.15 and 7.16). These results are low compared to the best results of Belinkov et al. (2014) for English (88.7%) and Arabic (82.6%), which are obtained using a semantic vectors model including prepositions (Head-Prep-Child model) enriched with syntactic and semantic information.

Despite the low scores results, if they are compared with the Belinkov et al. (2014) Head-Child model using the same kind of information as the experiment presented here (i.e. lexical information about the head and child involved in the PP-attachment with delexicalized preposition), the results are found to be better in 2 points at least and in 13 points at most. Belinkov et al. (2014) achieve 59% accuracy, whereas the models proposed in this experiment score 61% in Spanish for filtered sentences model, 63% in Spanish for non-filtered sentences model, 63% in Catalan for filtered sentences and 72% in Catalan for non-filtered sentences model.

A closer look at these results shows that the precision is low (i.e. under 50% in Spanish and under 60% in Catalan), whereas the recall gets high scores in the majority of models. The distributional models learned and the classification strategy followed in this experiment makes it possible to capture a large number of solutions, because all the answers of the classifier are accepted (except for the ambiguous ones), although some are more acceptable (strongly similar answers) than others (weakly similar answers), as shown in tables 7.15 and 7.16.

When comparing the performance of the model of filtered sentences and the model of non-filtered sentences, there are no significant differences in either language (tables 7.15 and 7.16). Precision scores are almost the same in both models for each language. On the other hand, recall metric improves significantly in non-filtered sentences, i.e., in Spanish the performance of the non-filtered sentences model increases 7.47 points and in Catalan 23.86 points. The fact that all relations in a sentence are captured by the model positively influences the recall.

With regard to the classes of similarity, strongly similar pairs are predicted better than the weakly similar pairs. Strongly similar pairs are the prototypical PP-attachment prediction, so the fact that they get better scores indicates that the models learned successfully capture the similarities. However, as the low score results show, the models need to be improved in order to increase their precision and recall.

In addition, the comparison of the results of the disambiguation of the patterns NP_1 NP_2 and VP NP_2 shows that the v-attached solution is performed better because the precision and the recall scores are higher than the n-attached solution. In the previous section (§7.2.3), the results of the evaluation showed that naive supervised classifiers were limited in solving PP-attachments to VP because they could not predict the variability of the context when PP is attached to VP. The results of this experiment point to the idea that distributional models tend to overcome this kind of limitation, although this experiment needs to be extended to empirically prove this statement.

All these observations point to the fact that, despite the low score results, word embeddings are a suitable method to improve the PP-attachment. Semantic similarities are captured by the distributional models learned in this experiment. However, as the results show, the precision of these models needs to be improved significantly along with the recall when learning with filtered sentences and without decreasing the recall scores when learning with non-filtered sentences.

The lexical information seems not to be expressive enough to provide a satisfactory disambiguation. For this reason, the PP-attachment disambiguation with context-predicting vectors needs to be enriched with new layers of more abstract linguistic information (e.g. subcategorization, semantic features, etc.), as (Aguilar et al., 2011; Belinkov et al., 2014) argue. Training data which has been extracted from Wikicorpus (Reese et al., 2010) is already annotated with WordNet

Model	Similarity	Pattern	Gold	System	Correct	P	R	F1
Filtered Sentences	Total		13,213	22,582	10,993	**0.4868**	0.8320	0.6142
	Strong	NN	6,044	8,431	4,013	0.4760	0.6640	0.5545
		VN	7,169	4,886	3,147	0.6441	0.4390	0.5221
	Weak	NN	6,044	3,790	1,318	0.3478	0.2181	0.2680
		VN	7,169	5,465	2,510	0.4593	0.3501	0.3973
	Ambiguous	NN	6,044	5	3	0.6000	0.0005	0.0010
		VN	7,169	5	2	0.4000	0.0003	0.0006
Non-filtered Sentences	Total		13,213	24,781	11,980	0.4835	**0.9067**	**0.6306**
	Strong	NN	6,044	9,090	4,303	0.4734	0.7119	0.5687
		VN	7,169	4,704	3,083	0.6554	0.4300	0.5193
	Weak	NN	6,044	4,128	1,373	0.3326	0.2272	0.2700
		VN	7,169	6,855	3,220	0.4697	0.4492	0.4592
	Ambiguous	NN	6,044	2	2	1.0000	0.0003	0.0007
		VN	7,169	2	0	0	0	0

TABLE 7.15: Evaluation scores of Spanish cosine similarity classification

Model	Similarity	Pattern	Gold	System	Correct	P	R	F1
Filtered Sentences	Total		13,000	15,326	9,006	0.5876	0.6928	0.6359
	Strong	NN	6,535	5,891	3,434	0.5829	0.5255	0.5527
		VN	6,465	3,762	2,632	0.6996	0.4071	0.5147
	Weak	NN	6,535	2,569	1,264	0.4920	0.1934	0.2777
		VN	6,465	3,102	1,675	0.5400	0.2591	0.3502
	Ambiguous	NN	6,535	1	1	1.0000	0.0002	0.0003
		VN	6,465	1	0	0	0	0
Non-filtered Sentences	Total		13,000	20,599	12,108	**0.5878**	**0.9314**	**0.7208**
	Strong	NN	6,535	7,194	4,273	0.5940	0.6539	0.6225
		VN	6,465	4,091	2,947	0.7204	0.4558	0.5584
	Weak	NN	6,535	3,858	1,897	0.4917	0.2903	0.3651
		VN	6,465	5,446	2,986	0.5483	0.4619	0.5014
	Ambiguous	NN	6,535	5	2	0.4000	0.0003	0.0006
		VN	6,465	5	4	0.8000	0.0006	0.0012

TABLE 7.16: Evaluation scores of Catalan cosine similarity classification

senses (Fellbaum, 1998) with the UKB algorithm (Agirre and Soroa, 2009). This information can be mapped to Top Concept Ontology features and Semantic File following (Aguilar et al., 2011). A layer of subcategorization information can also be added using AnCora-Verb lexicon (Aparicio et al., 2008) and SenSem lexicon (Vázquez and Fernández-Montraveta, 2015), because both resources are also annotated with WordNet senses (Castellón et al., 2003).

Moreover, the preliminary experiment described in §7.1 and the proposal of Belinkov et al. (2014) demonstrate that the lexical value of the preposition is a determining factor in the disambiguation. Therefore, new word embedding models for PP-attachment need to be learned with the preposition lexicalized. Lexicalized prepositional models can be extended in the two classes of models that Belinkov et al. (2014) build (§7.2.1), the Head-Prep-Child-Ternary model and the Head-Prep-Child model.

7.3 Integration of PP-attachment Knowledge in FDGs

The results of the word embeddings PP-attachment experiment are not expected to improve the FreeLing Dependency Grammars (FDGs). However, the knowledge learned has been implemented in order to determine if the current version of the grammars can handle successfully this kind of information in the future.

As detailed in the description about the state of the art of PP-attachment disambiguation proposals (§3.2.2), there are proposals that are parsing-aware in such a way that the solutions of the learning task are directly integrated in the parser (Foth and Menzel, 2006; Agirre et al., 2008; Henestroza and Candito, 2011).

Currently, FDGs have not implemented a mechanism that detects the PP-attachment errors during the parsing task as the system of Agirre et al. (2008) does, and that automatically proposes a solution for the PP-attachment errors that are detected using the information that the context-predicting vectors model learned provide.

Despite this, the knowledge learned in the previous supervised and unsupervised learning tasks (§7.2.3 and §7.2.4) can be integrated into FDGs. The current architecture of the grammars allows access to paired lemmas classified by a specific syntactic or semantic criterion as explained in the architecture of attachment rules (§5.2.1).

In particular, lists of paired lemmas n-attached or v-attached can be integrated in the section <*PAIRS*> and accessed by the rules such as in the figures 7.1 and 7.2. Therefore, the predictions n-attached and v-attached of the several supervised classifiers (§7.2.3) and the two word embedding models §7.2.4 have been integrated in FDGs as lists of paired lemmas.

On the other hand, in order for attachment rules to work with PP-attachment knowledge, the classes were declared in the rules. As a consequence of the integration, 28 rules were modified and a total of 7 new rules were added to handle the PP-attachment classes. In particular, 2 of these new rules make use of the knowledge learned about PP-attachment, a rule that attaches the PP to the preceding NP (figure 7.1) and a rule that attaches the PP to the preceeding VP (figure 7.2).

```
114 - - (sn,grup-sp) n-sim::(L.lemma,R:sn.lemma) top_left RELABEL -:-
```

FIGURE 7.1: PP-attachment rule for n-attached solution

```
102 - - (grup-verb,sp-de) v-sim::(L.lemma,R:sn.lemma) top_left RELABEL -:-
```

FIGURE 7.2: PP-attachment rule for v-attached solution

The rule that implements the n-attached solution (figure 7.1) states that at priority 114 a PP (*grup-sp*) whose head corresponds to any preposition other than 'de' ('of') is attached to the preceding NP (*sn*) when the lemma of the preceding NP and the lemma of the NP inside of the PP (*L.lemma,R:sn.lemma*) are paired lemmas in the list of strongly similar terms (*n-sim*), which is defined in the section *<PAIRS>*.

Similarly, the rule for the v-attached solution (figure 7.2) attaches a PP (*sp-de*) whose head has the lexical value 'de' ('of') to the preceding VP (*grup-verb*) at priority 102 if both lemmas of the preceding VP and of the NP inside of the PP (*L.lemma,R:sn.lemma*) are in the list of strongly v-attached similar paired lemmas (*v-sim*) in the section *<PAIRS>*.

As observed in the evaluation of the initial FDGs developed in this proposal (§6.4), the grammar performs accurately on n-attached PP when the lexical value of the preposition is 'de' ('of'). On the other hand, v-attachment PP is handled successfully with prepositions other than 'de' ('of') in the head position. For this reason, these two attachments are treated as the default in this new version of the grammar, so no additional knowledge is added to the rules. In addition, this criterion is supported by the data provided by Aguilar et al. (2011) about attachment and distribution of prepositions in the AnCora Corpus. Since 2% of the occurrences of preposition 'de' are attached to the VP, most of the occurrences of this preposition go with a NP and the other prepositions attach to a VP the majority of times.

Therefore, the prepositional attachments that need to be handled with the knowledge integrated are two. Firstly, a new rule for n-attached PP whose head is a preposition other than 'de' ('of') which is illustrated in figure 7.1. Secondly, a new rule to deal with v-attached PP when the lexical value of the head is the preposition 'de' ('of') which corresponds to the rule of figure 7.2. Both rules are more restrictive than the default rules for PP-attachment. For this reason, they are applied before the default rules by assigning a higher priority than the default rules.

7.4 Evaluation of PP-attachment Performance

In this section, the results of an evaluation task of FDGs using PP-attachment knowledge are presented in order to empirically assess the performance of the grammars. Different versions of the FDGs have been created in both languages with the predictions made by several naive supervised classifiers (§7.2.3) and the word embeddings models (§7.2.4). In particular, parsed dependency trees performed by FDGs are compared to the dependency trees of the Tibidabo Treebank (Marimon et al., 2014) and of the ParTes test suite (Lloberes et al., 2014, 2015b) in order to provide a quantitative and qualitative analysis of the grammars' performance.

The knowledge integrated in the grammars is a result of the first of a series of experiments, so the results are expected to be improved in future experiments. The performance of the embedding models learned are promising, but they still need to be tuned-in with new linguistic information and new patterns need to be learned, as stated in the previous section (§7.2). For this reason, the results of the evaluation of FDGs are likely not to be very explanatory about the accuracy of PP-attachment knowledge integration in parsing. However, they can confirm the observations of the evaluation of the models (§7.2.3 and §7.2.4) and point to specific issues related to PP-attachment that remain to be solved in FDGs and that need to be handled in the models of future experiments.

7.4.1 Evaluation Experiments

Six versions of FDGs have been tested for each language according to several models that have been learned about PP-attachment disambiguation (§7.2.3 and §7.2.4).

- **Bare**. Version of the grammar presented in §5 and evaluated in §6.4 which runs with rules without extra linguistic information about PP-attachment disambiguation.
- **PP-Weight**. Version of the grammar running with the lists of n-attached and v-attached pairs learned with the Weight Classifier (§7.2.3).
- **PP-Frequent**. Version of the grammar running with the lists of n-attached and v-attached pairs learned with the Most Frequent Classifier (§7.2.3).
- **PP-Monoclass**. Version of the grammar with the lists of n-attached and v-attached pairs learned with the Single Class Classifier (§7.2.3).
- **PP-Similar-F**. Version of the grammar running with the lists of n-attached and v-attached pairs generated from the model of filtered sentences learned with an unsupervised method of word embeddings and classified by the similarity of their vectors cosine (§7.2.4).
- **PP-Similar-NF**. Version of the grammar running with the lists of n-attached and v-attached pairs generated from the model of non-filtered sentences learned with an unsupervised method of word embeddings and classified by the similarity of their vectors cosine (§7.2.4).

Grammar	**NP$_1$ NP$_2$**	**VP NP$_2$**
PP-Weight	245	251
PP-Frequent	2552	251
PP-Monoclass	2960	0
PP-Similar-F	1221	1005
PP-Similar-NF	1336	1146

TABLE 7.17: Number of PP-attachment paired lemmas of Tibidabo Treebank

Except Bare Grammar (2,547 attachment rules), the rest of grammars are identical versions of the same grammar (2,554 attachment rules), but running with the different paired lemmas lists about PP-attachment described in the previous list.

Since FDGs are not prepared for detecting and solving PP-attachment errors during the parsing process (§7.3), this evaluation task emulates this process. Data used in this evaluation correspond to the harmonized dependency trees of the Tibidabo Treebank whose sentences correspond to a sub-set of the AnCora Corpus sentences. Because AnCora Corpus has been used in the test data to learn the models about PP-attachment (§7.2.2.3), n-attached and v-attached paired lemmas are also most probably in the Tibidabo. However, some annotation criteria are different, which make both resources different in the analysis. Therefore, learned paired lemmas extracted from AnCora and n-attached and v-attached PP of Tibidabo Treebank may not correspond. For example, 462 PP-attachments of *VP NP$_1$ P NP$_2$* of Tibidabo Treebank do not coincide with either n-attached or v-attached paired lemmas originally from AnCora Corpus.

For this reason, both supervised and unsupervised experiments have been reproduced with a list of aligned n-attached and v-attached lemmas of the Tibidabo Treebank containing the ambiguous pattern *VP NP$_1$ P NP$_2$* (table 7.17). This test data set have been created following the same methodology applied to set the of test pairs from AnCora (§7.2.2.3). In addition, the supervised and the unsupervised tasks are performed using the same parameters and tools of the naive learning experiment and the word embeddings learning experiment in order to reproduce the experiments under the same conditions.

The evaluation method relies on the empirical principles for evaluating NLP tools explained in §6.1. With regard to the evaluation of Spanish grammars, both quantitative and qualitative analyses are carried out. In the quantitative evaluation, the sentences of the Tibidabo Treebank (§6.3.1.2) which have been analysed with every version of FDGs (system analysis) are compared to the sentences of the treebank which have been manually annotated with syntactic information and harmonized to the syntactic criteria of the FDGs (gold analysis). To perform a qualitative analysis, several versions of the grammar are analysed with the data of the test suite ParTes (§6.3.2) and compared to the manually annotated test sentences of the test suite.

Grammar	LAS	UAS	LAS2
Bare	81.52	89.57	83.95
PP-Weight	81.58	89.59	83.90
PP-Frequent	81.76	89.42	84.14
PP-Monoclass	81.84	89.50	84.22
PP-Similar-F	81.61	89.46	83.96
PP-Similar-NF	81.61	89.46	83.96

TABLE 7.18: Quantitative evaluation scores in Spanish

Grammar	LAS	UAS	LAS2
Bare	80.99	90.11	81.94
PP-Weight	80.99	90.11	81.94
PP-Frequent	80.99	90.11	81.94
PP-Monoclass	80.99	90.11	81.94
PP-Similar-F	80.99	90.11	81.94
PP-Similar-NF	80.99	90.11	81.94

TABLE 7.19: Qualitative evaluation scores in Spanish

Concerning Catalan FDG evaluation, the grammars can only be evaluated qualitatively with the data of ParTes. ParTes test data is controlled in size, so it cannot be used for learning new models with evaluation data. For this reason, the n-attached and v-attached paired lemmas classes learned in the supervised experiment (§7.2.3) and in the unsupervised experiment (§7.2.4) are used.

Since PP-attachment knowledge is integrated for improving the construction of syntactic structure, attachment rules of the grammars are only evaluated and statistical metrics concerning attachment accuracy are only used (§6.2).

7.4.2 Evaluation Results

The quantitative global results in table 7.18 show that *Bare* grammar perform dependency trees in medium-high accuracy (*LAS*). The addition of simple PP-attachment knowledge does not contribute to a big increment of the accuracy, as *LAS* improves 0.32 points at most in PP-Monoclass grammar.

On the other hand, the performance of FDG on building dependency structures is high because all the grammars receive a high-accuracy score in *UAS* (table 7.18). *Bare* grammar scores 89.57%. Despite this, the rest of grammars working with PP-attachment knowledge does not necessarily get better results or improve the accuracy significantly compared to *Bare* grammar. In particular, only *PP-Weight* grammar exceeds the *Bare* grammar accuracy by 0.02 points. The two other grammars using knowledge learned by the naive supervised classifiers, *PP-Frequent* and *PP-Monoclass*, score lower than *Bare*. Furthermore, the grammars running with PP-attachment knowledge learned from word embeddings, *PP-Similar-F* and *PP-Similar-NF*, score lower than the *Bare* grammar and better than *PP-Frequent*.

A closer look at attachment accuracy confirms the tendency observed in the global accuracy scores (table 7.20). PP-attachment scores are low in all grammars with PP-attachment added or subtracted and only vary between 76% and 77% of accuracy. *Bare*, *PP-Weight* and *PP-Monoclass* grammars obtain the best scores. Of these the *PP-Weight* is the grammar providing most right

Grammar	Gold	System	%
Bare	5837	4506	77
PP-Weight	5837	4512	77
PP-Frequent	5837	4444	76
PP-Monoclass	5837	4476	77
PP-Similar-F	5837	4460	76
PP-Similar-NF	5837	4460	76

TABLE 7.20: Quantitative UAS results about PP-attachment in Spanish

Grammar	Gold	System	%
Bare	49	39	80
PP-Weight	49	39	80
PP-Frequent	49	39	80
PP-Monoclass	49	39	80
PP-Similar-F	49	39	80
PP-Similar-NF	49	39	80

TABLE 7.21: Qualitative UAS results about PP-attachment in Spanish

Grammar	LAS	UAS	LAS2
Bare	79.41	88.24	80.88
PP-Weight	79.23	88.05	80.70
PP-Frequent	79.23	88.05	80.70
PP-Monoclass	79.23	88.05	80.70
PP-Similar-F	79.41	88.24	80.88
PP-Similar-NF	79.41	88.24	80.88

TABLE 7.22: Qualitative evaluation scores in Catalan

Grammar	Gold	System	%
Bare	42	30	71
PP-Weight	42	29	69
PP-Frequent	42	29	69
PP-Monoclass	42	29	69
PP-Similar-F	42	30	71
PP-Similar-NF	42	30	71

TABLE 7.23: Qualitative UAS results about PP-attachment in Catalan

answers. On the other hand, *PP-Similar-F*, *PP-Similar-NF* and *PP-Frequent* score one point less and grammars with knowledge learned with word embeddings, *PP-Similar-F* and *PP-Similar-NF*, are in the fourth position with exactly the same amount of right answers.

The evaluation results of qualitative data with ParTes do not vary among the several grammars with knowledge about PP-attachment added or subtracted. *UAS* remains at 90.11% (table 7.19) and specific results of PP-attachment stay at 80% (table 7.21). These results are not very explicative without a deeper analysis. For this reason, in the following section (§7.4.3), an explanation of the result of PP-attachment knowledge integration in FDGs is provided.

Concerning Catalan FDG, the versions of the grammars working with PP-attachment knowledge added or subtracted follow a similar tendency as the Spanish grammars (tables 7.22 and 7.23). Table 7.22 shows that Catalan FDGs global accuracy scores (*LAS*) is medium-high (79.41% in *Bare*, *PP-Similar-F* and *PP-Similar-NF*, and 79.23% in *PP-Weight*, *PP-Frequent* and *PP-Monoclass*). With regard to attachment accuracy (table 7.23), *UAS* metric scores high (88.24% in *Bare*, *PP-Similar-F* and *PP-Similar-NF*, and 88.05% in *PP-Weight*, *PP-Frequent* and *PP-Monoclass*).

The attachment qualitative results (table 7.23) are parallel to the qualitative general results (table 7.22). *Bare*, *PP-Similar-F* and *PP-Similar-NF* grammars perform better than *PP-Weight*, *PP-*

Frequent and *PP-Monoclass*.

The results of the grammars with PP-attachment knowledge learned with word embeddings (*PP-Similar-F* and *PP-Similar-NF*) are the same as *Bare* grammar, and are slightly better than the grammars with PP-attachment knowledge learned with naive classifiers. Despite this, the knowledge learned with word embeddings does not contribute to a significant improvement of the FDGs accuracy. In the following section (§7.4.3), the analysis of the results will provide a deeper explanation of these results.

7.4.3 Analysis of the Results

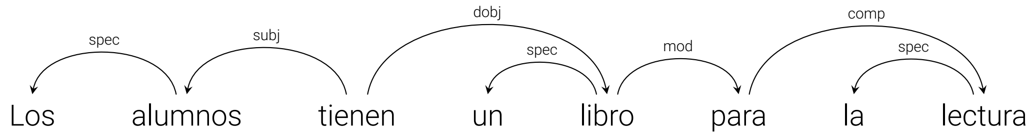
As pointed out in the previous section (§7.2), learning context-predicting vectors from simple linguistic information such as lexical information limits the disambiguation of PP-attachment. For this reason, the results of the models learned are considerably low (§7.2.4).

The evaluation results of the integration of the PP-attachment knowledge in the FDGs do not show that the addition of linguistic knowledge learned automatically helps much in the improvement of the grammars' performance accuracy neither (§7.4.2).

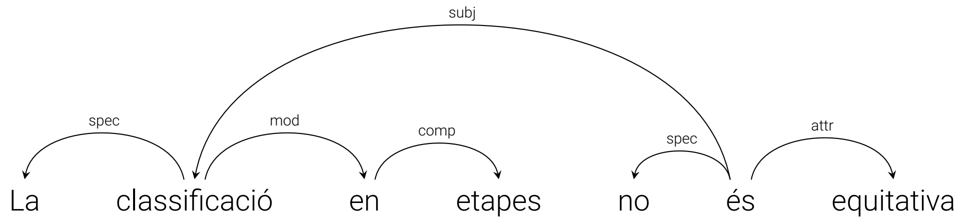
Apparently, it would seem that the low accuracy of the word embedding vectors learned from lexical information is responsible for the small impact in the parsing performance of FDGs. However, a deeper analysis explains the main reasons of these results.

ParTesEs and ParTesCa contain only a test sentence containing the pattern *VP* NP_1 *P* NP_2, which concerns a PP n-attached case (6) and which FDG performs wrong (7). In addition, since the test sample is not enough for the results to be analysed, five sentences in Spanish from the Tibidabo Treebank with the pattern *VP* NP_1 *P* NP_2 have been selected. In total, the Spanish test set contains six sentences in which three of them have a n-attached PP and the other ones have a v-attached PP.

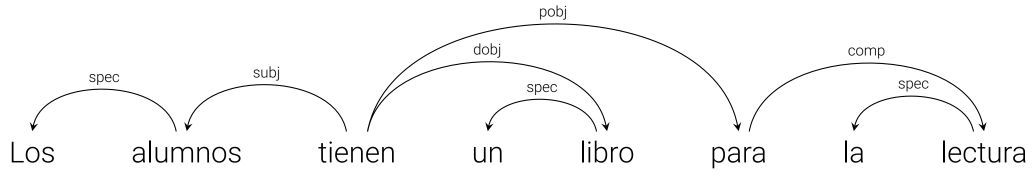
(6) a. Los alumnos tienen un libro *para la lectura*
'The students have a book for reading'



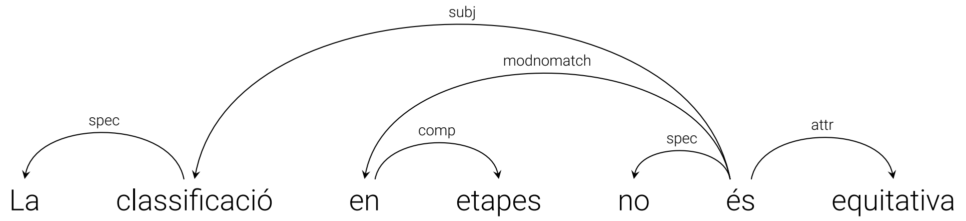
b. La classificació *en etapes* no és equitativa
The classification in phases is not equitable



(7) a. Los alumnos tienen un libro *para la lectura*
'The students have a book for reading'



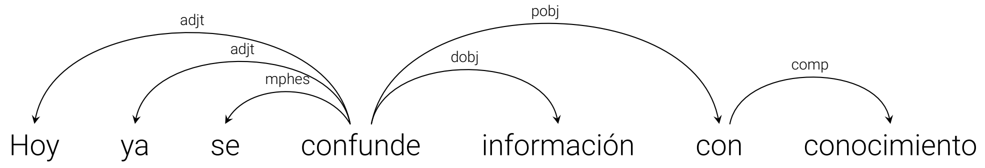
b. La classificació *en etapes* no és equitativa
The classification in phases is not equitable



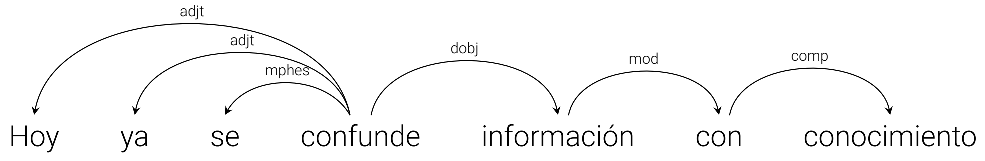
The PP-attachment in n-attached test sentences is wrong because the paired lemmas are missing in the classes of PP-attachment knowledge learned. For example, the lemmas 'libro' ('book') and 'lectura' ('reading') of the sentence (6-a) are not paired in the Tibidabo Treebank, although word embeddings models predict this paired lemmas as n-attached. For this reason, the PP with preposition 'para' ('for') is not identified by the rule of n-attached PP with prepositions other than 'de' (figure 7.1) and the default rule for v-attached PP is applied. If the pairs noun head and noun child are in the lists of lemmas, then the right attachments are performed.

In the case of v-attached PP test sentences such as (8-a), two situations are observed. While a pair of verbal head and nominal child is not in any of the models, the other two pairs are in the models, but wrongly classified and, consequently, the sentences where they occur are parsed wrong such as (8-b).

(8) a. Hoy ya se confunde información *con conocimiento*
'Today it is mixed up already information and knowledge'



b. Hoy ya se confunde información *con conocimiento*
'Today it is mixed up already information and knowledge'



For example, the sentence in (8) is captured by models learned by word embeddings, but it is wrongly represented. The cosine similarity of the vectors of the pair 'información' ('information') and 'conocimiento' ('knowledge') is higher than the similarity of the vectors of the pair 'confundir' ('to mix up') and 'conocimiento' ('knowledge'). Furthermore, since preposition neither subcategorization information is represented in the vectors of the unsupervised experiment (§7.2.4), models cannot predict that the verb 'confundir' ('to mix up') subcategorizes a prepositional argument with the preposition 'con' ('with'). For these reasons, the PP-attachment of (8-a) is classified as a strongly similar n-attached pair and, consequently, the FDG applies the rule of n-attached PP that occur with a preposition different than 'de' ('of').

From the issues analysed in these examples, several observations can be stated to implement them in the remaining set of experiments. The problems for capturing similarity relations are mainly due to two factors, the need of extending the training data and the fact of better representing the problem of PP-attachment.

Firstly, some cases are not parsed well by the FDG because they are not predicted by the models. Then, new data need to be added in the training set in order to extend the power of the embedding models.

Secondly, as observed in the results, the lexical value of the preposition and information of subcategorization are required to capture relations right. Therefore, word representation vectors need to be enriched with these layers of information, as Aguilar et al. (2011) and Belinkov et al. (2014) argue. In addition, semantic information such as Top Concept Ontology features and Semantic File used in WordNet (Fellbaum, 1998) can contribute decisively to discriminate the kind of attachment in the cases observed here (6) and (8). Lexical information limits the decisions in the learning task, but adding a layer of semantic information enriches the process of making generalizations, which have a direct effect on PP-attachment learned (Aguilar et al., 2011; Belinkov et al., 2014).

On the other hand, the PP-attachment pattern *VP* NP_1 *P* NP_2 targeted in this experiments is very specific comparing it to the global scope of PP-attachment. In the data used in the experiments (§7.2.2.3), the instances containing the pattern *VP* NP_1 *P* NP_2 in the training data (i.e. 18,873 pattern instances) corresponds to 26.97% of the total PP-attachment in the AnCora Corpus (i.e. 69,984 PP-attachment instances). Consequently, if the disambiguation of PP-attachment is improved in the learning task, this contribution can only be rather than small in the parsing performance. Therefore, to have a positive impact in the grammar performance, new patterns containing different grammatical categories need to be captured by the models.

Recapitulation

This chapter focused on the first part of a series of experiments for automatically disambiguating PP-attachment with word embeddings in order to improve the accuracy of FDGs. Firstly, a preliminary experiment based on a supervised classifier have been described (§7.1). Then, the experiments carried out in this proposal have been explained (§7.2). The main methodology and trends of distributional models have been presented in order to set the methodology of the experiments (§7.2.1).

After this explanation, the rest of the chapter focused on the development and results of the experiments, describing the data used (§7.2.2), the naive classifiers designed (§7.2.3) and the word embedding models learned (§7.2.4). Finally, the integration in FDGs of the knowledge learned in the experiments have been described. A complete evaluation task of FDGs with PP-attachment knowledge added has been provided in order to validate how statistical knowledge integration contributes to the accuracy has been detailed.

The results of the unsupervised experiment and the evaluation of FDGs working with statistical information show that PP-attachment can be disambiguated following the strategy proposed here. Word embeddings have made it possible to build language models whose recall is high, but whose precision is low. Despite this, the integration of the knowledge learned in FDGs have not improved grammars' accuracy.

More precisely, the results point out that context-predicting vectors need to be enriched with syntactic and semantic information in order to provide more robust generalizations. Moreover, the representation of the pattern which has been studied was very simple. The analysis of errors has demonstrated that the pattern needs to be represented with the preposition's lexical value. Finally, it has been observed that a very specific pattern has been targeted. In order to improve the grammars' performance, the number of patterns studied needs to be widen.

Chapter 8

Improving Argument Recognition

In the chapter §3, the ambiguities in natural language have been discussed and, in particular, the difficulties that NLP tools have to overcome some of them have been described. This chapter focuses on the limitations of parsers to detect the arguments of the verb predicate and to assign to them the correct dependency relation. As previously noted in §3.3, a parser by itself has a lot of difficulties to distinguish the different nature of the prepositional phrase in some contexts, e.g. 'a los alumnos' in the sentences 'La profesora lleva *a los alumnos* lecturas nuevas' (indirect object) and 'La profesora lleva *a los alumnos* al teatro' (prepositional object).

The rule-based dependency grammars developed in this proposal need to handle this issue. The parse trees that the grammars generate express the verb predicate structure represented by dependency relations. This operation is performed by a set of rules called labelling rules in the FDGs. The evaluation of the grammars shows that the Spanish and Catalan versions have a high accuracy, but they have some limitations on the argument recognition and, specifically, on some arguments like the prepositional object and the predicative.

This chapter is focused on providing a solution in order to improve the argument recognition in FDGs. In particular, the experiment described in this chapter aims to prove the first hypothesis of this proposal about the fact that linguistic knowledge contributes to improve the accuracy of rule-based dependency grammars such as FDGs. For this reason, the approach taken in this issue is one that is based on the acquisition of subcategorization patterns from existing resources of the languages included in this proposal (§8.1).

In order to answer this hypothesis, subcategorization frames have been automatically extracted from the SenSem Corpus (Vázquez and Fernández-Montraveta, 2015) and organized in coarse-grained classes following Alonso et al. (2007) (§8.1.1). Later, the subcategorization classes have been redesigned in order to integrate more fine-grained classes in order to test them in the grammars (§8.1.2), and they have been integrated in the grammars as explained in section §8.1.3. Finally, an exhaustive quantitative and qualitative evaluation has been carried out to argue for the contribution of the subcategorization information integration in the performance of the grammar (§8.2).

After the evaluation task, FDGs with highest accuracy scores are compared to other grammars and statistical parsers. The comparison will determine the status of the grammars developed in this proposal with the state of the art proposals in parsing (§8.3).

8.1 Acquisition of Subcategorization Information

A parser is extremely limited in the recognition of the verb predicate structure and, consequently, the probability of assigning the right dependency relation is low, as argued in section §3.3. This assumption has been confirmed in the evaluation of the FDGs detailed in §6.4. In that evaluation task, it has been pointed that particular syntactic realizations of arguments and adjuncts (i.e. prepositional phrases and adjectival phrases) are difficult to capture by the rules of FDGs, so the grammars tend to assign a wrong dependency relation in those cases.

Following the hypothesis that subcategorization frames improve the parsing performance (Carroll et al., 1998; Zeman, 2002), in this proposal a lexicon of subcategorization frames for both Spanish and Catalan languages is developed from automatically acquiring linguistic knowledge from syntactically annotated corpus.

Currently, there are two resources available containing both languages considered in this proposal: AnCora-Verb (Aparicio et al., 2008) and SenSem Verbal Lexicon (Vázquez and Fernández-Montraveta, 2015). As shown next, both resources are computational verbal lexicons with syntactic and semantic information which have been automatically extracted from manually annotated text.

AnCora Verb
This lexicon has been built by extracting the verb subcategoritzacion frames from the verbal forms occurrences syntactically and semantically annotated in the AnCora Corpus (Taulé et al., 2008). They are automatically encoded with the verb aspect class which they belong to, the set of arguments encoded with the PropBank tags (Palmer et al., 2005), the syntactic functions and semantic roles realized by the arguments. The Spanish version of the lexicon contains 5516 verbal entries corresponding to 2820 verbal lemma distributed in 88 subcategorization frame patterns, which correspond to a ratio of 62.68 verbs per subcategorization frame pattern. Concerning the Catalan version, the lexicon includes 4635 verbal entries corresponding to 2248 verbal lemma distributed in 66 subcategorization frames, which is a ratio of 70.23 verbs for every frame.

SenSem Verbal Lexicon
This lexicon contains 30,000 sentences for the 250 most frequent Spanish verbs which are distributed in 69 subcategorization frames in a ratio of 67.17 verbs per frame. Every lexical entry is encoded with several levels of linguistic information: syntactic (constituents and syntactic functions expressing the arguments of the predicate structure), semantic (verb sense corresponding to a synset of WordNet, semantic roles) and discursive (language register). The Catalan version of the lexicon has been automatically translated and manually

reviewed. It includes 20,000 sentences which correspond to 2680 verbal occurrences with different subcategorization frames distributed in 71 verb subcategorization patterns with a ratio of 37.75 verbs per frame.

Both lexicons are suitable for the acquisition of subcategorization frames for the FDGs. However, as a starting point, this proposal carried out the acquisition with the SenSem Verbal Lexicon for two reasons. The lexicon is based on the 250 most frequent verbs of Spanish, so the FDGs can cover the most frequent verbs. Furthermore, every verb is equally represented in the linguistic examples, while AnCora-Verb only includes the patterns found in the corpus, which causes the subcategorization patterns to be unevenly represented. For this reason, the acquisition of subcategorization frames with SenSem Verbal Lexicon is prioritized.

8.1.1 Initial Subcategorization Frames

To integrate subategorization frames in FDGs (Lloberes et al., 2010), the syntactic patterns and their verbs associated with the SenSem Verbal Lexicon (Vázquez and Fernández-Montraveta, 2015) have been automatically extracted. By the time this work was carried out, the Catalan version of the SenSem Verbal Lexicon was not available. For this reason, the Catalan data of Volem Multilingual Lexicon (Fernández et al., 2002) has been used and, specifically, the subcategrization frames annotated with syntactic information (arguments and morphosyntactic categories) have been automatically extracted from the database.

The extracted frames have been classified following the conclusions pointed out by Alonso et al. (2007). These authors propose several experiments to automatically classify verbs according to subcategorization classes using morphosyntactic categories, syntactic functions and semantic roles isolated or combined. Their experiments show that the combination of morphosyntactic categories and syntactic functions allows one to capture reliable subcategorization classes, while adding semantic roles to the morphosyntactic categories and syntactic functions increases the sparseness and boundaries between classes are less clear. Furthermore, as a result of the automatic classification task, some subcategorization classes emerged with regard of the distribution of the number and type of arguments (transitive, intransitive, ditransitive and verbs with prepositional arguments).

According to this, the classes of Alonso et al. (2007) have been applied over the SenSem Verbal Lexicon (Vázquez and Fernández-Montraveta, 2015). In addition, six new classes have been added in order to capture copulative constructions, sentences with arguments realized as an adjective phrase, impersonal verbs realized with any arguments, movement verbs which subcategorize two prepositional arguments, verbs expressing indirect object and verbs that subcategorize a subordinate clause. The final lexicon applied to the FDGs has 11 subcategorization classes containing a total of 1326 Spanish verbal lemmas and 2937 Catalan verbal lemmas with a different subcategorization frame (table 8.1).

A first experimental evaluation of the Spanish Grammar with the initial subcategorization lexicon (Lloberes et al., 2010) described in §6.3.1.1 showed that incorporating subcategorization

Frames	Spanish	Catalan
intransitive	43	60
transitive	38	831
ditransitive	158	549
prepositional-1	355	330
prepositional-2	11	35
movement	148	116
indirect object	130	31
predicative	11	14
copulative	3	3
impersonal	5	166
argument *wh-*	153	153

TABLE 8.1: Initial Subcategorization Frames in numbers

information promises to improve the performance of the grammars on recognizing arguments. In that task, *LAS* score 73.88% and some arguments and the adjunct are poorly recognized in the FDGs, such as prepositional argument (50%), indirect object (58.82%) and adjunct (52.18%). On the other hand, some subcategorization classes are underpopulated (e.g. intransitive and transitive verbs in Spanish) compared to other classes (e.g. prepositional verbs and predicative verbs). Classes such as intransitive and transitive verbs need to be better represented since they tend to be frequent in real data (Alonso et al., 2007)

8.1.2 Redesign of Subcategorization Frames

According to the evaluation results of the grammars with the initial subcategorization lexicon included, the lexicon has been redesigned (Lloberes et al., 2015a), proposing a set of more fine-grained subcategorization frame classes in order to represent verb subcategorization in the dependency rules in a controlled and detailed way.

New syntactic-semantic patterns have been automatically extracted directly from the SenSem Corpus (Fernández and Vàzquez, 2014). The idea of extracting the subcategorization patterns from the SenSem Verbal Lexicon (Vázquez and Fernández-Montraveta, 2015) has been abandoned. In this lexicon, the information is not compacted in the lexical entry and syntactic and semantic patterns (e.g. syntactic functions and semantic roles) need to be inferred (e.g. the annotation 'topicalization of the logical subject' corresponds to a pattern with an argument realized as a subject and another argument realized as direct object), which makes the acquisition process more complex and more expensive in time and effort. Furthermore, in the lexicon it is not possible to access the information about the pronominal particle *es*, which also determines the type of construction and its arguments (e.g. pronominal verb with pronominal particle, pronominal passive or impersonal).

Subcategorization patterns and their corresponding verbal lemma are extractedd according to the idea that every verbal lemma with a different subcategorization frame expresses a different meaning. Therefore, a new lexicon entry is created every time an annotated verbal lemma with a different frame is detected.

The lexicon acquired, named Computational Lexicon of Verb Subcategorization (CompLex-VS), contains 3102 syntactic patterns in the Spanish lexicon and 2630 patterns in the Catalan lexicon (see section 8.1.2 for detailed numbers). These are organized into 15 subcategorization frames as well as into 4 subcategoriztion classes. This organization configures the lexicon into a hierarchical structure of three levels in which subcategorization classes are the highest level, subcategorization frames depend on the subcategorization classes and syntactic patterns are the lowest level (figure 8.1).

The acquisition process identifies patterns with different word order belonging to the same frame (e.g. a pattern *subject – direct object – verb* is considered the same as *subject – verb – direct object* under the class of transitive verbs). On the other hand, certain patterns have been discarded because they are non-prototypical in the corpus (e.g. clitic left dislocations), they alter the sentence order (e.g. relative clauses), or they involve controversial argument classes (e.g. prepositional phrases seen as arguments or adjuncts depending on the context).

As figure 8.1 shows, the extracted patterns (*<verb>*) have been classified into *<frame>* classes according to the whole set of argument structures occurring in the corpus (*subj* for intransitive verbs, *subj,dobj* for transitive verbs, etc.). Simultaneously, frames have been organized in *<subcategorization>* classes (monoargumental, biargumental, triargumental and quatriargumental).

```
<subcategorization
    class="monoargumental"
    ref="1" freq="0.188480">
    <frame class="subj" ref="1" freq="0.188480">
        <verb lemma="pensar" id="2531" ref="1:1" fs="subj" cat="np" rs="exp"
            head="null" construction="active" se="no" freq="0.000070"/>
    </frame>
</subcategorization>
<subcategorization
    class="biargumental"
    ref="2" freq="0.733349">
    <frame class="subj,dobj" ref="2"
        freq="0.617452">
        <verb lemma="agradecer"
            id="454" ref="2:2" fs="subj,dobj" cat="np,complsc" rs="ag_exp,t"
            head="null,null" construction="active" se="no" freq="0.000140"/>
    </frame>
</subcategorization>
```

FIGURE 8.1: Example of the CompLex-VS

Every lexicon entry contains the syntactic function of every argument (*fs*), the grammatical category of the head of the argument (*cat*) and the thematic role (*rs*). The type of *construction* (e.g. active, passive, impersonal, etc.) has been inferred from the predicate and aspect annotations available in the SenSem Corpus.

Two non-annotated lexical items of the sentence have also been inserted into the subcategorization frame because the information that they provide is crucial for the argument structure configuration (e.g. the particle '*se*' and the lexical value of the prepositional phrase *head*).

In addition, meta-linguistic information has been added to every entry: a unique *id* and the relative frequency of the pattern in the corpus (*freq*). A threshold frequency has been established at $7{\cdot}10^{-5}$ (Spanish) and at $8.5{\cdot}10^{-5}$ (Catalan). Patterns below this threshold have been considered marginal in the corpus and they have been discarded.

Every pattern contains a link to the frame and subcategorization class that they belong to (*ref*). For example, if an entry has the reference *1:1*, it means that the pattern corresponds to a monoargumental verb whose unique argument is a subject.

8.1.3 Integration of CompLex-VS in the FDGs

From the CompLex-VS, two derived lexicons per language containing the verbal lemmas for every recorded pattern have been created to be integrated into the FDGs (Lloberes et al., 2015a). The CompLex-SynF lexicon contains the subcategorization patterns generalized by the syntactic function (Table 8.2). The CompLex-SynF+Cat lexicon collects the syntactic patterns combining syntactic function and grammatical category (adjective, noun or prepositional phrase, and infinitive, interrogative or completive clause).

The addition of grammatical categories makes it possible to restrict the grammar rules. For example, a class of verbs containing the verb 'quedarse' ('to get') whose argument is a predicative and a prepositional phrase allows the rules to identify that the prepositional phrase of the sentence 'Se ha quedado *de piedra*' ('$Ø_{3sg}$ got shocked') is a predicative argument. Furthermore, it allows for discarding the prepositional phrase of the sentence 'Aparece *de madrugada*' ('$Ø_{3sg}$ shows up at late night') being a predicative argument, although *aparecer* belongs to the class of predicative verbs but conveying a noun phrase as argument.

While in the CompLex-SynF lexicon the information is more compacted (1054 syntactic patterns classified in 15 frames), in the CompLex-SynF+Cat lexicon the classes are more granular (1356 syntactic patterns organized in 77 frames).

Only subcategorization patterns corresponding to lexicon entries referring to the active voice have been integrated in the FDGs, since they involve non-marked word order. Both lexicons also exclude information about the thematic role, although they take into account the value of the head (if the frame contains a prepositional argument) and the pronominal verbs (lexical entries that accept 'se' particle whose value neither is reflexive nor reciprocal).

Two versions of the Spanish dependency grammar and two versions of the Catalan dependency grammar have been created. One version contains the CompLex-SynF lexicon and the other one the CompLex-Synf+Cat.

Frames	**Spanish**	**Catalan**
subj	203	386
subj,att	3	7
subj,dobj	440	230
subj,iobj	37	61
subj,pobj	126	93
subj,pred	45	31
subj,attr,iobj	2	1
subj,dobj,iobj	113	72
subj,dobj,pobj	42	34
subj,dobj,pred	21	18
subj,pobj,iobj	2	1
subj,pobj,pobj	14	9
subj,pobj,pred	1	0
subj,pred,iobj	4	5
subj,dobj,pobj,iobj	1	0

TABLE 8.2: CompLex-SynF lexicon in numbers

The initial lexicon classes of the FDGs (§8.1.1) have been replaced with the new ones. Specifically, this information has been inserted in the part of the labelling rules about the syntactic properties of the parent node (observe *p.class* in figure 8.2 repeated in this section as 8.2).

```
grup-verb   dobj
            d.label=grup-sp
            p.class=trans
            d.side=right
            d.lemma=a|al
            d:sn.tonto=Human
            d:sn.tonto!=Building|Place
```

FIGURE 8.2: Labelling rule with subcategorization information

Finally, new rules have been added for frames of CompLex-SynF and CompLex-SynF+Cat that are not present in the initial lexicon. Furthermore, some rules have been disabled for frames of the initial lexicon that do not exist in the CompLex-SynF and CompLex-SynF+Cat lexicons (see table 8.3 for the detailed size of the grammars).

Grammar	Spanish	Catalan
Bare	392	509
Baseline	525	780
SynF	604	917
SynF+Cat	602	917

TABLE 8.3: Labelling rules in the grammars evaluated

8.2 Evaluation of Dependency Relations Labelling

An evaluation task has been carried out to test empirically how the FDGs performance changes when subcategorization information is added or subtracted (Lloberes et al., 2015a). Several versions of the grammars including or not subcategorization information have been tested using the two data sets to perform this task from a quantitative and qualitative point of view, the Tibidabo Treebank harmonized to FDGs (only available in Spanish) and the ParTes test suite (Spanish and Catalan), respectively. For this reason, this evaluation provides quantitative and qualitative results of the FDGs performance. Specifically, it focuses on the explanation of the reasons why particular arguments are recognized better and other ones not in FDGs. Quantitative and qualitative analyses complement and help to provide a global picture of the performance of the Spanish and Catalan dependency grammars.

8.2.1 Evaluation Experiments

Four versions of both Spanish and Catalan grammars are tested in order to assess the differences in performance depending on the subcategorization information added.

- **Bare**. Version of the grammar previously presented in §6.4 running without subcategorization frames.
- **Baseline**. Version of the grammar running with the initial lexicon for the FDGs (§8.1.1).
- **SynF**. Version of the grammar running with the new CompLex-SynF including syntactic functions (§ 8.1.2).
- **SynF+Cat**. Version of the grammar running with the new CompLex-Synf+Cat including syntactic functions and grammatical categories (§8.1.2).

These versions contain a different number of labelling rules (table 8.3) according to the number of rules created for integrating the subcategorization classes.

Grammar	LAS	UAS	LAS2
Bare	81.52	89.57	83.95
Baseline	82.34	89.57	84.82
SynF	84.26	89.57	86.77
SynF+Cat	84.26	89.57	86.77

TABLE 8.4: Quantitative evaluation scores in Spanish

Grammar	LAS	UAS	LAS2
Bare	80.99	90.11	81.94
Baseline	82.70	90.11	83.84
SynF	83.27	90.11	84.41
SynF+Cat	83.08	90.11	84.22

TABLE 8.5: Qualitative evaluation scores in Spanish

The method applied to evaluate the grammars corresponds to the empirical method established for this proposal (§6.1). Since the experiment presented here is focused on the implementation of subcategorization information for argument recognition in dependency relations, only the **labelling rules** are discussed in this section. In particular, the analysis of the precision and the recall of the grammars is based on the results of the dependency relations with correct attachments. Therefore, the errors in labelling as a consequence of wrong attachments are discarded. Although the focus is placed on the labelling rules, metrics related to **linking rules** are also mentioned to provide a general description of the FDGs.

8.2.2 ACCURACY RESULTS

The global results of the FDGs evaluation (LAS) show that the whole set of evaluated grammars scores over 80% accuracy in Spanish (tables 8.4 and 8.5) and around 80% in Catalan (table 8.6).

In the four Spanish grammar versions, the correct head (UAS) has been identified in 89.57% of the cases in the quantitative analysis (table 8.4) and 90.11% of the cases in the qualitative evaluation (table 8.5).

On the other hand, the tendency changes in dependency relation labelling (LAS2). The *Baseline* establishes that 84.82% of tokens have the correct syntactic function tag in the quantitative analysis and 83.84% tokens in the qualitative analysis. However, *Bare* drops 0.87 points in the quantitative evaluation and 1.9 points in the qualitative evaluation. *SynF* and *SynF+Cat* improve 1.95 points with respect to the baseline in the quantitative data, while in the qualitative evaluation *SynF* increases 0.57 points with respect to the baseline and *SynF+Cat* rises up to 0.38 points in performance.

A parallel behaviour is observed in Catalan. The scores are slightly lower than Spanish, though (table 8.6). The four Catalan grammars score 88.24% in attachment (UAS). The *Baseline* scores 83.64% in syntactic function assignment (LAS2). Once again FDGs performs worse without subcategorization information (2.76 points less in *Bare* grammar), but also with subcategorization information from CompLex-VS (1.65 points less in *SynF* and 2.02 points less in *SynF+Cat*).

From a general point of view, accuracy metrics show a medium-high accuracy performance of all versions of FDGs in both languages. Specifically, these first results highlight that subcat-

Grammar	LAS	UAS	LAS2
Bare	79.41	88.24	80.88
Baseline	81.80	88.24	83.64
SynF	80.15	88.24	81.99
SynF+Cat	79.78	88.24	81.62

TABLE 8.6: Qualitative evaluation scores in Catalan

Tag	Bare	Baseline	SynF	SynF+Cat
adjt	61.04	56.34	60.38	60.38
attr	92.57	92.12	92.21	92.21
dobj	82.01	83.61	85.14	85.14
iobj	80.49	64.97	82.86	82.86
pobj	21.39	28.88	60.45	60.45
pred	37.41	83.33	66.67	66.67
subj	89.86	89.83	89.83	89.83

TABLE 8.7: Quantitative labelling precision scores in Spanish

Tag	Bare	Baseline	SynF	SynF+Cat
adjt	49.57	67.23	78.65	78.65
attr	87.47	83.37	85.64	85.64
dobj	69.53	68.87	72.86	72.86
iobj	28.95	37.43	33.92	33.92
pobj	84.93	63.11	79.43	79.43
pred	62.65	6.02	55.42	55.42
subj	78.55	79.38	81.78	81.78

TABLE 8.8: Quantitative labelling recall scores in Spanish

egorization information helps with the syntactic function labelling. However, a deeper analysis of the precision and recall results will reveal how subcategorization influences the grammars' performance (sections 8.2.3 and 8.2.4).

8.2.3 PRECISION RESULTS

As observed in the accuracy results (section 8.2.2), in both languages most of the syntactic function assignments drop in precision when subcategorization classes are blocked in the grammar (table 8.4, table 8.5 and table 8.6), whereas syntactic function labelling tends to improve when subcategorization is available.

In particular, the results of prepositional object (*pobj*) varies extremely depending on the blocking or incorporation of subcategorization classes. The precision for both languages drops drastically when subcategorization is disabled (*Bare*). On the contrary, the precision improves significantly when the rules include subcategorization information (*Baseline*). Furthermore, the introduction of more fine-grained frames (*SynF* and *SynF+Cat*) helps the grammars to increase the precision in 31.57 in the quantitative evaluation and in 20 points in the qualitative evaluation.

This tendency is also observed in subject, direct object, attribute and predicative. Despite these improvements, some items differ from this general tendency.

In Spanish, the improvement of the copulative verbs (*attr*) is due to lexical information in the *Bare* FDG, while they quite remain stable in *Baseline*, *SynF* and *SynF+Cat*. The lower scores are

Tag	Bare	Baseline	SynF	SynF+Cat
adjt	39.47	44.64	45.45	45.45
attr	90.32	96.55	93.33	93.33
dobj	76.19	78.05	80.95	80.49
iobj	100.00	100.00	100.00	100.00
pobj	33.33	50.00	70.00	70.00
pred	20.00	100.00	33.33	33.33
subj	95.12	95.24	95.45	95.45

TABLE 8.9: Qualitative labelling precision scores in Spanish

Tag	Bare	Baseline	SynF	SynF+Cat
adjt	38.46	64.10	64.10	64.10
attr	100.00	100.00	100.00	100.00
dobj	82.05	82.05	87.18	84.62
iobj	28.57	28.57	28.57	28.57
pobj	90.91	72.73	63.64	63.64
pred	50.00	50.00	50.00	50.00
subj	76.47	78.43	82.35	82.35

TABLE 8.10: Qualitative labelling recall scores in Spanish

due to some prepositional phrases labelled as attribute instead of prepositional object in the case of *SynF*, *SynF+Cat* and *Baseline* itself.

Concerning the indirect object, the *Bare* grammar performs precisely, although the incorporation of fine-grained subcategorization classes in *SynF* and *SynF+Cat* increases the precision. There is a drop in the *Baseline* because of problems in recognizing dative clitics in that particular grammar. Precision remains the same for the indirect object (*iobj*) in the qualitative approach because the cases of indirect object detected correspond to dative clitics in singular and morphological information is enough for FDG.

The improvement of predicative (*pred*) in the grammars is related to the lack or addition of subcategorization classes similar to the prepositional object. It should be expected that fine-grained classes of *Synf* and *Synf+Cat* grammars perform better than coarse-grained classes of *Baseline* grammar as the qualitative analysis shows. However, the quantitative results are better for the *Baseline* grammar than the *Synf* and *Synf+Cat* grammars. A closer observation of these results reveals that the *Baseline* is precise on recognizing this type of dependency relation, but the recall is very low (6.02% as shown in table 8.8). Actually, the low recall involves that the *Baseline* is very precise in recognizing the very few verbs of the ParTes that subcategorize a prepositional object and predicative. As will be observed in §8.2.4 about recall results, this situation has also strong consequences for the number of verbs detected.

Adjunct (*adjt*) and, specifically, adjuncts realized by a prepositional phrase are implemented with general rules that make use of very little subcategorization information. Consequently, at the moment that the grammar handles adjuncts, it is expected that arguments realized by a preposition phrase are recognized already. Therefore, any prepositional phrase that the grammar does not recognize as an argument tends to be labelled as an adjunct. Actually, this particular treatment of the adjunct causes the low precision on adjunct recognition.

Concerning the Catalan version of FDGs, it shows a similar behaviour to Spanish. Catalan grammars follow the general tendency that subcategorization information contributes to the performance, but it is slightly diffused (table 8.11). The fine-grained classes in *SynF* and *SynF+Cat*

Tag	Bare	Baseline	SynF	SynF+Cat
adjt	56.00	59.52	51.02	51.02
attr	90.00	78.26	78.26	72.73
dobj	72.55	84.78	85.00	82.93
iobj	100.00	75.00	100.00	100.00
pobj	45.83	60.00	55.56	55.56
pred	22.22	100.00	25.00	33.33
subj	84.85	87.50	87.88	82.86

TABLE 8.11: Qualitative labelling precision scores in Catalan

Tag	Bare	Baseline	SynF	SynF+Cat
adjt	46.67	83.33	83.33	83.33
attr	90.00	90.00	90.00	80.00
dobj	88.10	92.86	80.95	80.95
iobj	66.67	100.00	66.67	66.67
pobj	84.62	69.23	38.46	38.46
pred	100.00	50.00	50.00	50.00
subj	65.12	65.12	67.44	67.44

TABLE 8.12: Qualitative labelling recall scores in Catalan

helps to increase the precision of arguments like *subj* (only in *SynF*), *dobj*, *iobj* and *pobj*. Once more the prepositional object (*pobj*) performance increases when subcategorization frames are available.

While in general the Spanish grammars with the fine-grained classes show better performance than the grammars with coarse-grained classes, in the Catalan FDG this distinction is not obvious. The qualitative analysis shows that errors are due to missing verbs in the subcategorization classes integrated in the Catalan grammars.

Despite this, there is a drop in precision in *Bare* with respect to the *Baseline* in the majority of arguments and the adjunct, except for the attribute (*attr*) and the indirect object (*iobj*). These two arguments are still precisely recognized without subcategorization information because *Bare* grammar uses lexical information in the attribute recognition and morphological information for recognizing dative clitics.

The results of *SynF* and *SynF+Cat* are almost identical in both languages. There are some minor changes in the subcategorization classes as a result of the different distribution of verbs in classes. For example, the verb 'resultar' in Catalan ('to be' or 'to be considered') is in the class of copulative verbs subcategorizing an argument realized as an adjective phrase and verbs subcategorizing a predicative argument as a prepositional phrase. This behaviour can only be caught when subcategorization adds grammatical information. For this reason, the precision rises from 25% in SynF to 33.33% in *Synf+Cat*. Despite this, in the majority of cases, the addition of grammatical information in the subcategorization classes does not have any effect. Therefore, in general grammatical categories do not make a contribution to the improvement of the precision, except in very particular cases.

8.2.4 RECALL RESULTS

The addition of subcategorization information in the FDGs also contributes to the improvement of the recall, as can be observed in the quantitative results (table 8.8) of *SynF* and *SynF+Cat* regarding the adjunct (*adjt*), the direct object (*dobj*) and the subject (*subj*). However, the *Baseline* shows some limitations in capturing the dependency relations of the arguments and the adjunct because the classes integrated in the grammar need to be better populated.

In the qualitative evaluation, the tendency that subcategorization can improve the recall of the arguments is present in the majority of cases, except for the prepositional object (*pobj*) and the direct object (*dobj*) in *SynF+Cat*. The fact that the results are better in the qualitative results than in the quantitative results shows that the lower results in recall are due to the lack of verbs in the classes.

In the results for the Catalan FDG (table 8.12), this situation is reflected explicitly. The overlap between the verbs of the CompLex-VS lexicon and the set of verbs in ParTes is small. For this reason, the results of the grammars *SynF* and *SynF+Cat* are low, even lower than the *Bare* grammar sometimes (e.g. direct object, indirect object, prepositional object and predicative). In this respect, the *Baseline* grammar has better populated classes for some arguments (e.g. attribute, direct object and indirect object) and for the adjunct than *SynF* and *SynF+Cat*.

In particular, the *Bare* grammar gets higher scores on the recall of the attribute (*attr*), the prepositional object (*pobj*) and the predicative (*pred*). These results are possible because lexical information in the copulative verbs allows it to capture successfully the attribute, and in the case of the prepositional object and the predicative the rules that detect these arguments behave as general rules when the subcategorization information is blocked. This is the reason why the rules do not successfully capture the adjunct. Since rules for prepositional object and predicative are generic and apply before adjunct rules, most adjuncts are wrongly labelled as a prepositional object or a predicative.

Once again there are no significant differences between *SynF* and *SynF+Cat* in Catalan. This reinforces the idea that grammatical categories are not decisive for capturing new argument and adjuncts.

8.2.5 Analysis of the Results

The whole set of experiments demonstrates that subcategorization improves significantly the performance of the rule-based FDGs. Therefore, the first hypothesis of this proposal about stating that linguistic knowledge contributes to an improvement in a dependency grammar's accuracy is confirmed.

However, this statement needs to be detailed. Subcategorization does not work the same way for every argument nor for the adjunct. Some arguments and, specifically, the prepositional object and the predicative, are difficult to capture without subcategorization information. As observed, the addition of subcategorization classes in the rules makes it possible to assign them the correct dependency relation, otherwise they are misslabelled most of the time (e.g. with the adjunct in the FDGs).

On the other hand, there are other arguments, such as the attribute, the direct object, the indirect object and the subject that are not as strictly restricted by the addition of subcategorization classes in the rules in order to be labelled successfully. Copulative verbs tend to belong to a closed reduced list, so they can be successfully handled only with lexical information. The subject, the direct object and the indirect object recognition is guaranteed by the labelling rules itself (e.g. the majority of subjects can be captured by defining their position in the sentence).

In general, the fine-grained classes of CompLex-VS lexicon added to the grammars *SynF* and *SynF+Cat* perform better than the initial coarse-grained classes of the FDGs (*Baseline*). This argument does not implicitly point to the fact that the more fine-grained the subcategorization classes are the better the performance of a rule-based grammar on the dependency relations recognition. If this was true, the *SynF+Cat* grammar, which contains information generalized by syntactic function and grammatical categories, would perform better than the grammar *SynF*, in which the subcategorization classes are generalized only by syntactic functions. These two grammars perform almost identically which shows that the grammatical categories are not relevant in the argument recognition in the FDGs.

The classification of frames proposed in the CompLex-VS resource is consistent with the methodology and ensures high performance of the dependency rule-based grammars that have been developed. Furthermore, it is an essential resource for the grammars tested since it ensures medium-high precision results in *LAS* metric (compared to grammars without subcategorization classes, which always score lower). On the other hand, the CompLex-VS lexicon shows some limitations on the recall of the FDGs because it needs to be populated with new verbs, since some arguments are not captured properly because the verb is missing in the lexicon.

8.3 Comparison of FDGs

In this section, best FDGs in accuracy is compared with other rule-based approaches and dependency statistical parsers to determine the status of the grammars developed in this proposal with the state of the art proposals in parsing. In order to perform an accurate comparison, the same gold standard needs to be used in the evaluation and in the training task, in the case of statistical parsers. Despite this, among the proposals observed, a different evaluation corpus is used. For this reason, the comparison is an approximative description of the current state of the art of parsing performance.

From the results of experiments about PP-attachment (§7) and argument recognition (§8), best accuracy results tend to be observed in the dependency grammar SynF, i.e., FDG working with linguistic knowledge about verb subcategorization frames generalized by syntactic function. For this reason, FDG-SynF is considered the best dependency grammar developed in this proposal to compare with other rule-based proposals and other dependency parsing proposals from the statistical approach perspective.

With regard to state of the art of rule-based approach in Spanish, as noticed in the revision of grammars and parsers available in Spanish (§2.3), the list is short and only six proposals are found: HISPAL (Bick, 2006), Connexor (Tapanainen and Järvinen, 1997), DILUCT (Calvo and Gelbukh, 2006), DepPattern (Gamallo, 2015), Slot Unification Parser (Ferrández and Moreno, 2000) and Spanish Resource Grammar (Marimón, 2010). Among these rule-based solutions, some of them have been evaluated (i.e. HISPAL, Connexor, DILUCT and DepPattern).

The most similar evaluation to the method proposed in this thesis is the task carried out Gamallo (2015). However, it uses a small sub-set of AnCora Corpus (3,000 sentences). HISPAL evaluation is not comparable because uses a very small evaluation data set and the object observed (i.e. syntactic function in-clause, syntactic function subclause and boundness) are completely different than the object evaluated here (i.e. LAS, UAS and LAS2). Connexor and DILUCT have been evaluated as well using a reduced data set (190 random sentences from a previous version of AnCora Corpus), that seems more suitable for a qualitative evaluation than a quantitative evaluation.

Grammar	**LAS**	**UAS**
FDG	**84.26**	**89.57**
DepPattern	— — —	84.60
Connexor	55.00	— — —
DILUCT	47.00	— — —
HISPAL	— — —	— — —
SUP	— — —	— — —
SRG	— — —	— — —

TABLE 8.13: Comparison of Spanish FDG with existent Spanish rule-based grammars

A ranking of rule-based proposals and FDG (table 8.13) shows that FDG is in the top position in both metrics (i.e. LAS and UAS). The attachment accuracy in DepPattern only has been evaluated and stays 4.97 points below FDG. On the other hand, Connexor and DILUCT score very low in LAS. Furthermore, this ranking shows the current status of evaluation methods in rule-based approaches. Not all the proposals are evaluated. In the case that the grammars are evaluated, a quantitative evaluation is carried out. However, they are measured by a small data set, which makes it difficult to inform about the real performance of the grammar. These evaluations are also partial because in any case both metrics that inform about the accuracy of attachments and labellings are calculated.

FDG has also been compared to Spanish statistical parsers. More precisely, this comparison includes a Maximum Entropy ranker of SRG trained with Tibidabo Treebank (Marimon et al., 2014), MaltParser trained with Tibidabo Treebank (Marimon et al., 2014), and all the parsers that took part of the CoNLL edition of 2009 trained with AnCora Corpus (Hajič et al., 2009).

The Maximum Entropy parser (MaxEnt$_{Tibidabo}$) is an implementation of Max-Ent-based parser ranker of Toutanova et al. (2005). The trees of SRG are ranked automatically by the parser and the best analysis is selected automatically. The results of the parser are evaluated with 1,428 sentences of the corpus.

Furthermore, the same authors (Marimon et al., 2014) train the MaltParser (Nivre, 2003) with a sub-set of the Tibidabo Corpus (Malt$_{Tibidabo}$) and evaluate this parser with the same 1,428 sentences of the corpus used in the MaxEnt$_{Tibidabo}$ evaluation.

Spanish FDG, MaxEnt$_{Tibidabo}$ and Malt$_{Tibidabo}$ use the same resource to evaluate, Tibidabo Treebank. However, the sub-corpus used in the evluation of FDG and both parsers is different and annotated with different syntactic criteria. Consequently, the comparison between these tools is approximative.

On the other hand, the results of CoNLL 2009 task (Hajič et al., 2009) are relevant compared to the previous editions of the competition. In particular, the edition in 2009 has been focused on dependency parsing and semantic role labelling. Although both tasks are performed together, in the evaluation task the accuracy results of dependency parsing are provided isolated from the semantic role labelling task. The evaluation data used corresponds to 1,725 sentences of the AnCora Corpus.

Finally, another work in statistical parsing in Spanish (Ballesteros et al., 2014) since it presents a complete different perspective of syntactic dependencies. It generating deep syntactic dependency structures from 3,036 sentences AnCora-UPF corpus (Mille et al., 2013). This parser has been evaluated with 258 sentences of the same corpus.

Grammar	LAS	UAS	LAS2
MaxEnt$_{Tibidabo}$	95.40	96.80	97.60
Malt$_{Tibidabo}$	92.00	95.00	94.50
Merlo	87.64	— — —	— — —
Bohnet	87.19	— — —	— — —
Che	87.33	— — —	— — —
Chen	86.29	— — —	— — —
FDG	**84.26**	**89.57**	**86.77**
Lluís	83.09	— — —	— — —
Zhang	82.69	— — —	— — —
Brown	82.46	— — —	— — —
Asahara	81.74	— — —	— — —
Li	77.21	— — —	— — —
Ren	76.11	— — —	— — —
Vallejo	73.07	— — —	— — —
Dai	71.64	— — —	— — —
Ballesteros	68.31	77.31	80.47
Zeman	65.98	— — —	— — —

TABLE 8.14: Comparison of Spanish FDG with Spanish statistical parsers

Parser	LAS	UAS	LAS2
Merlo	87.86	— — —	— — —
Che	86.56	— — —	— — —
Bohnet	86.35	— — —	— — —
Chen	85.88	— — —	— — —
Zhang	82.67	— — —	— — —
Brown	82.61	— — —	— — —
FDG	**80.15**	**88.24**	**86.77**
Asahara	79.48	— — —	— — —
Dai	77.85	— — —	— — —
Ren	77.84	— — —	— — —
Vallejo	77.33	— — —	— — —
Li	75.68	— — —	— — —
Lluís	64.21	— — —	— — —
Zeman	67.68	— — —	— — —

TABLE 8.15: Comparison of Catalan FDG with Catalan statistical parsers

The observation of the Spanish results (table 8.14) shows that the MaxEnt$_{Tibidabo}$ parser is on the top position of the ranking followed by the Malt$_{Tibidabo}$. FDG is in the seventh position from the top parser and ten statistical dependency parsers score below FDG. FDG is 11.14 points far from the best ranked parser and 18.28 points separate FDG and the lowest-ranked parser (Zeman).

Concerning Catalan, FDG is the only rule-based grammar evaluated. Although it has been only evaluated with the qualitative test sentences of ParTes, the results are compared to the statistical parsers of CoNLL 2009 (evaluated with sentences of AnCora Corpus) in order to measure the status of FDG in the state of the art. Likewise in Spanish parsers, FDG in Catalan is in the seventh position of a total of fourteen proposals. It scores 7.71 points less than the top-ranked parser of Merlo and 12.47 points higher than the lowest-ranked parser (Zeman).

Recapitulation

The argument recognition was pointed to as one of the important ambiguity issues which parsing needs to deal with in order to increase the accuracy of the parse trees (§3.3). Concerning this limitation, the FDGs in Spanish and in Catalan are not an exception. Initially they had limitations in the recognition of some specific arguments like the prepositional argument and the predicative (§6.4).

As a result of this situation, a new subcategorization lexicon, CompLex-VS, has been designed and which replaced the initial classes of FDGs (§8.1.1). While initial lexicon was classified in coarse-grained classes, CompLex-VS uses a different generalization of frames which have been acquired from the SenSem Corpus (§8.1.2). The new frames have been integrated into the labelling rules of the FDGs (§8.1.3). Finally, a set of experiments has been carried out to test how the subcategorization information improves the performance of these grammars (§8.2).

The results show that subcategorization frames ensure a high degree of accuracy performance. In most cases, the initial coarse-grained frames integrated in the grammars and the new CompLex-VS frames show an improvement. However, the increment is more evident in some arguments (prepositional object and the predicative) than others, like the complement in attributive verbs. These results indicate that some arguments necessarily need subcategorization information to be disambiguated, while others can be disambiguated just with the rules without any information added or with simple lexical information.

Furthermore, the new frames of CompLex-VS provide better results than the initial ones in the majority of cases. Therefore, more fine-grained frames (CompLex-SynF) contribute to raise the accuracy of the grammars. Despite this evidence, fine-grained classes do not necessarily mean improvement of the parser's performance. The most fine-grained lexicon (CompLex-SynF+Cat), which combines syntactic function and grammatical category information, neither improves nor worsens the results of the FDGs.

In the last part of this chapter, the results of the Spanish and Catalan grammars with highest accuracy scores have been compared to the results of other grammars available which have been evaluated. Furthermore, the performance of FDGs have been contrasted to statistical parsers performance. FDG is the grammar with the highest accuracy scores among the grammars analysed, and is ranked in the seventh position in the comparison with statistical parsers.

Chapter 9

Conclusions

Despite the recent advances in parsing, significant efforts are needed to improve current parsers' performance. For this reason, this thesis has focused on providing solutions for improving dependency grammars' performance. In order to achieve this main aim, this thesis has tried to resolve two of the major problematic linguistic phenomena faced by parsers namely that of high ambiguity. In particular, two highly controversial phenomena, prepositional phrase attachment and argument recognition, are studied and empirically experimented on in detail.

The strategy followed relied on the acquisition and integration of statistical and linguistic knowledge in dependency grammars for Spanish and Catalan. These grammars have been specifically developed for this research following the main principles of the dependency grammar formalism and establishing linguistically motivated criteria for these constructions controversial from the representational point of view.

In order to improve the performance of both grammars by knowledge integration, two experiments based on knowledge acquisition of prepositional phrase attachment and argument recognition were carried out. Disambiguation of prepositional phrase attachment was conducted by automatically learning word distributions using word embeddings from lexical representations of words involved in this kind of attachment. On the other hand, the task of recognizing arguments was handled by acquiring verbal subcategorization classes organized in a lexicon of subcategorization frames created specifically for this experiment.

The knowledge of both experiments was integrated and tested with the grammars by performing an empirical evaluation. This research is based on the idea that the assessment of parsing performance needs to offer a global perspective of the results in order to propose robust solutions for parsing accuracy. For this reason, this thesis designed an evaluation method in which quantitative and qualitative analyses are present and complementing each other.

In the following sections, the results of this research are analysed and confronted with the initial hypothesis stated of this study. Moreover, this research contributions are listed. In the last section of this chapter, new questions and lines of research are discussed according to the conclusions.

Hypothesis Revisited

In the introductory chapter, three different hypotheses and their respective sub-hypotheses were formulated as the basis of this research. This section aims at revisiting these hypotheses and checking whether they have been confirmed.

Hypothesis 1

Syntactic grammar rules provide an acceptable solution for the majority of constructions except for ambiguous syntactic phenomena

This hypothesis has been answered by evaluating the dependency grammars developed in this proposal (§6.4). The evaluation of the grammars working exclusively with rules without knowledge added showed an acceptable performance of these grammars.

However, both quantitative and qualitative analyses demonstrated that accuracy drops in specific constructions and dependency relations which are highly ambiguous. Among ambiguities in syntactic structures, prepositional phrase, adverbs, coordination and interjections are poorly performed. On the other hand, among the ambiguities in dependency relations, the prepositional phrase tends to appear as one of the structures with a higher error rate in dependency relation assignment (e.g. indirect object, prepositional object and adjunct).

These results leads to state that syntactic grammar rules on their own cover the basic syntactic phenomenon of the sentence, but they are limited in disambiguating phenomenon where more than one answer is possible.

Hypothesis 2

Statistical knowledge integrated in the grammar improves the accuracy of the grammar's performance

This statement can be broken down into the following specific points:

> Hypothesis 2.1
>
> **Unsupervised learning makes it possible to capture more consistently unpredicted data**
>
> A preliminary experiment about automatic learning of PP-attachment disambiguation using a classifier (§7.1) showed that, while supervised learning with a classifier enriched with semantic information ensures precise disambiguated PP-attachments, it is limited in making predictions for unseen data.
>
> For this reason, an unsupervised method based on word embeddings was tested (§7.2). The results of this experiment reveal that word embeddings are a robust method for making predictions of unseen data. It classifies all the results in a gradation of similarity instead of discriminating them as the supervised method does.
>
> Therefore, the experiments confirm that unsupervised learning based on word embeddings are capable of performing more consistent predictions than supervised learning methods.

Hypothesis 2.2

Language models learned by simple information such as lexical information provide a language representation of the PP-attachment which is not precise enough to disambiguate it. Therefore, enriched vectors with more complex information such as syntactic and semantic information ensure an improvement in the disambiguation task.

In the experiments for testing the word embeddings models learned, it has been observed that this method performs better than a supervised classifier. Despite this, models learned only with lexical information are achieve far from acceptable results. Furthermore, when PP-attachment disambiguated data is integrated in parsing, there is no significant change in the accuracy performance of the dependency grammars.

A qualitative analysis of the grammars' evaluation results has shown that new training data needs to be enlarged, the lexical value of the preposition needs to be added to the model and vectors have to be enriched with semantic and syntactic information in order to extend the power of word embeddings in precision and recall. In addition, new patterns of PP-attachment need to be learned in order that statistical knowledge learned and integrated in the grammars has a real impact on accuracy, especially since the pattern that has been studied covers a small sample of all the possible configurations of PP-attachments.

The hypothesis discussed here, then, is partially answered even as more experiments following the observations pointed out above are needed to prove that statistical knowledge can contribute to PP-attachment in parsing.

Hypothesis 3

Linguistic knowledge added to a rule-based grammar contributes to an improvement of the grammar's performance

To confirm this hypothesis two aspects formulated in the following sub-hypothesis were studied in the experiment found in chapter §8 about argument recognition:

Hypothesis 3.1

Subcategorization information has a great impact on highly ambiguous arguments

The integration of subcategorization frames in the labelling rules ensures an important increase in the quality of the grammars. However, subcategorization information has a different impact concerning the type of argument.

It is difficult for the rules to identify these arguments unless knowledge of subcategorization is integrated into the grammars. This is the reason why arguments realized as prepositional object and predicative are difficult to capture without subcategorization information and are significantly improved when subcategorization information is added in the labelling rules.

In addition, the results have indicated that the addition of linguistic knowledge by classes tends to restrict some labelling rules that were overgeneralizing and capturing a large amount of ambiguous arguments, but at the same time assigning the wrong label in a lot of cases. Consequently, the recall is slightly affected in order to capture better the arguments and needs to be improved.

Therefore, linguistic knowledge expressing the subcategorization frames of verbs contributes to the grammar performance accuracy positively.

Hypothesis 3.2

Fine-grained subcategorization frame classes are able to capture arguments more precisely than coarse-grained subcategorization classes

Subcategorization classes with several granularities influence the performance of the grammars differently. The results of the experiments reveals that a high level of accuracy is not always ensured when parsing with coarse-grained classes of subcategorization. On the other hand, more fine-grained subcategorization classes makes grammars' ability to perform more accurately.

However, more granularity among the classes does not automatically increase the accuracy of the grammars. The types of syntactic information by which information is generalized are also relevant. Subcategorization classes generalized by syntactic function and grammatical category are more granular than classes generalized only by syntactic information. However, they do not have an impact on the improvement in the performance accuracy.

For this reason, it is possible to state that subcategorization information organized in fine-grained classes really assists in recognizing more precisely dependency relations, but the type of syntactic information expressed in the classes is essential to ensure this improvement. Therefore, fine-grained grained subcategorization classes generalized by syntactic functions provide the knowledge that the grammars need in order to recognize the arguments properly.

Research Contributions

The present study has provided a **quantitative and qualitative analysis of the improvement of dependency grammars accuracy**. The starting point of this empirical study was the experimentation with two kinds of knowledge, statistical and linguistic, to identify whether the addition of knowledge to two dependency grammars for Spanish and Catalan contributes to their performance.

As a result of this study, an empirical demonstration of the **positive impact of knowledge integration in parsing** has been provided. On the one hand, this research has showed that statistical knowledge based on word embeddings and implemented to disambiguate PP-attachment may contribute to resolving this linguistic ambiguity. However, the experiments have determined that **language models learned from vectors representing lexical information are limited**. Vectors need to be enriched with more abstract linguistic information such as semantic and syntactic information. In addition, the scope PP-attachment patterns analysed need to be enlarged in order to observe significant changes in parsing accuracy.

On the other hand, this thesis demonstrates that **linguistic knowledge expressing verbal subcategorization classes is decisive** in order to recognize the arguments of the verb predicate and, consequently, to identify the correct dependency relation in the parse tree. More precisely, it has been shown that the addition of linguistic knowledge contributes a great deal to amount in the grammars' ability to correctly identify highly ambiguous arguments. In addition, the results of the study demonstrate that fine-grained classifications assist in the process of capturing the differences between arguments, so they also contribute to the dependency grammars' accuracy.

In order to validate these statements, **two dependency grammars for Spanish and Catalan**, named FreeLing Dependency Grammars (FDGs), were developed. The methodology and the research carried out here has made it possible to achieve a high accuracy in both grammars performance. In particular, the Spanish FDG contributes to the framework of highly accurate Spanish dependency grammars next to DepPattern (Gamallo, 2015). The Catalan FDG is the first contribution made to Catalan dependency grammars. With regard to the framework of statistical dependency parsing in the languages of this research, both FDGs are among state of the art dependency parsers and dependency grammars, although that they are performing less accurately than best Spanish and Catalan parsers (Ballesteros and Carreras, 2015; Hall et al., 2007).

Together with the FDGs, a **linguistic criteria proposal about the nature of syntactic heads** has been developed. Despite the fact that FDGs are grammars based on the syntactic dependencies formalism, they are strongly critical in linguistically controversial syntactic constructions. For this reason, after the theoretical revision of these constructions, FDGs have established their set of criteria that aims to be an eclectic proposal between syntactic and semantic perspectives. On the other hand, a list of linguistically motivated dependency relations labels has been created and **these labels have been mapped to Universal Dependencies syntactic labels**.

At the evaluation level, this research has shown that **a complete parsing evaluation task needs to be complemented with a qualitative analysis**, apart from a quantitative measurement of the results. For this reason, this thesis has designed a global evaluation method where quantitative and qualitative evaluation tasks complement each other. In particular, because qualitative evaluation is omitted the majority of times or is carried out without appropriate tools, this research has created a **new linguistic resource, ParTes, to test qualitatively the performance of parsed syntactic trees by means of a hierarchically structured test suite of syntactic structure and word order phenomena in Spanish and Catalan**.

Finally, the majority of these contributions have been published in journals and proceedings in the area of Natural Language Processing, Computational Linguistics and Applied Linguistics (§9). Furthermore, the research related to this thesis has been presented in conference communications and posters, workshops and seminars.

New Research Directions

The work presented in this thesis is just a small step towards the improvement of dependency grammars' accuracy. Actually, there is still a long way to go in order to improve their performance. Here we offer a summary of those lines we would like to study further.

Firstly, as we stated above, integration of statistical knowledge can improve the performance of FDGs, but more experiments are needed in order to prove that statistical knowledge learned by word embeddings contribute to PP-attachment disambiguation. As the results of the experiments shown (§7.2), future experiments need to handle the automatic learning task with word embeddings including the lexical value of the preposition and enriching vectors with semantic and syntactic information.

Both aspects can be performed using the same Wikicorpus (Reese et al., 2010) used for building the training data set of the PP-attachment disambiguation experiment (§7.2.2). On the one hand, Belinkov et al. (2014) propose two ways of generating vectors including the lexical value of the preposition of the pattern *VP* NP_1 *P* NP_2: to consider the whole pattern including the preposition a block, or to make the context flexible by calculating the weight of the preposition with NP_1 and *VP*. On the other hand, Top Concept Ontology features and Semantic File can be used by linking WordNet senses (Fellbaum, 1998) annotated in the corpus in order to disambiguate n-attachments and v-attachments. On the other hand, subcategorization information can be added in the training data in order to disambiguate v-attachments by mapping verbal WordNet synsets to SenSem verbs using the Spanish SemCor corpus (Castellón et al., 2003).

The results of the experiments have also shown that the size of the training data set is small for capturing basic relations involved in PP-attachment. Wikicorpus is a large corpus of 120 million words, but, since the pattern studied in this research is specific, training data needs to be enlarged with new data. However, larger free Spanish and Catalan corpora are not available. For this reason, online digitalized data can be downloaded as Gala and Lafourcade (2006) propose, and pre-processed linguistically with FreeLing library (Padró et al., 2010) to provide automatic annotations to extract the patterns of PP-attachment aimed to be studied.

Furthermore, the analysis of the evaluation results of the FDG with PP-attachment knowledge (§7.4.3) has shown that the PP-attachment pattern studied (*VP* NP_1 P NP$_2$) represents a small part of the whole phenomenon. In order that statistical knowledge about PP-attachment has a real impact on parsing performance, more patterns of PP-attachment need to be captured such as PP-attachment inside of another PP (*VP* PP_1 PP_2), attachment to the adjective (*VP AP PP* or *VP NP AP PP*) and attachment to the adverb (*VP Adv PP* or *VP NP Adv PP*).

Secondly, the integration of linguistic knowledge by means of subcategorization classes ensured an increase in the accuracy of the dependency grammars on the recognition of arguments. Despite this, it has been observed that FDGs lack some accuracy in the number of arguments captured. However, this limitation can be overcome and new cases can be captured by implementing the strategies used in this experiment, i.e., acquiring new verbal subcategorization frames from linguistic resources previously annotated. In particular, another subcategorization lexicon, AnCora-Verb (Aparicio et al., 2008), is available for both languages of this study, and can be used following the same methodology used in the SenSem Corpus extraction (Vázquez and Fernández-Montraveta, 2015). In addition, Spanish subcategorization CompLex-VS lexicon can be extended with subcategorization frames of the TRL Spanish V-SUBCAT lexicon (Padró et al., 2011).

Thirdly, coordination has been shown to be a highly ambiguous structure in the evaluation of FDGs (§6.4). Despite this, this thesis has not focused on resolving the attachment ambiguities of coordinating constructions. Since this construction is frequent in language, a better treatment of the rules that handle coordinated structures needs to be provided. As a starting point, a deeper error analysis focused on the most frequent coordination configurations has to be performed in order to conduct experiments about automatic learning of coordinated structures (Kübler et al., 2009; Maier and Kübler, 2013).

Fourthly, as a consequence of the previous observation, ParTes necessarily needs to include a repertoire of coordinating constructions in order to perform a robust analysis of errors of FDGs. This repertoire has to be populated with the basic coordination configurations organized from simplest structures (i.e. lexical coordination) to the most complex (i.e. clausal and sentence coordination). In order to extend the ParTes systematically, manually annotated treebanks such as AnCora (Taulé et al., 2008) can be used to extract the structures automatically.

On the other hand, the current verion of ParTes only includes grammatical test cases, which are appropriate to evaluate the precision of the parsing tool that is being evaluated. Despite this, the architecture of ParTes is not prepared for testing overgeneralization in rules. In order to deal with overgeneration, the test suite needs to include ungrammatical test sentences of the linguistic phenomena handled in the resource (Lehmann et al., 1996; Oepen and Flickinger, 1998). Besides assessment of rules of overgeneralization, ungrammatical test sentences are an essential feature to evaluate qualitatively non-deterministic parsers, which generate several possible parse trees for a sentence.

Fifthly, the evaluation method designed in this proposal is intrinsic. In other words, the performance of FDGs is evaluated in isolation. A way to enrich the evaluation task can be to carry out an extrinsic evaluation, i.e., an evaluation of the FDGs embedded in another NLP application. There are two open-source tools where FDGs developed in this research were implemented successfully. On the one hand, VERTa is a qualitative evaluation metric for Machine Translation in English and Spanish that compares fluency or adequacy of texts by similarity metrics (Comelles et al., 2014). On the other hand, OpenTrad[1] is a Machine Translation platform that uses both Spanish and Catalan FDGs to build translations from a source language to a target language.

On a different note, we would also like to investigate how flexible the current version of FDGs are in different text domains. This issue has been studied from the point of view of statistical parsing (Sekine, 1997; Gildea, 2001; McClosky et al., 2010). Although, rule-based parsing and statistical parsing are based on different strategies, the domain adaptation of statistical Spanish parsers (Nivre, 2003; Ballesteros and Carreras, 2015) and Catalan parsers (Nivre, 2003; Carreras,

[1] `http://www.opentrad.com/`

2007) can be contrasted to the accuracy of FDGs evaluated across different domains.

Bibliography

Aduriz, I., Aranzabe, M., Arriola, J., Diaz-De-Ilarraza, A., Gojenola, K., and Oronoz, M. (2004). A Cascaded Syntactic Analyser for Basque. In *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, pages 124–135. Springer Verlag.

Agerri, R., Bermudez, J., and Rigau, G. (2014). IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3823–3828.

Agirre, E., Baldwin, T., and Martinez, D. (2008). Improving Parsing and PP Attachment Performance with Sense Information. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 317–325.

Agirre, E., Màrquez, L., and Wicentowsky, R. (2007). SemEval 2007. In *Proceedings of the Fourth International Workshop on Semantic Evaluations*.

Agirre, E. and Soroa, A. (2009). Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41.

Aguilar, N., Alonso, L., Lloberes, M., and Castellón, I. (2011). Resolving prepositional phrase attachment ambiguities in Spanish with a classifier. In *Proceedings of the workshop on Natural Language Processing and Web-based Technologies*.

Ait-Mokhtar, S., Chanod, J., and Roux, C. (2001). A Multi-Input Dependency Parser. In *Proceedings of the Seventh International Workshop on Parsing Technologies (IWPT-2001)*.

Ajdukiewicz, K. (1935). Die syntaktische konnexität. In *Polish Logic 1920–1939*, pages 207–231. Oxford University Press.

Alarcos, E. (1994). *Gramática de la lengua española*. Espasa Calpe, Real Academia Española.

Aldezabal, I., Aranzabe, M., Atutxa, A., Gojenola, K., and Sarasola, K. (2003). Patrixa: A unification-based parser for Basque and its application to the automatic analysis of verbs. In *Inquiries into the lexicon-syntax relations in Basque*. Euskal Herriko Unibertsitatea.

Alonso, L., Castellón, I., and Tincheva, N. (2007). Obtaining coarse-grained classes of subcategorization patterns for Spanish. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*.

Alsina, A., Badia, T., Boleda, G., Bott, S., Gil, A., Quixal, M., and Valentín, O. (2002). CATCG: Un sistema de análisis morfosintáctico para el catalán. *Procesamiento del Lenguaje Natural*, 29.

Altmann, G. and Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, 30(3):191–238.

Álvez, J., Atserias, J., Carrera, J., Climent, S., Laparra, E., Oliver, A., and Rigau, G. (2008). Complete and Consistent Annotation of WordNet using the Top Concept Ontology. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 1529–1534.

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

Arias, B., Bel, N., Fisas, B., Lorente, M., Marimon, M., Morell, C., Vázquez, S., and Vivaldi, J. (2014). The IULA Spanish LSP Treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

Atserias, J., Comelles, E., and Mayor, A. (2005). TXALA un analizador libre de dependencias para el castellano. *Procesamiento del Lenguaje Natural*, 35.

Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B., and Vossen, P. (2004). The MEANING Multilingual Central Repository. In *In Proceedings of the Second International WordNet Conference*, pages 80–210.

Attardi, G. (2006). Experiments with a Multilanguage Non-projective Dependency Parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 166–170.

Badia i Margarit, A. (1962). *Gramàtica catalana*. Gredos.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc.

Baker, C., Fillmore, C., and Lowe, J. (1998). The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 86–90.

Ballesteros, M., Bohnet, B., Mille, S., and Wanner, L. (2014). Deep-syntactic parsing. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1402–1413.

Ballesteros, M. and Carreras, X. (2015). Transition-based Spinal Parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 289–299.

Bar-Hillel, Y. (1953). A quasi-arithmetical notation for syntactic description. *Language*, 29:47–58.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 238–247.

Belinkov, Y., Lei, T., Barzilay, R., and Globerson, A. (2014). Exploring Compositional Architectures and Word Vector Representations for Prepositional Phrase Attachment. *Transactions of the Association for Computational Linguistics*, 2:561–572.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3:1137–1155.

Bengoetxea, K. and Gojenola, K. (2010). Application of Different Techniques to Dependency Parsing of Basque. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 31–39.

Berwick, R., Abney, S., and Tenny, C., editors (1991). *Principle-Based Parsing: Computation and Psycholinguistics*. Kluwer Academic Publishers.

Bick, E. (2006). A Constraint Grammar-Based Parser for Spanish. In *Proceedings of TIL 2006 - 4th Workshop on Information and Human Language Technology*.

Bies, A., Mott, J., Warner, C., and Kulick, S. (2012). English web treebank ldc2012t13.

Bikel, D. and Chiang, D. (2000). Two Statistical Parsing Models Applied to the Chinese Treebank. In *In Proceedings of the Second Chinese Language Processing Workshop*, pages 1–6.

Björkelund, A. and Nivre, J. (2015). Non-Deterministic Oracles for Unrestricted Non-Projective Transition-Based Dependency Parsing. In *Proceedings of the 14th International Conference on Parsing Technologies*, pages 76–86.

Boguraev, B. and Briscoe, T. (1987). Large Lexicons for Natural Language Processing: Utilising the Grammar Coding System of LDOCE. *Computational Linguistics*, 13(3–4):203–218.

Bohnet, B. and Nivre, J. (2012). A Transition-based System for Joint Part-of-speech Tagging and Labeled Non-projective Dependency Parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1455–1465.

Bojar, B. (2004). Problems of Inducing Large Coverage Constraint-Based Dependency Grammar for Czech. In *Proceedings of International Workshop on Constraint Solving and Language Processing*, pages 29–42.

Bonet, S. (2002). Les subordinades substantives. In *Gramàtica del Català Contemporani*, pages 2321–2387. Empúries.

Bonet, S. and Solà, J. (1986). *Sintaxi generativa catalana*. Bilioteca Universitària, Enciclopèdia Catalana.

Bosque, I. and Demonte, V. (1999). *Gramática Descriptiva de la Lengua Española*. Espasa Calpe.

Brent, M. (1993). From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax. *Computational Linguistics*, 19(2).

Bresnan, J. (1982). *The mental representation of grammatical relations*. MIT Press.

Brill, E. and Resnik, P. (1994). A Rule-based Approach to Prepositional Phrase Attachment Disambiguation. In *Proceedings of the 15th Conference on Computational Linguistics*, pages 1198–1204.

Briscoe, T. (1987). Deterministic Parsing And Unbounded Dependencies. In *3rd Conference of the European Chapter of the Association for Computational Linguistics*, pages 211–217.

Briscoe, T. and Carroll, J. (1997). Automatic extraction of subcategorization from corpora. In *Fifth Conference on Applied Natural Language Processing*, pages 356–363.

Brucart, J. (1997). La estructura del sintagma nominal: las oraciones de relativo. In *Gramática descriptiva de la lengua española*. Espasa.

Buchholz, S. and Marsi, E. (2006). CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*.

Büttcher, S., Clarke, C., and Cormack, G. (2010). *Information Retrieval: Implementing and Evaluating Search Engines*. The MIT Press.

By, T. (2004). English dependency grammar. In *Proceedings of the ACL Workshop on Recent Advances in Dependency Grammar*, pages 72–77.

Calvo, H. and Gelbukh, A. (2006). DILUCT: An Open-Source Spanish Dependency Parsers based on Rules, Heuristics, and Selectional Preferences. In Kop, C., Fliedl, G., Mayr, H., and Métais, E., editors, *Natural Language Processing and Information Systems, 11th International Conference on Applications of Natural Language to Information Systems (NLDB'06)*, volume 3999 of *Lecture Notes in Computer Science*, pages 164–175. Springer.

Candito, M., Crabbé, B., and Denis, P. (2010). Statistical French Dependency Parsing: Treebank Conversion and First Results. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 1840–1847.

Carrera, J., Castellón, I., Lloberes, M., Padró, L., and Tinkova, N. (2008). Dependency grammars in FreeLing. *Procesamiento del Lenguaje Natural*, 41:21–28.

Carreras, X. (2007). Experiments with a Higher-Order Projective Dependency Parser. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 957–961.

Carreras, X., Collins, M., and Koo, T. (2008). TAG, Dynamic Programming, and the Perceptron for Efficient, Feature-rich Parsing. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 9–16.

Carroll, J. (2003). Parsing. In Mitkov, R., editor, *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Carroll, J., Minnen, G., and Briscoe, T. (1998). Can Subcategorisation Probabilities Help a Statistical Parser? In *Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora*.

Castellón, I., Climent, S., Coll-Florit, M., Lloberes, M., and Rigau, G. (2003). Semantic Hand Tagging of the SenSem Corpus Using Spanish WordNet Senses. In *Proceedings of the 6th International Global Wordnet Conference*, pages 72–78.

Charniak, E. and Johnson, M. (2005). Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 173–180.

Chomsky, N. (1957). *Syntactic Structures*. Mouton de Gruyter.

Chomsky, N. (1959). On Certain Formal Properties of Grammars. *Information and Control*, 2(2):137–167.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press.

Chomsky, N. (1981). *Lectures on Government and Binding: The Pisa Lectures*. Mouton de Gruyter.

Chomsky, N. (1995). *The Minimalist Program*. MIT Press.

Choukri, K. and Nilsson, M. (1998). The European Language Resources Association. In *Proceedings of the First International Conference on Language Resources and Evaluation*, pages 153–158.

Church, K. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Civit, M. (2003). Criterios de etiquetación y desambiguación morfosintáctica de corpus en español. In *Colección de Monografías de la Sociedad Española para el Procesamiento del Lenguaje Natural: 8*. Sociedad Española para el Procesamiento del Lenguaje Natural.

Civit, M., Martí, M., and Bufí, N. (2006). Cat3LB and Cast3LB: from Constituents to dependencies. In *Advances in Natural Language Processing (LNAI, 4139)*. Springer Verlag.

Clark, S. (2015). Vector Space Models of Lexical Meaning. In Lappin, S. and Fox, C., editors, *Handbook of Contemporary Semantic Theory*, pages 493–522. Wiley-Blackwell.

Collins, M. (1999). *Head-Driven Statistical Models for Natural Language Parsing*. University of Pennsylvania.

Collins, M. (2000). Discriminative Reranking for Natural Language Parsing. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000)*, pages 175–182.

Collins, M. and Brooks, J. (1995). Prepositional Phrase Attachment through a Backed-Off Model. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 27–38.

Collins, M., Ramshaw, L., Hajič, J., and Tillmann, C. (1999). A Statistical Parser for Czech. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 505–512.

Collobert, R. and Weston, J. (2008). A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 160–167.

Comelles, E., Atserias, J., Arranz, V., Castellón, I., and Sesé, J. (2014). VERTa: Facing a Multilingual Experience of a Linguistically-based MT Evaluation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 2701–2707.

Consortium, B. N. C. (2007). *The British National Corpus*. Oxford University Computing Services.

Copestake, A. and Flickinger, D. (2000). An open source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings or Second Conference on Language Resources and Evaluation*, pages 591–600.

Costa, F. and Branco, A. (2010). LXGram: A Deep Linguistic Processing Grammar for Portuguese. In *Computational Processing of the Portuguese Language: 9th International Conference, PROPOR 2010*, pages 86–89. Springer.

Cowie, J. and Lehnert, W. (1996). Information Extraction. *Communication of the ACM*, 39(1):80–91.

Croft, W. (2008). Aspectual and causal structure in event representations. In *Routes to language development. Studies in honor of Melissa Bowerman*, pages 139–166. Erlbawm.

Davies, M. (2011). *Google Books (American English) Corpus (155 billion words, 1810-2009)*. Brigham Young University.

De Marneffe, M., Maccartney, B., and Manning, C. (2006). Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the Fifth Conference on International Language Resources and Evaluation*, pages 449–454.

De Marneffe, M. and Manning, C. (2012). Stanford typed dependencies manual. Technical report, Technical report, Stanford University.

Debusmann, R., Duchier, D., and Kruijff, G. (2004). Extensible Dependency Grammar: A New Methodology. In *Proceedings of the COLING 2004 Workshop on Recent Advances in Dependency Grammar*, pages 78–85.

Delbecque, N. and Lamiroy, B. (1999). La subordinación sustantiva: las subordinadas enunciativas en los complementos verbales. In *Gramática descriptiva de la lengua española*, pages 1965–2081. Espasa.

Dik, S. (1968). *Coordination. Its implications for the theory of general linguistics*. North-Holland Publishing Company.

Dowty, D. (1979). *Word Meaning and Montague Grammar*. Reidel.

EAGLES (1994). Draft Interim Report EAGLES. Technical report, Expert Advisory Group on Language Engineering Standards.

Edmonds, P. and Kilgarriff, A. (2002). Introduction to the special issue on evaluating word sense disambiguation systems. *Journal of Natural Language Engineering*, 8(4):279–291.

Eisner, J. (1996). Three New Probabilistic Models for Dependency Parsing: An Exploration. In *Proceedings of the 16th Conference on Computational Linguistics*, pages 340–345.

Erk, K. and Padó, S. (2008). A Structured Vector Space Model for Word Meaning in Context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 897–906.

Evert, S. (2005). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

Fabra, P. (1918). *Gramàtica catalana*. Institut d'Estudis Catalans.

Fellbaum, C. (1998). *WordNet. An electronic database*. MIT Press.

Fernández, A. and Vàzquez, G. (2014). The SenSem Corpus: an annotated corpus for Spanish and Catalan with information about aspectuality, modality, polarity and factuality. *Corpus Linguistics and Linguistic Theory*, 10(2).

Fernández, A., Vázquez, G., Saint-Dizier, P., Benamara, F., and Kamel, M. (2002). The VOLEM Project: A Framework for the Construction of Advanced Multilingual Lexicons. In *Proceedings of the Language Engineering Conference*.

Ferrández, A. and Moreno, L. (2000). Slot Unification Grammar and Anaphora Resolution. In Nicolov, N. and Mitkov, R., editors, *Recent Advances in Natural Language Processing II. Selected papers from RANLP 1997*. John Benjamins Publishing Co.

Flickinger, D., Nerbonne, J., and Sag, I. (1987). Toward Evaluation of NLP Systems. Technical report, Hewlett Packard Laboratories. Distributed at the 24th Annual Meeting of the Association for Computational Linguistics (ACL).

Foth, K. and Menzel, W. (2006). The Benefit of Stochastic PP Attachment to a Rule-based Parser. In *Proceedings of the 21st Internationa Conference. on Computational Linguistics*, pages 223–230.

Freitag, D. (2000). Machine Learning for Information Extraction in Informal Domains. *Machine Learning*, 39(2–3):169–202.

Gaifman, H. (1965). Dependency systems and phrase-structure systems. *Information and Control*, 8:304–337.

Gala, N. and Lafourcade, M. (2006). PP Attachment Ambiguity Resolution with Corpus-Based Pattern Distributions and Lexical Signatures. *ECTI-CIT Transactions on Computer and Information Technology*, 2(2):116–120.

Gamallo, P. (2015). Dependency Parsing with Compression Rules. In *Proceedings of the 14th International Conference on Parsing Technologies*, pages 107–117.

García, L. (1999). Relaciones paratácticas e hipotácticas. In *Gramática descriptiva de la lengua española*, pages 3507–3547. Espasa.

Gazdar, G., Klein, E., Pullum, J., and Sag, I. (1985). *Generalized phrase structure grammar*. Blackwell Publishing and Harvard University Press.

Gelbukh, A., Torres, S., and Calvo, H. (2005). Transforming a Constituency Treebank into a Dependency Treebank. *Procesamiento de Lenguaje Natural*, 35:145–152.

Gildea, D. (2001). Corpus Variation and Parser Performance. In *2001 Conference on Empirical Methods in Natural Language Processing*, pages 167–202.

Gildea, D. and Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Gillon, B. (1990). Ambiguity, generality, and indeterminacy: Tests and definitions. *Synthese*, 85(3):391–416.

Green, S. and Manning, C. D. (2010). Better Arabic Parsing: Baselines, Evaluations, and Analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 394–402.

Grishman, R., Macleod, C., and Meyers, A. (1994). Comlex Syntax: Building a Computational Lexicon. In *Proceedings of the 15th Conference on Computational Linguistics*, pages 268–272.

Haegeman, L. (1991). *Introduction to Government and Binding Theory*. Blackwell.

Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M., Màrquez, L., Meyers, A.and Nivre, J., Padó, S., Štěpánek, J., Straňák, P., Surdeanu, M., Xue, N., and Zhang, Y. (2009). The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 1–18.

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Bojar, O., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., , Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z., and Žabokrtský,

Z. (2012). Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, pages 3153–3160.

Hall, J., Nilsson, J., Nivre, J., Eryigit, G., Megyesi, B., Nilsson, M., and Saers, M. (2007). Single Malt or Blended? A Study in Multilingual Parser Optimization. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 933–939.

Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

Hays, D. (1964). Dependency theory: A formalism and some observations. *Language*, 40:511–525.

Hellwig, P. (1986). Dependency unification grammar. In *Proceedings of the 11th International Conference on Computational Linguistics*, pages 195–198.

Henestroza, E. and Candito, M. (2011). Parse Correction with Specialized Models for Difficult Attachment Types. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1222–1233.

Hernanz, M. (2002). L'oració. In Solà, J., Lloret, M., Mascaró, J., and Pérez Saldanya, M., editors, *Gramàtica del Català Contemporani*. Empúries.

Hernanz, M. and Brucart, J. (1987). *La sintaxis. Principios teóricos. La oración simple*. Editorial Crítica.

Hindle, D. and Rooth, M. (1993). Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120.

Hudson, R. (1984). *Word Grammar*. Blackwell Publishing.

Hutchins, W. and Somers, H. (1992). *An introduction to machine translation*. Academic Press.

Joanis, E., Stevenson, S., and James, D. (2008). A general feature space for automatic verb classification. *Natural Language Engineering*, 14:337–367.

Jurafsky, D. and Martin, J. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Prentice-Hall.

Kaplan, R. and Bresnan, J. (1995). Lexical-Functional Grammar: A Formal System for Grammatical Representation. In *Formal Issues in Lexical-Functional Grammar*, Center for the Study of Language and Information Series. The University of Chicago Press.

Karlsson, F. (1990). Constraint Grammar As a Framework for Parsing Running Text. In *Proceedings of the 13th Conference on Computational Linguistics*, pages 168–173.

Kay, M. (1983). Unification Grammar. Technical report, Xerox Palo Alto Research Cente.

King, M., Maegaard, B., Schütz, J., desTombes, L., Bech, A., Neville, A., Arppe, A., Balkan, L., Brace, C., Bunt, H., Carlson, L., Douglas, S., Höge, M., Krauwer, S., Manzi, S., Mazzi, C., Sieleman, A., and Steenbakkers, R. (1996). EAGLES Evaluation of Natural Language Processing Systems. Technical report, Center for Sprogteknologi, Cophenhaguen.

Kipper, K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis, University of Pennsylvania.

Klein, D. (2005). *The Unsupervised Learning of Natural Language Structure*. PhD thesis, Stanford University.

Klein, D. and Manning, C. (2003). Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press.

Kolz, B., Badia, T., and Saurí, R. (2014). From constituents to syntax-oriented dependencies. *Procesamiento del Lenguaje Natural*, 52:53–60.

Koo, T., Globerson, A., Carreras, X., and Collins, M. (2007). Structured Prediction Models via the Matrix-Tree Theorem. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 141–150.

Kübler, S., Maier, W., Hinrichs, E., and Klett, E. (2009). Parsing coordinations. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 406–414.

Kudo, T. and Matsumoto, Y. (2000). Japanese Dependency Structure Analysis Based on Support Vector Machines. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing*, pages 18–25.

Kummerfeld, J., Hall, D., Curran, J., and Klein, D. (2012). Parser Showdown at the Wall Street Corral: An Empirical Investigation of Error Types in Parser Output. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1048–1059.

Lehmann, S., Oepen, S., Regnier-Prost, S., Netter, K., Lux, V., Klein, J., Falkedal, K., Fouvy, F., Estival, D., Dauphin, E., Compagnion, H., Baur, J., Balkan, L., and Arnold, D. (1996). TSNLP - Test Suites for Natural Language Processing. In *Proceedings of the 16th Conference on Computational Linguistics*, volume 2.

Lei, T., Xin, Y., Zhang, Y., Barzilay, R., and Jaakkola, T. (2014). Low-Rank Tensors for Scoring Dependency Structures. In *Proceedgins of the 52nd Annual Meeting of the Association for Computational Linguistics*.

Lenci, A. (2014). Carving verb classes from corpora. *Word Classes: Nature, typology and representations*, 332:17–36.

Levin, B. (1991). Building a Lexicon: The Contribution of Linguistics. *International Journal of Lexicography. Special issue. Building a Lexicon*, 4(3).

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press.

Levy, R. and Manning, C. (2003). Is It Harder to Parse Chinese, or the Chinese Treebank? In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 439–446.

Li, J. and Brew, C. (2008). Which Are the Best Features for Automatic Verb Classification. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 434–442.

Lin, D. (1998a). Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 768–774.

Lin, D. (1998b). Dependency-Based Evaluation of MINIPAR. In *Workshop on the Evaluation of Parsing Systems, First International Conference on Language Resources and Evaluation*.

Lin, D. and Pantel, P. (2001). DIRT – Discovery of Inference Rules from Text. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 323–328.

Lloberes, M. and Castellón, I. (2011). Consideraciones sobre la naturaleza de los núcleos sintácticos. Hacia una representación sintáctica de dependencias. *Anuari de Filologia. Estudis de Lingüística*, 1.

Lloberes, M., Castellón, I., and Padró, L. (2010). Spanish FreeLing Dependency Grammar. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*.

Lloberes, M., Castellón, I., and Padró, L. (2015a). Enhancing FreeLing Rule-Based Dependency Grammars with Subcategorization Frames'. In *Proceedings of the 3rd Conference on Dependency Linguistics*, pages 201–210.

Lloberes, M., Castellón, I., and Padró, L. (2015b). Suitability of ParTes Test Suite for Parsing Evaluation. In *Proceedings of the 14th International Conference on Parsing Technologies*, pages 61–65.

Lloberes, M., Castellón, I., Padró, L., and Gonzàlez, E. (2014). ParTes. Test Suite for Parsing Evaluation. *Procesamiento del Lenguaje Natural*, 53.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical cooccurrence. *Behavior Research Methods, Instrumentation, and Computer*, 28:203–208.

Luque, F., Quattoni, A., Balle, B., and Carreras, X. (2012). Spectral Learning for Non-Deterministic Dependency Parsing. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 409–419.

Magerman, D. (1994). *Natural Language Parsing as Statistical Pattern Recognition*. PhD thesis, Stanford University.

Maier, W. and Kübler, S. (2013). Are All Commas Equal? Detecting Coordination in the Penn Treebank. In *The Twelfth Workshop on Treebanks and Linguistic Theories (TLT12)*, pages 121–133.

Manning, C. (1993). Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*.

Manning, C., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Marcus, M., Marcinkiewicz, M., and Santorini, B. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Marimon, M. and Bel, N. (2015). Dependency structure annotation in the IULA Spanish LSP Treebank. *Language Resources and Evaluation*, 49(2):433–454.

Marimon, M., Bel, N., and Padró, L. (2014). Automatic Selection of HPSG-parsed Sentences for Treebank Construction. *Computational Linguistics*, 40(3).

Marimon, M., Bel, N., and Seghezzi, N. (2007). Test-suite Construction for a Spanish Grammar. In *Proceedings of the GEAF 2007 Workshop*.

Marimón, M. (2010). The Spanish Resource Grammar. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*.

Màrquez, L., Carreras, X., Litkowski, K., and Stevenson, S. (2008). Semantic role labeling: an introduction to the special issue. *Computational linguistics*, 34(2):145–159.

Martins, A., Almeida, M., and Smith, N. (2013). Turning on the Turbo: Fast Third-Order Non-Projective Turbo Parsers. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 617–622.

Marton, Y., Habash, N., and Rambow, O. (2013). Dependency Parsing of Modern Standard Arabic with Lexical and Inflectional Features. *Computational Linguistics*, 39(1):161–194.

Maruyama, H. (1990). Structural Disambiguation with Constraint Propagation. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 31–38.

McClosky, D., Charniak, E., and Johnson, M. (2006). Effective Self-training for Parsing. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 152–159.

McClosky, D., Charniak, E., and Johnson, M. (2010). Automatic domain adaptation for parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 28–36.

McDonald, R. and Nivre, J. (2011). Analyzing and Integrating Dependency Parsers. *Computational Linguistics*, 37(1):197–230.

Mcdonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu, N., and Lee, J. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 92–97.

McDonald, R., Pereira, F., Ribarov, K., and Hajič, J. (2005). Non-projective Dependency Parsing Using Spanning Tree Algorithms. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 523–530.

McEnery, T. and Wilson, A. (1996). *Corpus Linguistics*. Edinburgh University Press, Edinburgh.

McLauchlan, M. (2001). Maximum Entropy Models and Prepositional Phrase Ambiguity. Master's thesis, University of Edimburgh.

Mel'čuk, I. (1988). *Dependency Syntax: Theory and Practice*. State U. Press of NY.

Mel'čuk, I. (1996). Lexical functions: a tool for the description of lexical relations in a lexicon. In *Lexical Functions in Lexicography and Natural Language Processing*. John Benjamins Publishing Co.

Mel'čuk, I. (2003). Dependency in Natural Language. In *Dependency in Linguistic Description. Studies in Language Companion Series*, pages 1–10. John Benjamins Publishing Co.

Merlo, P. and Ferrer, E. E. (2006). The Notion of Argument in PP Attachment. *Computational Linguistics*, 32(3):341–378.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Mille, S., Burga, A., and Wanner, L. (2013). AncoraUPF: A Multi-Level Annotation of Spanish. In *In Proceedings of the Second International Conference on Dependency Linguistics*, pages 217–226.

Mitkov, R. (2003). *The Oxford Handbook of Computational Linguistics*. Oxford Handbooks in Linguistics, Oxford University Press.

Moreno Cabrera, J. (2000). *Curso universitario de lingüística general: Teoría de la gramática y sintaxis general*. Springer.

Müller, S. and Kasper, W. (2000). HPSG Analysis of German. In *Verbmobil: Foundations of Speech-to-Speech Translation, Artificial Intelligence*, pages 238–253. Springer.

Nivre, J. (2003). An Efficient Algorithm for Projective Dependency Parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies*, pages 149–160.

Nivre, J. (2004). Incrementality in Deterministic Dependency Parsing. In *Proceedings of the Workshop on Incremental Parsing: Bringing Engineering and Cognition Together*, pages 50–57.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., and Yuret, D. (2007). The CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*.

Nivre, J., Hall, J., Nilsson, J., Eryiğit, G., and Marinov, S. (2006). Labeled Pseudo-projective Dependency Parsing with Support Vector Machines. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*.

Ó Séaghdha, D. and Copestake, A. (2008). Semantic Classification with Distributional Kernels. In *Proceedings of the 22Nd International Conference on Computational Linguistics*, pages 649–656.

Oepen, S. and Flickinger, D. (1998). Towards Systematic Grammar Profiling: Test Suite Technology Ten Years after. *Journal of Computer Speech and Language*, 12:411–435.

Olteanu, M. and Moldovan, D. (2005). PP-attachment Disambiguation Using Large Context. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 273–280.

Padó, S. and Lapata, M. (2007). Dependency-Based Construction of Semantic Space Models. *Computational Linguistics*, 33(2):161–199.

Padró, M., Bel, N., and Necsulescu, S. (2011). Towards the automatic merging of lexical resources: Automatic mapping. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 296–301.

Padró, L., Reese, S., Lloberes, M., and Castellón, I. (2010). FreeLing 2.1: Five years of open-source language processing tools. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

Pantel, P. and Lin, D. (2000). An Unsupervised Approach to Prepositional Phrase Attachment Using Contextually Similar Words. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 101–108.

Paroubek, P., Chaudiron, S., and Hirschman, L. (2007). Principles of Evaluation in Natural Language Processing. *Traitement Automatique des Langues*, 48(1):7–31.

Pereira, F. and Warren, D. (1986). Definite clause grammars for language analysis. In *Readings in Natural Language Processing*, pages 101–124. Morgan Kaufmann Publishers Inc.

Petrov, D. and Klein, D. (2007). Improved Inference for Unlexicalized Parsing. In *Proceedings of the Conference On Human Language Technology and North American chapter of the Association for Computational Linguistics (HTL-NAACL 2007)*, pages 404–411.

Pollard, C. (1984). *Generalized phrase structure grammars, head grammars, and natural language*. PhD thesis, Stanford University.

Pollard, C. (1985). Lecture notes on head-driven phrase-structure grammar. Center for the Study of Language and Information, unpublished.

Popel, M., Mareček, D., Štěpánek, J., Zeman, D., and Žabokrtský, Z. (2013). Coordination Structures in Dependency Treebanks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 517–527.

Pustejovsky, J. (1991). The generative lexicon. *Comput. Linguist.*, 17(4):409–441.

Qian, X. and Liu, Y. (2012). Joint Chinese Word Segmentation, POS Tagging and Parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 501–511.

Rabin, M. and Scott, D. (1959). Finite Automata and Their Decision Problems. *IBM Journal of Research and Development*, 3(2):114–125.

Rafferty, A. and Manning, C. (2008). Parsing Three German Treebanks: Lexicalized and Unlexicalized Baselines. In *Proceedings of the Workshop on Parsing German*, pages 40–46.

Ratnaparkhi, A. (1998). Statistical Models for Unsupervised Prepositional Phrase Attachment. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 1079–1085.

Ratnaparkhi, A., Reynar, J., and Roukos, S. (1994). A Maximum Entropy Model for Prepositional Phrase Attachment. In *Proceedings of the Workshop on Human Language Technology*, pages 250–255.

Reese, S., Boleda, G., Cuadros, M., Padró, L., and Rigau, G. (2010). Wikicorpus: A Word-Sense Disambiguated Multilingual Wikipedia Corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 1418–1421.

Ribarov, K. (2004). *Automatic Building of a Dependency Tree - The Rule-Based Approach and Beyond*. PhD thesis, Univerzita Karlova.

Rimell, L., Clark, S., and Steedman, M. (2009). Unbounded Dependency Recovery for Parser Evaluation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*.

Schröder, I., Pop, H., Menzel, W., and Foth, K. (2001). Learning the constraints weights of a dependency grammar using genetic algorithms. In *Proceedings of the 13th International conference on Domain Decomposition Methods*, pages 243–247.

Schulte im Walde, S. (2006). Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics*, 32(2):159–194.

Sebastián, N., Martí, M., Carreiras, M., and Cuetos, F. (2000). *LEXESP: Léxico Informatizado del Español*. Publicacions de la Universitat de Barcelona.

Sekine, S. (1997). The Domain Dependence of Parsing. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 96–102.

Sennet, A. (2016). Ambiguity. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Center for the Study of Language and Information (CSLI), Stanford University, spring 2016 edition.

Serra i Prunyonosa, T. (2002). La coordinació. In Solà, J., Lloret, M., Mascaró, J., and Pérez-Saldanya, M., editors, *Gramàtica del Català Contemporani*. Empúries.

Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company.

Siegel, M. and Bender, E. (2002). Efficient Deep Processing of Japanese. In *Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization at the 19th International Conference on Computational Linguistics*.

Silva, J., Branco, A., Castro, S., and Reis, R. (2010). Out-of-the-box Robust Parsing of Portuguese. In *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language*, pages 75–85.

Silveira, N., Dozat, T., De Marneffe, M., Bowman, S., Connor, M., Bauer, J., and Manning, C. (2014). A Gold Standard Dependency Corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 2897–2904.

Sleator, D. and Temperley, D. (1991). Parsing English with a Link Grammar. In *Third International Workshop on Parsing Technologies*.

Socher, R., Bauer, J., Manning, C., and Ng, A. (2013). Parsing With Compositional Vector Grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 455–465.

Solà, J., Lloret, M., Mascaró, J., and Pérez-Saldanya, M. (2002). *Gramàtica del Català Contemporani*. Empúries.

Sparck Jones, K. and Galliers, J. (1996). *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer-Verlag.

Spitkovsky, V., Alshawi, H., and Jurafsky, D. (2010). From Baby Steps to Leapfrog: How "Less is More" in Unsupervised Dependency Parsing. In *North American Association for Computational Linguistics - Human Language Technologies (NAACL-HLT)*.

Stetina, J. and Nagao, M. (1997). Corpus Based PP Attachment Ambiguity Resolution with a Semantic Dictionary. In *Proceedings of the Fifht Workshop on Very Large Corpora*, pages 66–80.

Sun, L. and Korhonen, A. (2009). Improving Verb Clustering with Automatically Acquired Selectional Preferences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 638–647.

Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L., and Nivre, J. (2008). The CoNLL-2008 Shared Task on Joint Parsing of Syntactic and Semantic Dependencies. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 159–177.

Surdeanu, M. and Turmo, J. (2005). Semantic Role Labeling Using Complete Syntactic Analysis. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 221–224.

Tapanainen, P. and Järvinen, T. (1997). A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 64–71.

Tapanainen, P. and Järvinen, T. (1998). Towards an Implementable Dependency Grammar. In *Proceedings of the ACL Workshop on Processing of Dependency-Based Grammars*, pages 1–10.

Taulé, M., Martí, M., and Recasens, M. (2008). AnCora: Multi level annotated corpora for Catalan and Spanish. In *6th International Conference on Language Resources and Evaluation, Marrakesh*.

Tesnière, L. (1959). *Élements de syntaxe structurale*. Klincksieck.

Titov, I. and McDonald, R. (2008). Modeling Online Reviews with Multi-grain Topic Models. In *Proceedings of the 17th International Conference on World Wide Web*, pages 111–120.

Toutanova, K., Manning, C., Flickinger, D., and Oepen, S. (2005). Stochastic HPSG parse disambiguation using the Redwoods corpus. *Research on Language and Computation*.

Toutanova, K., Manning, C., Shieber, S., Flickinger, D., and Oepen, S. (2002). Parse Disambiguation for a Rich HPSG Grammar. In *Proceedings of the First Workshop on Treebanks and Linguistic Theories*, pages 253–263.

Turian, J., Ratinov, L., and Bengio, Y. (2010). Word Representations: A Simple and General Method for Semi-supervised Learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394.

Turney, P. (2002). Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 417–424.

Turney, P. (2008). The Latent Relation Mapping Engine: Algorithm and Experiments. *Journal of Artificial Intelligence Research*, 33(1):615–655.

Turney, P. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

Vázquez, G. and Fernández-Montraveta, A. (2015). Constructions at argument-structure level in the SenSem Corpora. *Language Resources and Evaluation*, 49(3):637–658.

Verkuyl, H. (1989). Aspectual Classes and Aspectual Composition. *Linguistics and Philosophy*, 12:39–94.

Villalba, X. (2002). La subordinació. In *Gramàtica del Català Contemporani*, pages 2247–2319. Empúries.

Šuster, S. (2012). Resolving PP-attachment ambiguity in French with distributional methods. Master's thesis, University of Groningen and University of Lorraine.

Weischedel, R., Hovy, E., Marcus, M., Palmer, M., Belvin, R., Pradan, S., Ramshaw, L., and Xue, N. (2011). OntoNotes: A Large Training Corpus for Enhanced Processing. In Olive, J., Christianson, C., and McCary, J., editors, *Handbook of Natural Language Processing and Machine Translation*. Springer.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann Publishers Inc.

Yamada, H. and Matsumoto, Y. (2003). Statistical Dependency Analysis With Support Vector Machines. In *Proceedings of 8th International Workshop on Parsing Technologies*, pages 195–206.

Yuret, D. (1998). *Discovery of Linguistic Relations Using Lexical Attraction*. PhD thesis, Massachussetts Institute of Technology.

Zeman, D. (2002). Can Subcategorization Help a Statistical Dependency Parser? In *19th International Conference on Computational Linguistics*.

Zeman, D. (2009). A Simple Generative Pipeline Approach to Dependency Parsing and Semantic Role Labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL): Shared Task*, pages 120–125.

Appendix

In this section, the research activity related to this thesis is shown chronologically.

Publications

1. Lloberes, M., I. Castellón, L. Padró (2010). 'EsTxala y CaTxala: Gramáticas de dependencias para el español y el catalán', Actas del IX Congreso Internacional de Lingüística General. Valladolid:, p. 1406-1421. Lingüística General. Universidad de Valladolid. ISBN: 987-84-693-6785-5

2. Lloberes, M., I. Castellón, L. Padró (2010). 'Spanish FreeLing Dependency Grammar'. Nicoletta Calzolari et al. (ed.), Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), p. 693–699. ISBN 2–9517408–6–7

3. Lloberes,M. i I. Castellón (2011). 'Consideraciones sobre la naturaleza de los núcleos sintácticos. Hacia una representación sintáctica de dependencias', Anuari de Filologia. Estudis de Lingüística, 1:1, p. 101-134. ISSN: 2014-1408

4. Lloberes M., I. Castellón, L. Padró, E. González (2014). 'ParTes. Test Suite for Parsing Evaluation', Procesamiento del Lenguaje Natural, 2014:53, p. 87-94

5. Lloberes M., I. Castellón, L. Padró (2015). 'Suitability of ParTes Test Suite for Parsing Evaluation', Proceedings of the 14th International Conference on Parsing Technologies, p. 61-65

6. Lloberes M., I. Castellón, L. Padró (2015). 'Enhancing FreeLing Rule-Based Dependency Grammars with Subcategorization Frames', Proceedings of the 3rd Conference on Dependency Linguistics, p. 201-210

Conferences

1. Talk. XXIV Congreso Anual de la Sociedad Española para el Procesamiento del Lenguaje Natural, Universidad Carlos III, Madrid (September 10th–12th 2008)

2. Talk. III Jornadas sobre modelos y técnicas para el acceso a la información multilingüe y multimodal en la web, Universidad Carlos III, Madrid (February 5th–6th 2009)

3. Poster Session. VII International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta (May 17th–23rd 2010)

4. Talk. IX Congreso Internacional de Lingüística General, Valladolid (June 21st–23rd 2010)

5. Talk. XXVI Congreso Internacional de la Associación de Jóvenes Lingüistas, Universidad de Salamanca, Salamaca (March 9–11 2011)

6. Talk. III Congreso Internacional de Lingüística de Corpus, Universitat Politècnica de València, València (7–9 April 2011)

7. Talk. XXX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural, Universitat de Girona (September 17th–19th 2014)

8. Talk. 14th International Conference on Parsing Technologies, Euskal Herriko Unibertsitatea (Jully 22nd–24th 2015)

9. Short Talk and Poster Session. 3rd International Conference on Dependency Linguistics, Uppsala University (August 24th–26th 2015)

ESTXALA Y CATXALA: GRAMÁTICAS DE DEPENDENCIAS PARA EL ESPAÑOL Y EL CATALÁN

Marina Lloberes, Irene Castellón
GRIAL. Departamento de Lingüística General
Universitat de Barcelona

Lluís Padró
TALP. Departamento de Lenguajes y sistemas informáticos
Universitat Politècnica de Catalunya

Resumen

Presentamos dos gramáticas de dependencias computacionales, una para el español y otra para el catalán, desarrolladas en el entorno Freeling y que se distribuyen bajo la licencia GPL. El objetivo de ambas gramáticas es proporcionar análisis profundos y robustos para un amplio repertorio de fenómenos lingüísticos del español y del catalán. Además, son una respuesta a la escasez de recursos que existen para este par de lenguas. El foco de nuestra investigación actualmente reside en la metodología de evaluación de la gramática con el fin de precisar la cualidad de los recursos que presentamos y de aportar una evaluación exhaustiva.

1. Introducción

EsTxala y CaTxala[1] son gramáticas basadas en el formalismo de dependencias (Tèsnière, 1959; Mel'čuk, 1988) para el español y el catalán, respectivamente. Ambas gramáticas han sido diseñadas para el análisis sintáctico automático y, en concreto, para la librería de herramientas de Procesamiento de Lenguaje Natural (PLN) FreeLing (Padró et al., 2010).[2]

Muchas aplicaciones de PLN (traducción automática, extracción de información, etiquetaje automático de roles semánticos, etc.) requieren de una cierta profundidad en los análisis sintácticos para poder obtener una buena representación semántica sobre la que aplicar procesos posteriores. En consecuencia, durante los últimos años, este ámbito del PLN ha hecho grandes avances.

[1] ExTxala y CaTxala se han desarrollado en el marco de los proyectos KNOW (Ministerio de Educación y Ciencia, TIN2006–1549–C03–02) y KNOW2 (Ministerio de Ciencia y Educación, TIN2009–14715–C04–03, TIN2009–14715–C04–04), y también han sido utilizadas en los proyectos OpenTrad y EuroOpenTrad (Ministerio de Industria, Turismo y Comercio, Programa PROFIT, FIT–350401–2006–5), dos proyectos que tienen como objetivo desarrollar traductores basados en la transferencia para las lenguas oficiales del Estado Español (español, catalán, gallego y vasco) y para el Inglés.

[2] http://www.lsi.upc.edu/PNL ~/Freeling/

En lenguas como el inglés, se han desarrollado diversas gramáticas con este objetivo, como MaltParser (Nivre, 2006), Minipar (Lin, 1998), Connexor (Tapanainen y Järvinen, 1998) o Link Grammar (Sleator y Temperley, 1991). No obstante, en otras lenguas, esta efervescencia parece ser menor. En español, existen algunas gramáticas como HISPAL (Bick, 2006), Slot Unification Grammar (Ferrández et al., 2000) o Spanish Resource Grammar en HPSG (Marimón et al., 2010). En catalán, los recursos sintácticos desarrollados son menos abundantes y, en general, se trata de gramáticas con un nivel más alto de superficialidad (Alsina et al., 2002; Castellón et al., 1998)

En esta línea, las gramáticas de dependencias EsTxala y CaTxala se proponen como un recurso libre que proporciona árboles sintácticos profundos y robustos. Ambas gramáticas se componen de un conjunto de reglas heurísticas desarrolladas manualmente y basadas en conocimiento sintáctico. Además, el analizador de dependencias de FreeLing, TXALA (Atserias et al., 2005), permite disponer de una serie de recursos lingüísticos de diferente naturaleza asociados a cada gramática

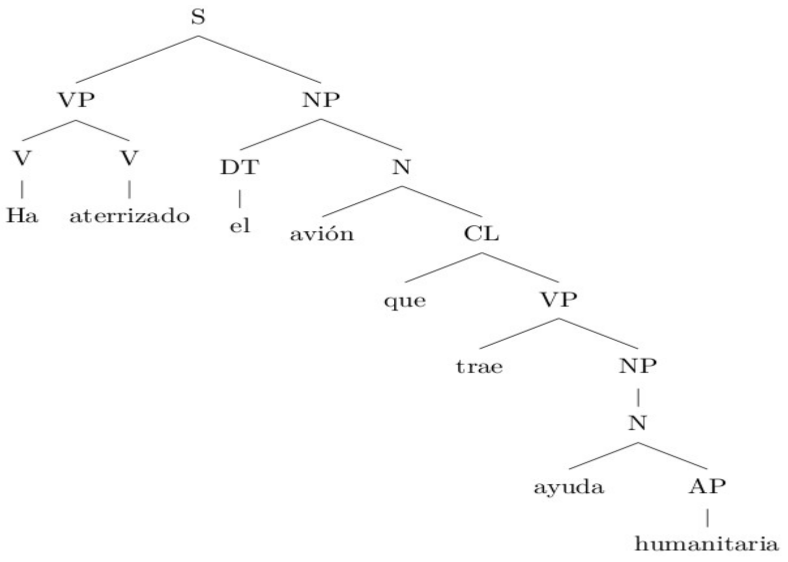
En el análisis automático del lenguaje, se han implementado diferentes formalismos sintácticos (Head–Driven Phrase Structure Grammar, Constraint Grammar, análisis de constituyentes, etc.), pero uno de los formalismos que en los útimos años está siendo utilizado es el análisis de dependencias que tiene como origen los postulados de la Sintaxis Estructural (Tèsnière, 1959).

El formalismo de dependencias (Tèsnière, 1959; Mel'čuk, 1988) parte de la idea que la oración es una red de conexiones entre las unidades léxicas ya que entre todas ellas se establecen relaciones. A diferencia del análisis de constituyentes (Chomsky, 1981), donde los nodos terminales se proyectan en unidades más complejas o constituyentes para expresar dichas relaciones (1b), las relaciones de dependencia se establecen entre los mismos nodos terminales (1c).

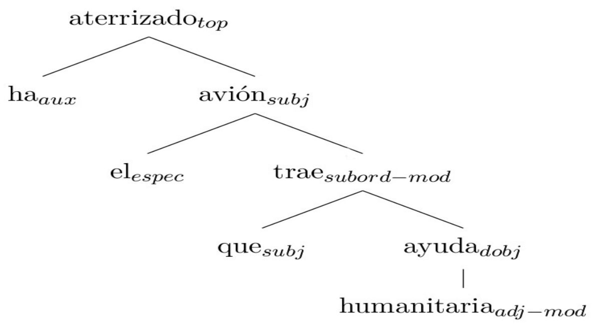
Cada nodo terminal se encuentra dependiendo o modificando otro nodo terminal que, en realidad, es su núcleo sintáctico. No obstante, mientras que un núcleo sintáctico puede tener uno o más modificadores, un nodo dependiente sólo pude modificar a un núcleo sintáctico (1c). Aunque el formalismo de dependencias clásico determina que para cada nodo dependiente existe un único núcleo sintáctico, hay autores (De Marnafee et al., 2006) que proponen un análisis en que los nodos dependientes pueden aceptar dos núcleos sintácticos. De esta manera, pueden explicar la doble funcionalidad del pronombre relativo. En todo caso, estas relaciones se llevan a cabo si la combinación de dos terminales hereda las propiedades sintácticas y semánticas del núcleo sintáctico.

(1)

a. Ha aterrizado el avión que trae ayuda humanitaria.



b.



c.

El análisis que propone el formalismo de dependencias va más allá de la sintaxis ya que las dependencias sintácticas son, en realidad, una representación muy próxima a la estructura semántica del predicado (Mel'čuk, 1988). De modo que los analizadores automáticos que buscan una cierta representación semántica encuentran en el análisis de dependencias el formalismo idóneo para hacer un análisis profundo y completo de la sintaxis de la oración.

En las últimas décadas, ha crecido el interés en el desarrollo de sistemas de análisis sintáctico automático basados en las dependencias sintácticas (Tapanainen y Järvinen, 1998; Collins, 2000; De Marnafee et al., 2006; Nivre, 2006) con el objetivo de superar o, almenos, de mejorar el nivel sintáctico en PLN.

Lenguas como el inglés tienen a disposición un amplio repertorio de analizadores de dependencias (Tapanainen y Järvinen, 1998; Collins, 2000; De Marnafee et al., 2006; Nivre, 2006). A pesar de eso, hay otras lenguas, como el español y el catalán, que progresivamente están quedando al margen de estos avances tecnológicos ya que gozan de pocos sistemas de análisis sintáctico automático, gramáticas y recursos. Existen muy pocos analizadores basados en este formalismo para el español, como DILUCT (Calvo y Gelbukh, 2006), DepPattern (Gamallo y González, 2009), MaltParser (Nivre, 2006) y Connexor (Tapanainen y Järvinen, 1998). A su vez, esta escasez es mayor al observar los analizadores y gramáticas para el catalán (Alsina et al., 2002; Castellón et al., 1998).

El resto del artículo se estructura de la siguiente forma. En primer lugar, en la sección 2, tratamos la arquitectura del analizador en el cual se han integrado las gramáticas EsTxala y CaTxala para, posteriormente, describir cada una de ellas (sección 3). A continuación, presentamos la evaluación realizada de EsTxala (sección 4). Finalmente, exponemos las conclusiones de esta investigación (sección 5).

2. Arquitectura del analizador de dependencias TXALA

TXALA (Atserias et al., 2005) es un analizador de dependencias integrado en FreeLing (Padró et al., 2010). Puesto que TXALA es el último módulo que actúa en la cadena de herramientas de FreeLing, parte de los árboles sintácticos de constituyentes parciales generados por el analizador TACAT (Castellón et al., 1998).

El objetivo principal de TXALA es, dada una gramática, obtener análisis profundos y no ambiguos para cualquier cadena de entrada. Así, la robustez es una de las características de este analizador. Para ello, TXALA construye el análisis en tres etapas. En primer lugar, construye un árbol completo a partir de los fragmentos analizados por TACAT. Simultáneamente, convierte este árbol de constituyentes en un árbol de dependencias. Por último, etiqueta las funciones sintácticas de cada relación de dependencia.

Las dos primeras operaciones se llevan a cabo mediante un conjunto de reglas heurísticas definidas manualmente y que combinan los subárboles adyacentes resultantes del análisis parcial. Para poder controlar estas combinaciones, el analizador emplea un sistema de prioridades numérico integrado en las reglas de cada gramática que determina el orden de aplicación del conjunto de reglas.

Una vez completado el árbol de constituyentes, TXALA transforma esta estructura a una estructura de dependencias, obteniendo un árbol donde se explicitan los núcleos y los dependientes aún sin función sintáctica.

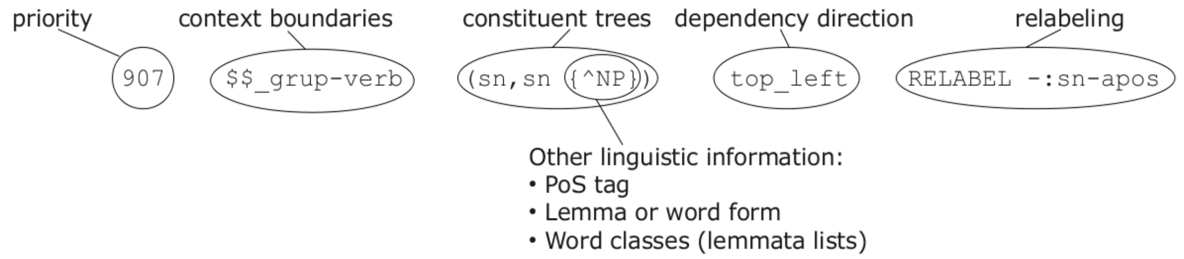


Figura 1. Regla de estructura para sintagmas nominales apositivos

En la figura 1, se observan las diferentes partes que forman una regla, la prioridad y una serie de condiciones. La regla ejemplificada en la figura 1 tiene asignada la prioridad *907* y se aplica si dos sintagmas nominales adyacentes (sn), uno de los cuales es un nombre propio (NP), se encuentran en posición preverbal ($$_grup–verb). El primero de los dos sintagmas se convierte en el núcleo sintáctico de la relación de dependencia (top_left) y el nodo dependiente se recategoriza como *sn–apos*.

Por último, se añaden las etiquetas de función sintáctica al árbol de dependencias. Para ello, se aplican reglas funcionales, también desarrolladas manualmente, que tienen acceso a diferentes recursos o estructuras informativas.

Estas reglas permiten la exploración de los niveles del árbol (inferiores o superiores), el acceso a fuentes de información como la información morfosintáctica de las unidades (categoría y lema), y el acceso o consulta a clases de palabras externas a la gramática construidas a priori (por ejemplo, información sobre la subcategorización de las unidades). Por último, TXALA permite realizar consultas al recurso Multilingual Central Repository y, en concreto, permite navegar por la jerarquía de WordNet y consultar los rasgos de la Top Concept Ontology (TCO).

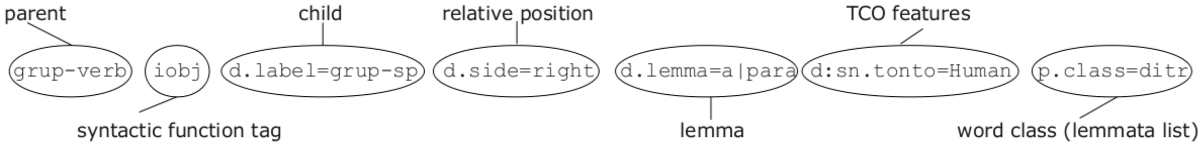


Figura 2. Regla de etiquetaje funcional para objetos indirectos

En la figura 2, se recoge un ejemplo de regla con las condiciones asociadas. En este caso, se establece que el sintagma preposicional (grup–sp) que depende del núcleo verbal (grup–verb) sea etiquetado como objeto indirecto (iobj) si se cumplen las condiciones siguientes: (1)

el sintagma preposicional es postverbal (d.side=right), (2) tiene como terminal el lema *a* (d.lemma=a), (3) contiene un sintagma nominal con valor *humano* según TCO (d:sn.tonto), y (4) el núcleo verbal pertenece a la clase de verbos ditransitivos (p.class=ditr).

3. Las gramáticas EsTxala y CaTxala

En el momento de desarrollar las gramáticas del español (EsTxala) y del catalán (CaTxala), optamos por desarrollarlas en paralelo, dada la proximidad de las lenguas y el gran número de estructuras similares.

EsTxala incluye un total de 4.487, de las cuales 3.810 son estructurales y 677 explicitan las funciones sintácticas. La gramática CaTxala está formada por 3.011 reglas, que se distribuyen en 2.409 reglas de estructura sintáctica y en 602 reglas de función sintáctica. Ambas gramáticas son capaces de resolver estructuras intransitivas, transitivas, ditransitivas, oraciones con argumentos preposicionales e impersonales. Un ejemplo del análisis se puede observar en las figuras 3 y 4.

Las funciones utilizadas en ambas gramáticas parten de las propuestas de I.A. Mel'čuk en *Dependency Syntax: Theory and Practice* (1988). En cuanto a la estructura, se ha optado por una solución ecléctica. Si bien la esencia de las gramáticas recoge el trabajo de Mel'čuk (1988), algunos aspectos de gran importancia no se tratan mediante los postulados de las dependencias sintácticas, sino más bien mediante los criterios de la sintaxis generativa (Chomsky, 1981).

La justificación de estos cambios viene dada por la correspondencia entre los niveles sintáctico y semántico. Además, este criterio permite estructurar la oración de forma que elementos del mismo nivel, como el caso de la coordinación, no dependan el uno del otro. De esta manera, se manifiesta de una forma clara su codependencia. El formalismo que EsTxala y CaTxala toman de base concede importancia a las unidades léxicas y no a los tradicionalmente llamados relatores o categorías funcionales.

Por ese motivo, las unidades léxicas que la sintaxis estructural (Tesnière, 1959) considera relacionales y, en consecuencia, no nucleares (como preposiciones, conjunciones subordinadas y conjunciones coordinantes), en EsTxala y CaTxala, son tratadas como nodos nucleares.

Según este criterio, por ejemplo, en el sintagma preposicional, el núcleo del sintagma debe ser la preposición, ya que, por un lado, la preposición es necesaria para la existencia de este sintagma. Por otro lado, esta unidad es la pieza léxica generalmente subcategorizada por otros núcleos. En otras palabras, en los ejemplos de (2) el verbo 'parlar' ('hablar') (2a) e 'ir' (2b) requieren un argumento necesariamente preposicional con valores concretos, un argumento "regido", por lo que una representación como la que proponemos creemos que representa más claramente este fenómeno (figuras 3 y 4).

(2)

a. El Joan parla de política. ('Juan habla de política.')

b. Los periodistas van a Berlín.

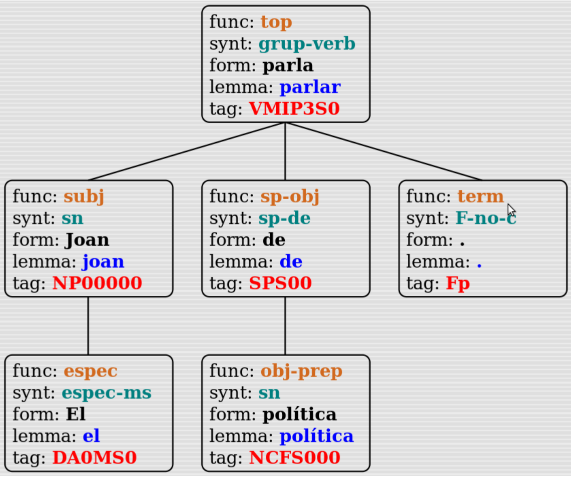


Figura 3. Análisis de CaTxala de (5a)

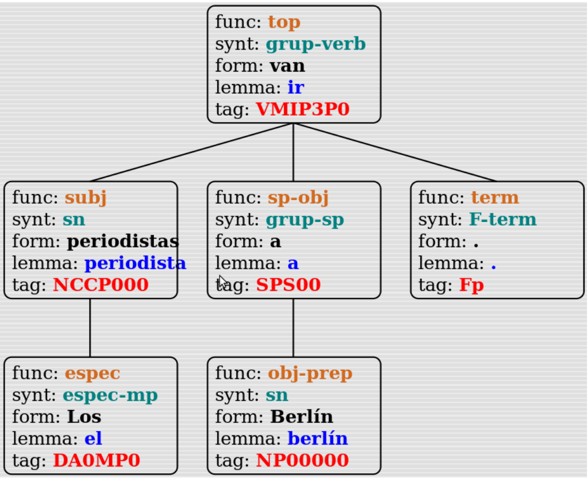


Figura 4. Análisis de EsTxala de (5b)

Además de las reglas, las dos gramáticas, EsTxala y CaTxala, utilizan conocimiento lingüístico de diferente nivel externo al analizador:

- Conocimiento semántico mediante consultas a WordNet y a la TCO (Top Concept Ontology), este acceso permite comprobar la pertenencia de las palabras a determinadas clases semánticas.
- Conocimiento sintáctico como información sobre subcategorización verbal o nominal extraída del corpus SenSem (Alonso et al., 2007) en el caso del español, o del lexicón Volem Multilingüe (Fernández et al., 2002) en el caso del catalán.
- Conocimiento léxico, es decir, acceso a clasificaciones de unidades basadas en su naturaleza semántica (nombres propios de persona, de lugares, etc.) o bien a listas de unidades que informan de su comportamiento en la oración, por ejemplo, un léxico de marcadores del discurso.

Como es lógico, los fenómenos que hemos tratado con especial interés en estas gramáticas afectan tanto a la estructura como a la función. Respecto a la estructura, nos hemos centrado, con especial interés entre otros, en la identificación de los núcleos de los que dependen los sintagmas preposicionales (nominales o verbales) (Lloberes et al., 2010) dado que es una de las estructuras más ambiguas (Figura 5). En la asignación de funciones sintácticas, por ejemplo, hemos trabajado la distinción entre complemento directo de persona y el complemento indirecto para el español (Figura 6).

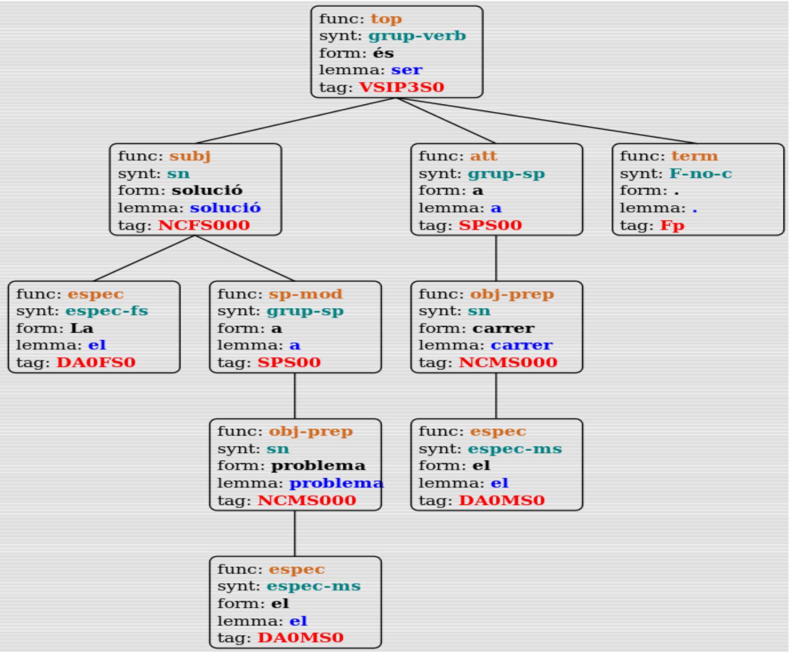


Figura 5. Análisis en CaTxala de agrupación del sintagma preposicional

(La solució al problema és al carrer. 'La solución al problema está en la calle')

4. Evaluación

Dado que los recursos que desarrollamos son herramientas para ser empleadas en otras aplicaciones del PLN (como traducción automática, extracción de información, etc.), es necesario conocer la cualidad de estos recursos. Para ello, se ha llevado a cabo una evaluación rigurosa de la gramática. Esta tarea tiene en cuenta tanto el análisis cuantitativo, que permite determinar la cobertura lingüística, como también el análisis de errores, donde se ponen de manifiesto las principales debilidades del sistema.

La evaluación cuantitativa no es simple, dado que en general los sistemas presentan diferentes *tagsets* (conjunto de categorías) y muchas veces no se pueden establecer correspondencias unívocas entre las categorías de los diferentes sistemas. Por otro lado, los criterios aplicados en dos sistemas de análisis sintáctico pueden diferir e incluso las estructuras o la organización arbórea de las oraciones o sintagmas pueden ser muy diferentes.

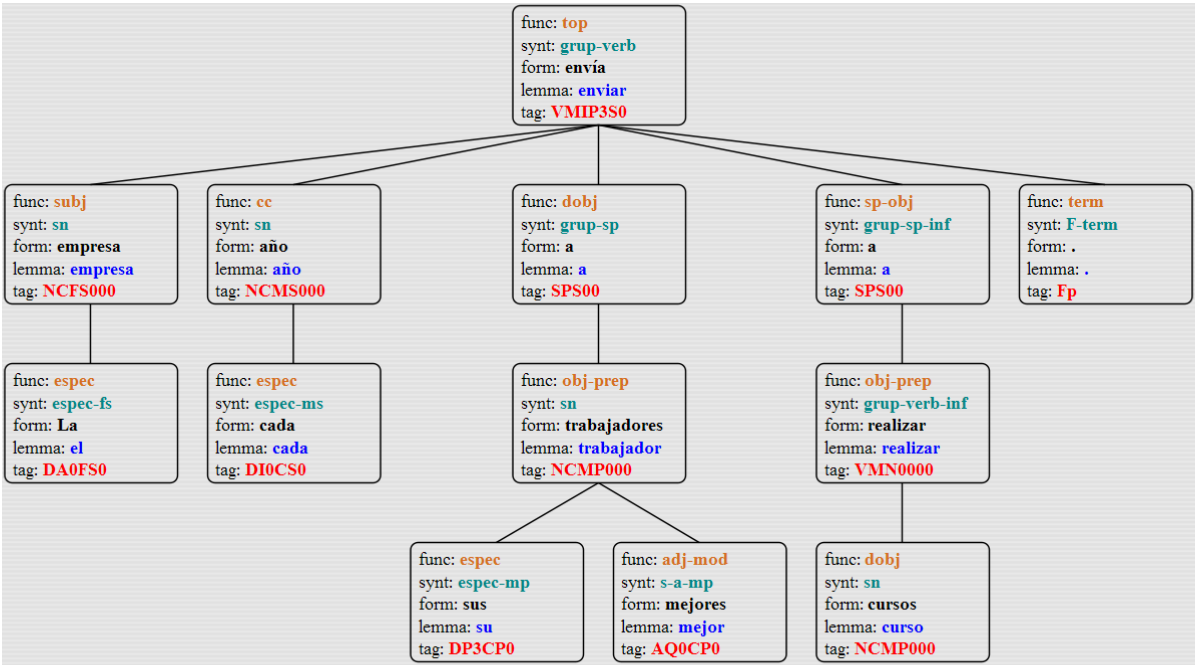


Figura 6. Análisis en EsTxala del reconocimiento objetos directos
(La empresa envía cada año a sus mejores trabajadores a realizar cursos.)

Dada esta complejidad, hasta el momento únicamente estamos trabajando en la evaluación de EsTxala y más adelante nos planteamos extenderla a CaTxala. Por lo tanto, la evaluación que presentamos se debe considerar preliminar en cuanto a la evaluación cuantitativa, pero ha sido bastante exhaustiva en el análisis de errores.

Para la evaluación y análisis de errores hemos utilizado dos corpus, AnCora (Martí et. al., 2007) y SenSem (Alonso et al., 2007), el primero nos permite realizar la evaluación cuantitativa, dado que es un corpus de 550.000 palabras analizado manualmente en el formalismo de dependencias. Para el análisis de errores hemos usado dos corpus, el mismo AnCora y SenSem.

4.1. Evaluación cuantitativa

Como hemos apuntado, ha sido necesario un trabajo extenso para establecer las correspondencias entre las categorías y criterios de AnCora y de EsTxala. Consideramos los datos obtenidos como provisionales dado que la correspondencia aún no está finalizada.

Para esta evaluación hemos utilizado el sistema de evaluación de CoNLL de las tareas compartidas de la edición de 2006. Este sistema tiene en cuenta tres métricas:

- Etiquetaje de funciones y detección de núcleo (*Labeled Attachment*, 'LA'): mide la cantidad de árboles que tienen asignado el núcleo correcto y la función sintáctica correcta.
- Detección de núcleo *(Unlabeled Attachment,* 'UA'): mide el conjunto de árboles que tienen asignado el núcleo correcto.
- Etiquetaje de funciones (*Label Accuracy*, 'LAcc'): mide la cantidad de árboles en los cuales se ha asignado la función sintáctica correcta.

Actualmente y evaluando contra AnCora, nuestro sistema tiene un 73,88% de precisión en relación con la detección del núcleo y función, un 81,13% de los árboles tienen asociado el núcleo correcto y un 78,81% han sido etiquetados con la función sintáctica correcta. Si comparamos con otros sistemas evaluados con este mismo corpus, se observa que, aunque no alcanzamos a los mejores analizadores estadísticos, que se encuentran por encima del 80% de precisión en cuanto a estructuras y funciones sintácticas correctas, la gramática se sitúa ya en un nivel aceptable.

4.2. Análisis de errores

Para analizar el conjunto de errores que comete la gramática y la motivación de ellos, se ha llevado a cabo una evaluación controlada de los datos de AnCora y SenSem. Además, se ha generado un corpus derivado de cada uno de ellos para poder evaluar determinados fenómenos lingüísticos que incluye EsTxala, por un lado, y para probar la transportabilidad de la gramática entre diferentes corpus, algo que los analizadores estadísticos no consiguen con un nivel aceptable.

En total partimos de cuatro corpus de pequeño tamaño para poder llevar a cabo dicho análisis:

- AnCoraR: 25 frases reales del corpus AnCora (sin modificar) que son representativas de diferentes fenómenos lingüísticos.
- SenSemR: 25 frases reales del corpus SenSem (sin modificar) que son representativas de diferentes fenómenos lingüísticos.

- AnCoraS: corpus de frases simples producido a partir de las 25 frases del corpus AnCoraR.
- SenSemS: corpus de frases simples producido a partir de las 25 frases del corpus SenSemR.

El hecho de generar los corpus de frases simples nos interesa para aislar fenómenos y poder detectar donde el sistema falla en aspectos sintácticos básicos. Mientras que los corpus reales nos informan de los errores que provocan las construcciones más complejas, por ejemplo, las oraciones subordinadas y las coordinaciones oracionales.

Además, para comprobar que la muestra es representativa, evaluamos cuantitativamente los resultados de aplicar la gramática EsTxala a estos cuatro corpus, los resultados se pueden ver en la tabla 1.

Corpus	LA	UA	LAcc
AnCoraR	73.88	81.13	78.81
AnCoraS	85.46	92.22	87.37
SenSemR	74.33	80.93	77.28
SenSemS	85.02	91.82	85.85

Tabla 1. Resultados de la evaluación cualitativa de EsTxala

Como se observa en la tabla 1, las cifras aumentan en las variantes de los corpus simples (AnCoraS y SenSemS) si los comparamos con los corpus reales AnCoraR y SenSemR. Ello es debido a la diferencia de complejidad de las oraciones de estos corpus. Por lo tanto, la simplicidad en la oración es un factor a favor en vez de la complejidad (por ejemplo, en las oraciones subordinadas).

Respecto a los errores observados, desde el punto de vista estructural las gramáticas tienen menor índice de acierto tratando la adjunción de los sintagmas preposicionales, el alcance de la coordinación y la subordinación. En cuanto a la distinción funcional, la distinción entre complemento y adjunto para algunos sintagmas preposicionales es uno de los puntos más débiles de la gramática.

Tanto los resultados de la evaluación cuantitativa como los del análisis de errores pueden ser mejorados. Puesto que la versión actual de la gramática contiene básicamente información sintáctica, nuestra hipótesis actual es que la adición de información semántica puede mejorar estos niveles de acierto.

Igualmente, con el fin de solucionar estos errores, tenemos la intención de integrar un modelo estadístico para alguno de los fenómenos, haciendo que TXALA sea un analizador híbrido. Otras posibles soluciones, como integrar la desambiguación semántica y la construcción de un modelo lingüístico de restricciones selectivas también serán exploradas.

5. Conclusiones

En este artículo, hemos presentado dos gramáticas computacionales, EsTxala y CaTxala que se basan en el formalismo de dependencias y en reglas con conocimiento lingüístico, básicamente sintáctico pero también semántico. Tanto EsTxala como CaTxala tratan de manera profunda y robusta un amplio repertorio de fenómenos del español y del catalán, incluyendo aquellos fenómenos lingüísticos complejos de resolver des del punto de vista del PLN (agrupación del sintagma preposicional, formación de la subordinación y de la coordinación, reconocimiento de argumentos y adjuntos, etc.).

Puesto que este par de recursos están diseñados para ser un módulo de otras aplicaciones computacionales, estamos evaluando exhaustivamente su cualidad. La tarea de evaluación, como se ha apuntado, es compleja y, por esa razón, hemos elaborado una metodología para llevar a cabo esta tarea con garantías.

Esta metodología nos ha permitido obtener los resultados cuantitativos del proceso de evaluación y elaborar un análisis de errores inicial. El primer conjunto de resultados informa que, aunque los analizadores sintácticos automáticos más precisos se sitúan por encima del 80% de precisión (en relación con los análisis con asignación de núcleo y de etiquetaje de funciones correctos), EsTxala se aproxima a éstos. Los resultados cualitativos determinan que la complejidad de la oración (subordinadas y coordinadas), la detección del núcleo del sintagma preposicional y el reconocimiento de argumentos (sobre todo aquellos que están regidos por una preposición) son fenómenos que en etapas posteriores deben ser resueltos.

Actualmente, estamos trabajando para mejorar los resultados obtenidos en las tareas de evaluación y estamos detallando las correspondencias de los recursos que utilizamos para evaluar con el objetivo de llevar a cabo una evaluación completa en relación con el análisis cuantitativo y el análisis de errores.

Igualmente, estamos planteando nuevas vías de investigación para incrementar la precisión de las gramáticas. Éstas están encaminadas a tratar de manera diferente la información asociada a las gramáticas, a incluir otros tipos de conocimiento lingüístico (semántica), y a emplear conocimiento estadístico en aquellos aspectos en que el conocimiento lingüístico no puede aportar más información.

Referencias

ALONSO, Laura, Joan Antoni CAPILLA, Irene CASTELLÓN, Ana FERNÁNDEZ y Glòria VÁZQUEZ, 2007. «The SenSem project: Syntactico–semantic annotation of sentences in Spanish», en Nicolas Nicolov, Kalina Bontcheva, Galia Angelova y Ruslan Mitkov (ed.), *Recent Advantages in Natural Langage Processing IV. Selected papers from RANLP 2005*, Amsterdam–Philadelphia: John Benjamins Publishing Co. 89–98.

ALSINA, Àlex, Toni BADIA, Gemma BOLEDA, Stefan BOTT, Àngel GIL, Martí QUIXAL y Oriol VALENTÍN, 2002. «CATCG: Un sistema de análisis morfosintáctico para el catalán», *Procesamiento del Lenguage Natural*, 29. 309–310.

ATSERIAS, Jordi, Elisabeth COMELLES y Aingeru MAYOR, 2005. «TXALA un analizador libre de dependencias para el castellano», *Procesamiento del Lenguaje Natural*, 35. 455–456.

BICK, Eckhard, 2006. «A constraint grammar–based parser for Spanish», en *Proceedings of TIL 2006, 4th Workshop on Information and Human Language Technology*.

CALVO, Hiram y Alexander GELBUKH, 2006. «DILUCT: An open–source Spanish dependency parsers based on rules, heuristics, and selectional preferences», en Christian Kop, Günther Fliedl, Heinrich C. Mayr y Elisabeth Métais (ed.), *Natural Language Processing and Information Systems. Lecture Notes in Computer Science*. Berlin: Springer–Verlag, 3999. 164–175.

CASTELLÓN, Irene, Montse CIVIT y Jordi ATSERIAS, 1998. «Syntactic parsing of unrestricted Spanish text», en *Proceedings of First International Conference on Language Resources and Evaluation*.

CHOMSKY, Noam, 1981. *Lectures on Government and Binding. The Pisa Lectures*. Holland: Foris Publications. Reprint. 7th Edition. Berlin and New York: Mouton de Gruyter, 1993.

COLLINS, Michael, 2000. «Discriminative reranking for natural language parsing», en *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000)*. 175–182.

FERNÁNDEZ, Ana, Patrik SAINT–DIZIER, Glòria. VÁZQUEZ, Farah BENAMARA , Mouna KAMEL, 2002. «The VOLEM Project: a Framework for the Construction of

Advanced Multilingual Lexicons», en *Proceedings of the Language Engineering Conference*. 89–98.

FERRÁNDEZ, Antonio, Manuel PALOMAR y Lídia MORENO, 2000. «Slot unification grammar and anaphora resolution», en Nicolas Nicolov y Ruslan Mitkov (ed.), *Recent Advances in Natural Language Processing II. Selected papers from RANLP 1997*. Amsterdam–Philadelphia: John Benjamins Publishing Co.155–166.

GAMALLO, Pablo y Isaac GÓNZALEZ, 2009. «Una gramática de dependencias basada en patrones de etiquetas», *Procesamiento del Lenguaje Natural*, 43. 315–323.

LIN, Dekang, 1998. «Dependency–based evaluation of MINIPAR», en *Proceedings of the LREC Workshop on the Evaluation of Parsing Systems*.

LLOBERES, Marina, Irene CASTELLÓN y Lluís PADRÓ, 2010. «Spanish FreeLing Dependency Grammar», en Nicoleta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner y Daniel Tapias (ed.), *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA).

MARIMÓN, Montserrat, 2010. «The Spanish Resource Grammar», en Nicoleta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner y Daniel Tapias (ed.), *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA).

de MARNAFEE, Marie–Catherine, Bill MAcCARTNEY y Christopher D. MANNING, 2006. «Generating typed dependency parses from phrase structure parses», en *Proceedings of the Fifth Conference on International Language Resources and Evaluation (LREC'06)*. 449–454.

MARTÍ, M. Antònia, Mariona TAULÉ, Lluís MÀRQUEZ y Manuel BERTRAN, 2007. *AnCora: Multilingual and multilevel annotated corpora*. http://www.clic.ub.edu.

MEL'ČUK, Igor A, 1988. *Dependency Syntax: Theory and Practice*. Albany, New York: State U. Press of NY.

NIVRE, Joakim, 2006. *Inductive Dependency Parsing*. Dordrecht: Springer–Verlag.

PADRÓ, Lluís, Samuel REESE, Marina LLOBERES y Irene CASTELLÓN, 2010. «FreeLing 2.1: Five years of open–source language processing tools», en Nicoleta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner y Daniel Tapias (ed.), *Proceedings of the Seventh Conference on International Language*

Resources and Evaluation (LREC'10). Valletta, Malta: European Language Resources Association (ELRA).

SLEATOR, Daniel D.K. y Davy TEMPERLEY, 1991. «Parsing English with a Link Grammar», en *Third International Workshop on Parsing Technologies*.

TAPANAINEN, Pasi y Timo JÄRVINEN, 1998. «Towards an implementable dependency grammar», en *Proceedings of the ACL Workshop on Processing of Dependency–Based Grammars*. 1–10.

TESNIÈRE, Lucien, 1959. *Eléments de syntaxe structurale*. Paris: Klincksieck.

Spanish FreeLing Dependency Grammar

Marina Lloberes*, Irene Castellón*, Lluís Padró†

*GRIAL Research Group, Universitat de Barcelona
Gran Via de les Corts Catalanes 585, Barcelona
{marina.lloberes,icastellon}@ub.edu

†TALP Research Center, Universitat Politècnica de Catalunya
Jordi Girona 1-3, Barcelona
padro@lsi.upc.edu

Abstract

This paper presents the development of an open-source Spanish Dependency Grammar implemented in FreeLing environment. This grammar was designed as a resource for NLP applications that require a step further in natural language automatic analysis, as is the case of Spanish-to-Basque translation. The development of wide-coverage rule-based grammars using linguistic knowledge contributes to extend the existing Spanish deep parsers collection, which sometimes is limited. Spanish FreeLing Dependency Grammar, named EsTxala, provides deep and robust parse trees, solving attachments for any structure and assigning syntactic functions to dependencies. These steps are dealt with hand–written rules based on linguistic knowledge. As a result, FreeLing Dependency Parser gives a unique analysis as a dependency tree for each sentence analyzed. Since it is a resource open to the scientific community, exhaustive grammar evaluation is being done to determine its accuracy as well as strategies for its manteinance and improvement. In this paper, we show the results of an experimental evaluation carried out over EsTxala in order to test our evaluation methodology.

1. Introduction

Spanish FreeLing Dependency Grammar (EsTxala) was developed as a resource for FreeLing[1], an open-source multilingual NLP library (Atserias et al., 2006). It was designed for those NLP applications that require need deeper syntactic representation or certain level of semantic representation.

Because of deep parsing importance in NLP, a wide range of resources has been developed from different approximations and linguistic formalisms. For languages like English, large amount of deep parsers exists such as MaltParser (Nivre, 2006), Minipar (Lin, 1998), Connexor (Järvinen and Tapanainen, 1998) or Link Parser (Sleator and Temperley, 1991).

However, few broad-coverage parsers and grammars are developed for languages like Spanish, such as Constraint-Grammar for HISPAL parser (Bick, 2006), Slot Unification Grammar developed by Ferrández et al. (2000) or Spanish Resource Grammar in the framework of HPSG (Marimón et al., 2007).

Further, although dependency formalism was implemented in NLP (By, 2004), there are few dependency parsers for Spanish, MaltParser (Nivre, 2006), DILUCT (Gelbukh et al., 2005) and Connexor (Järvinen and Tapanainen, 1998).

One additional problem is that few resources for Spanish are open-source. While MaltParser and DILUCT are totally open-source, Connexor grants a restrictive licence to researchers and HISPAL provides only parsed texts.

On the other hand, among deep parsers for Spanish, most of them are based on statistical knowledge, while Txala (the FreeLing Dependency Parser) relies on hand–written heuristic rules based on linguistic knowledge (Atserias et al., 2005).

Txala parser and its first Spanish grammar was developed in the framework of OpenTrad and EuroOpenTrad, two Open-Source Machine Translation projects aiming to develop transfer translators for all official languages in Spain (Spanish, Catalan, Galician, and Basque), as well as English.

EsTxala grammar has been extended in KNOW project. One goal of the KNOW project is the development of wide-coverage, deep parsing grammars whose outcome will be open to the scientific community.

The rest of the paper is structured as follows. Section 2 introduces the main features of Txala parser and briefly describes Freeling Dependency Grammars development. Section 3 surveys Spanish FreeLing Dependency Grammar and strategies followed to solve some complex linguistic phenomena. Experimental evaluation results are presented in section 4 and conclusions and further work in section 5.

2. FreeLing Dependency Parser

Txala parser is a module in FreeLing processing chain which acts after sentence splitting, morphological analysis, tagging and shallow parsing.

The main aim of Txala parser and EsTxala grammar is to provide deeper and more robust parse trees, solving attachment ambiguity for all structure levels, and always providing a syntactic analysis for any structure. In order to satisfy these two goals, Txala parser carries on three steps, starting from partial trees produced by FreeLing Shallow Parser:

- Build full syntactic tree.
- Convert the full tree into a dependency tree.
- Label the syntactic function of each dependency.

The first step is dealt with a set of manually defined heuristic rules (that describe language structures, not structures

[1] http://www.lsi.upc.edu/~nlp/freeling/

included in corpora) by combining each two adjacent subtrees of a linguistic chain. To attach consecutive subtrees a priority value is assigned to each rule. The rule with highest priority is applied and the pair of subtrees are merged into one.

Apart from priority, rules also express conditions that each subtree head must meet. These conditions can be related to:

- Morphology: PoS tag.
- Lexicon: word form, lemma.
- Syntax: context boundaries of the pair of subtrees, word classes defined as a lemmata lists.
- Semantics: word classes.

Also, the head node is marked on the rules becoming the parent of all subtrees below.

```
907  $$_grup-verb -
     (sn,sn{^NP})
     top_left RELABEL sn-apos
```

Figure 1: Parsing Rules Structure. Example of noun phrase aposition before main verb.

For instance, the rule in Figure 1 has priority 907, and states that when two adjacent noun phrase (sn) chunks –the second having a proper noun (NP) as head– are found with a verb group (grup-verb) immediately to their right, the second noun phrase becomes a child of the first, and the root of the resulting tree is relabeled as sn-apos.

When the tree-completion task is completed, the tree is straightforwardly transformed to a dependency structure. This is possible because the head of each rule is explicitly marked by the shallow parser and by the tree-completion step.

Finally, each dependence is labeled with its syntactic function by another set of rules. They are applied when specific conditions are met by both head and dependent nodes.

At this level, conditions refer to:

- Morphology: PoS tag.
- Lexicon: lemma.
- Syntax: relative position, word classes.
- Semantics: word classes, WordNet semantics files, EuroWordNet top-ontology features.

```
grup-verb subj
     d.label=sn* d.side=right
     p.class=intr
```

Figure 2: Labeling Rules Structure. Example of right subject with intransitive verbs

The example labeling rule in Figure 2 states that a node depending of the head of a verb group (grup-verb) will be labeled as subject (subj) if it is the head of a noun phrase (sn*), located at the right of the verb phrase, and the class for the verb is intransitive (intr).

As a result of the steps described above, Txala parser gives a unique analysis as a dependency tree for each sentence analyzed.

The version of the parser presented in this paper includes some improvements respect to the version described in Atserias et al. (2005), which include:

1. About tree attachment rules:
   - Extension of the catalogue of subtree-fusion operations.
   - Possibility of specifying form, lemma, PoS or word class conditions on subtree heads.
   - Possibility of specifying context conditions (stated as labels corresponding to subtrees).
   - Defining word classes via lists in external files.
2. Labeling rules also accept new conditions regarding:
   - EWN Top Ontology properties.
   - WN semantic file.
   - Synonyms.
   - Hypernyms.

Txala parser also includes dependency grammars for English, Catalan and Galician, but this paper describes the development of FreeLing Spanish Dependency Grammar, EsTxala, which is currently at the most advanced stage of development.

3. EsTxala Grammar

EsTxala includes a set of 4,408 rules. Of those, 3,808 relate to full parsing tree construction, and 600 are used to define dependency relations by labeling each dependency.

The former are used to handle recursion and attachments between phrases, finite clauses (headed by conjunctions or relative pronouns), non-finite clauses (headed by infinitive, participle or gerund), simple coordinations (i.e. between phrases), and passive, among other structures.

Among the latter, labeling rules carry on intrachunk relations and external chunk relations.

Intrachunk relations include labeling determiners and modifiers, which doesn't require much rules.

External chunk relations are based on argument and adjunct recognition, as well as argument or adjunct types distinction, which are cases usually complex to solve. To be able to perform external chunk labeling, EsTxala distinguishes among structures like transitive, intranstitive, ditransitive, prepositional (singled or doubled), and impersonal.

To carry out wide-coverage full syntactic analysis of natural language sentences, complex phenomena (prepositional phrase attachment, coordination, prepositional arguments and prepositional adjuncts) have to be solved. Rules themselves will not succeed without some sort of additional knowledge.

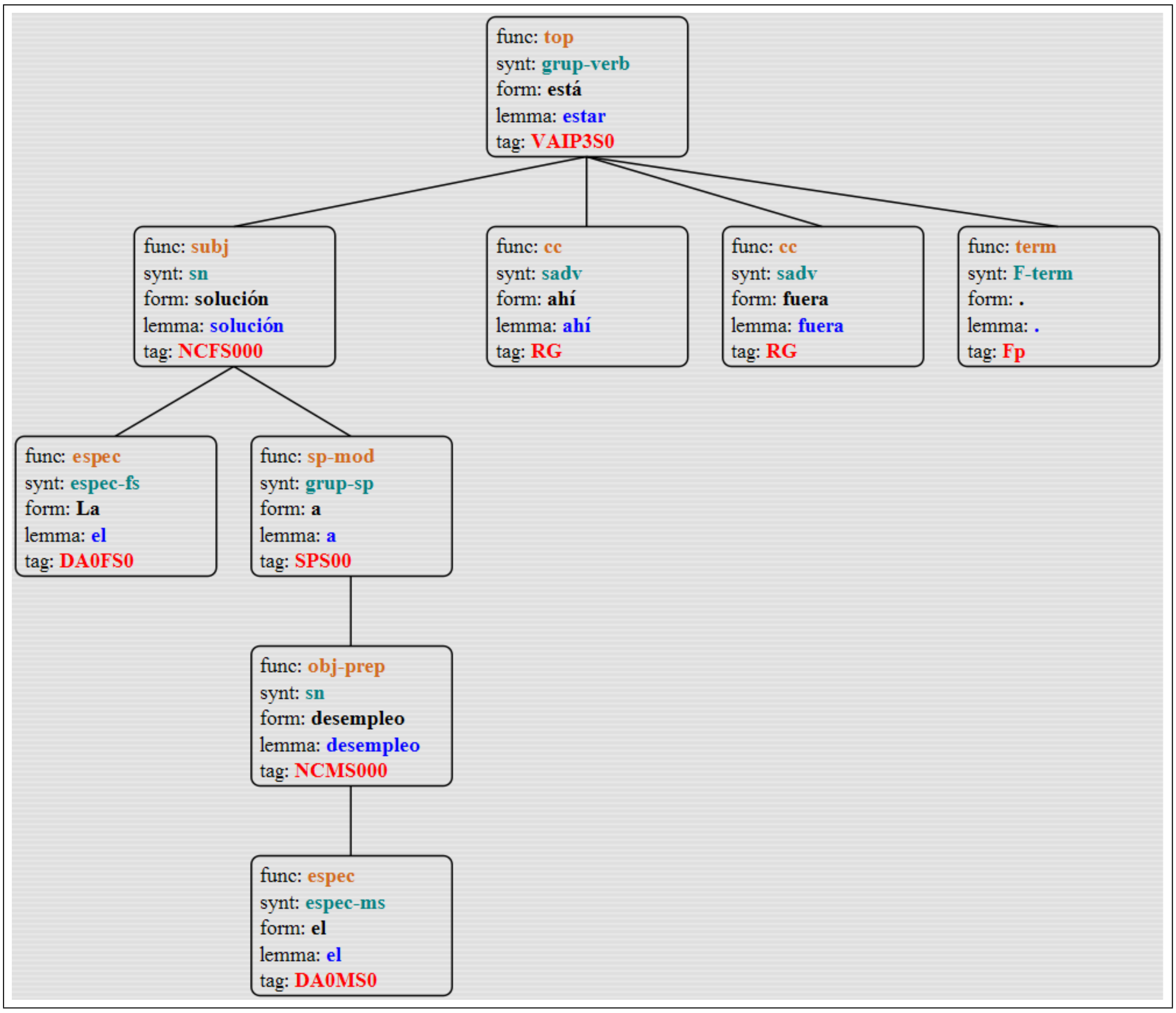


Figure 3: Preposition Phrase Attachment to Noun Phrase – *La solución al desempleo está ahí fuera* ('The solution to unemployment is out there').

EsTxala includes external modules as linguistic knowledge used by rules:

- Semantic knowledge (WordNet, EuroWordNet Top-ontology features).
- Syntactic knowledge: SenSem Corpus (Alonso et al., 2007) has been used to represent verbal subcategorization classes.
- Lexical information: EsTxala include a lexicon of prototypical discourse markers.

One of the main complex phenomena to be solved at EsTxala was prepositional phrase attachment. In Spanish prepositions can modify either a noun phrase –e.g. *La solución al desempleo está ahí fuera* ('The solution to unemployment is out there')– as it is illustrated at Figure 3 or a verb phrase –e.g. *Mi vecina piensa en cambiar de casa* ('My neighbour is thinking about moving to another flat').

Most problems are related to preposition *de* ('of'/'from', genitive among others) because is commonly used as noun phrase modifier –*El libro de Cervantes es bien conocido* ('Cervantes's book is well-known')– as well as argument –*Mi hijo viene del mercado* ('My son comes from the market')– or adjunct –*Empezaron la excursión de madrugada* ('They began the excursion at daybreak').

Nevertheless, adding information about both verb and noun behaviour and defining immediate syntactic context of prepositional phrase allow to partly account for these problematic cases (s. Figure 4).

Preposition phrases in Spanish also are problematic when labeling dependencies. Sometimes they act as argument, sometimes as adjunct, and there are also several arguments whose head is a preposition. It seems that some prepositions accept to be used in more contexts than others, as preposition *a* ('to'/'for').

When a preposition phrase headed by *a* is an argument, it can be a prepositional argument –e.g. *Disfruta yendo al cine cada domingo por la noche* ('He enjoys going to the cinema every Sunday evening')–, indirect object –e.g. *El presidente presentó la ley a los diputados.* ('The president presented the law to the congressmen.'), or direct object referring human entities –e.g. *El juez convocó al empresario* ('The judge summoned the company manager')–.

EsTxala labeling rules decide which phrases are verb arguments and which others are adjuncts by resorting to external linguistic knowledge linked to EsTxala. Sometimes it is

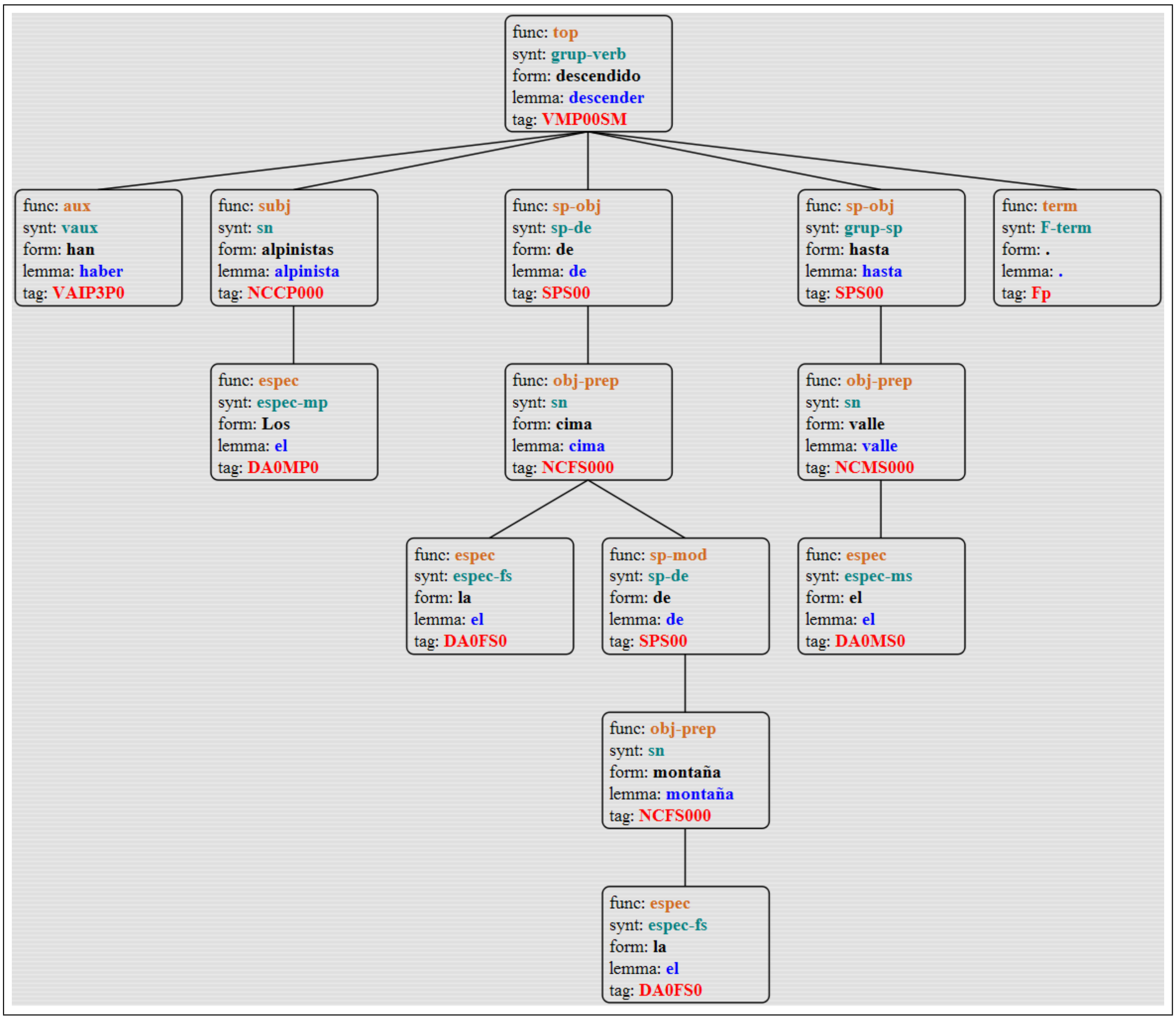


Figure 4: Preposition Phrase Attachment to Verb Phrase – *Los alpinistas han descendido de la cima de la montaña hasta el valle.* (‘Climbers descended from the summit to the valley.’)

necessary to combine informations from different resources depending on phenomena complexity. For example, carrying out direct object referring human entities it is required to consult TCO features and verb diathesis (s. Figure 5).

4. EsTxala Evaluation

Rigorous and exhaustive grammar evaluation requires qualitative and quantitative analysis in order to observe which phenomena fail and which failures are relevant for significatively improving the grammar. In this paper, we present results obtained from experimental evaluation.

Two evaluation corpora are used on this task, AnCora (Martí et al., 2007) and SenSem (Alonso et al., 2007). On this evaluation stage, 25 sentences were randomly selected from AnCora (AnCoraR) and 25 from SenSem (SenSemR) as real corpora samples. On the other hand, EsTxala evaluation statistics were obtained using ‘CoNLL-X Shared Task (2006): Multi-lingual Dependency Parsing’ evaluation script and three metrics are taken into account:

- Labeled Attachment (LA): the amount of trees that are assigned the correct head and dependency relation.
- Unlabeled Attachment (UA): the amount of trees that are assigned the correct head.
- Label Accuracy (LAcc): the amount of trees that are assigned the correct dependency relation.

EsTxala scores satisfactorily in both evaluation corpora (s. Table 1). In AnCoraR, 73.88% of the trees receive correct head and dependency relation jointly (LA), 81.13% are well-headed (UA) and 78.81% are labeled with correct dependency relation (LAcc). Regarding SenSemR, similar scores are obtained: 74.33% of the trees have correct head and dependency relation jointly (LA), 80.93% have correct head (UA) and 77.28% get correct dependency relation.

In order to determine whether these rather low scores are caused motivated by sentence complexity (i.e. finite and non-finite clauses), complex sentences were isolated by hand from other linguistic phenomena present at both evaluation corpora. AnCoraR and SenSemR were transformed into two single-clause samples: AnCoraS and SenSemS.

As expected, the simple sentences corpora AnCoraS and SenSemS obtain best scores (s. Table 1), increasing about 10 points at the three metrics taken into account. Therefore, while trees are well-built in single clauses context, complex

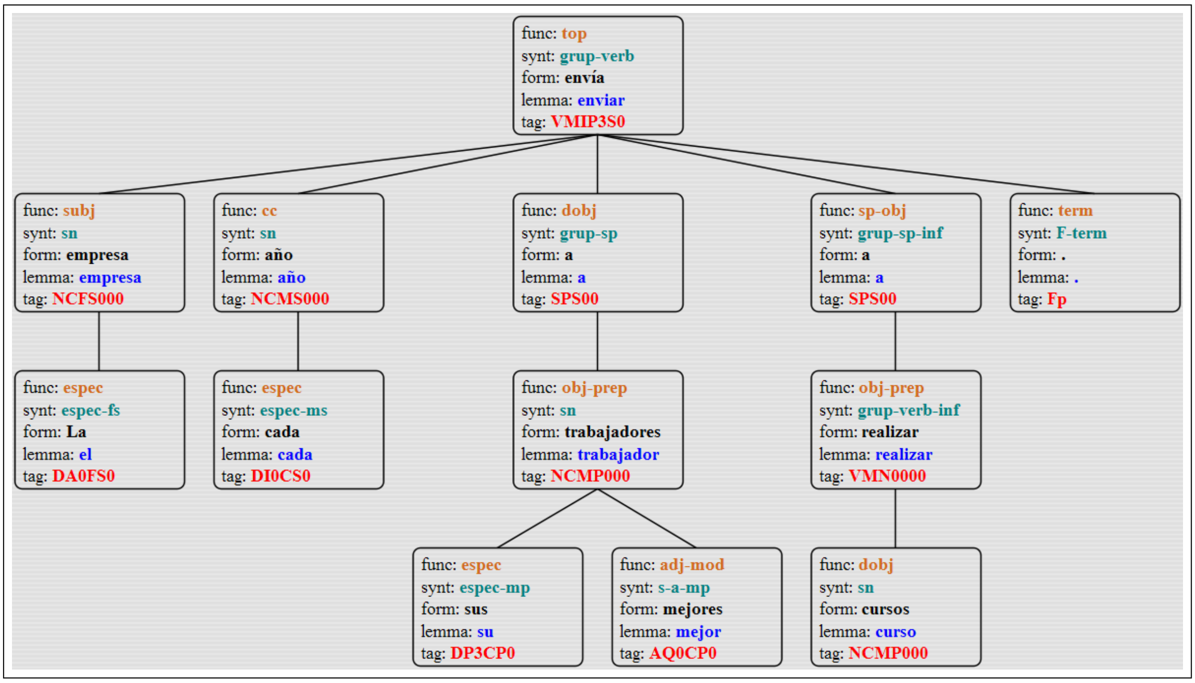


Figure 5: Argument/adjunct recognition – *La empresa envía cada año a sus mejores trabajadores a realizar cursos.* ('The company sends every year their best employees to take courses.')

corpora	LA	UA	LAcc
AnCoraR	73.88	81.13	78.81
AnCoraS	85.46	92.22	87.37
SenSemR	74.33	80.93	77.28
SenSemS	85.02	91.82	85.85

Table 1: EsTxala Accuracy Scores

	AnCora		SenSem	
PoS	Real	Simple	Real	Simple
ADJ	94.74	97.30	91.43	91.43
COOR	44.44	71.43	23.81	42.86
CONJ	58.82	33.33	55.00	50.00
DET	95.83	98.32	99.15	99.15
NOUN	91.16	94.54	90.71	94.27
PRON	81.48	92.59	92.59	97.10
REL	62.50	50.00	43.48	0.00
ADV	53.85	74.07	83.33	96.15
PREP	71.07	83.05	69.23	80.92
VERB	72.73	96.55	78.26	95.87

Table 2: EsTxala UA Accuracy

clauses formation still is problematic to deal with in EsTxala.

In terms of unlabeled attachment score (s. Table 2), best results in both corpora and their single clauses variants are found on those nodes placed near to terminal nodes like determiner (DET), noun (NOUN) or adjective (ADJ).

Also, phenomena usually difficult to solve in NLP are quite problematic in EsTxala (s. Table 2): Coordination (COOR) and clauses (CONJ and REL) are quite low a part from prepositional phrase attachment. Finite clauses (CONJ) score 58.82% in AnCoraR and 55.00% in SemSemR. Relative clauses (REL) are frequently problematic (62.50% in AnCoraR and 43.48% in SenSemR)[2]. Coordination rules succeed in few cases (44.44% in AnCoraR, 23.81% in SenSemR and 42.86% in SemSemS), but are built quite satisfactorily (71.43%) in AnCoraS. However, prepositional phrase attachment (PREP) is well-built (71.07% in AnCoraR, 69.23% in SenSemR, 83.05% in AnCoraS and 80.92% SenSemS) in more cases than coordination or clauses.

Regarding labeled attachment score, best results (s. Table 3) are found in those tags related to internal phrase relations like some noun modifiers (adj-mod, sp-mod, subord-mod), determiners (espec), or auxiliaries (aux).

On the other hand, dependency labels for relations between main verb and its children show some problems. Relations like subject (subj), patient as subject (subj-pac) and direct object (dobj) succeed satisfactorily. However, most difficulties are found in prepositon-headed arguments or adjuncts. Regarding verb arguments, indirect object (iobj) scores 58.82% in AnCoraR and 44.44% in SenSemR, and prepositional argument (sp-obj) scores 50% in AnCoraR and 45.28% in SenSemR. Accuracy in predicate adjuncts (cc) reaches 59.77% in AnCoraR and 56.64% in SenSemR, and sentence adjuncts (ador) score 52.18% in AnCoraR and 40% in SenSemR. Simple clauses samples are similar to real corpora samples, although they slightly increase accuracy scores.

5. Conclusions & further work

In this paper we presented an open-source dependency grammar for Spanish, implemented in FreeLing environ-

[2] Some conjunctions and relative pronouns appear in AnCoraS and SenSemS, but these occurrences are not considered clause markers.

	AnCora		SenSem	
Function	Real	Simple	Real	Simple
adj-mod	91.36	91.89	87.81	92.10
ador	52.18	53.33	40.00	20.00
att	53.33	78.57	50.00	76.92
aux	100.00	100.00	100.00	94.74
cc	59.77	64.37	56.64	64.96
co-n	73.69	94.12	76.19	85.71
co-sp	-	-	44.44	40.00
co-v	41.38	-	75.68	-
dep	66.67	75.00	100.00	100.00
dobj	79.02	86.42	69.03	83.76
dprep	100.00	100.00	100.00	100.00
dverb	100.00	100.00	92.31	92.31
es	82.35	75.00	91.67	91.67
espec	95.49	96.99	98.82	99.22
iobj	58.82	66.67	44.44	66.67
obj-prep	94.17	96.99	94.86	99.24
sn-mod	76.92	84.62	51.85	51.85
sp-mod	88.00	90.00	77.65	80.46
sp-obj	50.00	43.24	45.28	43.34
subj	82.86	96.15	70.77	85.19
subj-pac	50.00	80.00	-	-
subord-mod	89.66	100.00	74.28	-
top	69.39	97.50	65.31	97.30
vsubord	93.11	100.00	92.31	-

Table 3: EsTxala LAcc F1

ment. EsTxala was developed as a broad-coverage rule-based grammar relying on linguistic information. We have also described the most recent update of the Txala parser, which features a number of improvements over its predecessor (Atserias et al., 2005).

Finally, we exposed results from a limited experimental evaluation: 73.88% (Labeled Attachment Accuracy), 81.13% (Unlabeled Attachment Accuracy), 78.81% (Label Accuracy) in AnCora, and 74.33% (Labeled Attachment Accuracy), 80.93% (Unlabeled Attachment Accuracy), 77.28% (Label Accuracy) in SenSem. Results from first expermients encourage developing an exhaustive evaluation.

Because evaluating precision and coverage of grammars like EsTxala is a complex task, we are still developing experiments to determine EsTxala accuracy in terms of qualitative and quantitative analysis. These experiments also aim to find out whether the use of linguistic knowledge improves grammar accuracy.

One of the most important evaluation topic is to test external syntactic knowledge included in EsTxala (i.e. verb subcategorization classes), since a large amount of labeling rules depend on it. Studying EsTxala resources evaluation we will verify if syntax knowledge is enough or semantic information is required to improve grammar accuracy.

However, before carrying out quantitative evaluation, we must solve linguistic criteria differences between evaluation corpora and EsTxala. We are developing a mapping between EsTxala and AnCora labeling tags and structures which allows to evaluate EsTxala on CoNLL shared tasks datasets.

This empirical evaluation methodology will also allow to mantain and improve EsTxala in the future.

6. Acknowledgements

This research has been partially funded by the Spanish Science and Innovation Ministry through the project KNOW2 (TIN2009-14715-C04-03, TIN2009-14715-C04-04).

7. References

L. Alonso, J.A. Capilla, I. Castellón, A. Fernández, and G. Vázquez. 2007. The sensem project: Syntactico-semantic annotation of sentences in spanish. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing IV. Selected papers from RANLP 2005*, pages 89–98. John Benjamins Publishing Co., Amsterdam and Philadelphia.

J. Atserias, E. Comelles, and A. Mayor. 2005. Txala un analizador libre de dependencias para el castellano. *Procesamiento del Lenguaje Natural*, 35:455–456.

J. Atserias, B.Casas, E. Comelles, M. González, L. Padró, and M. Padró. 2006. Freeling 1.3: Syntactic and semantic services in an open-source nlp library. In *Proceedings of the Fifth international conference on Language Resources and Evaluation, LREC'06*, Genoa, Italy, May.

E. Bick. 2006. A constraint grammar-based parser for spanish. In *Proceedings of TIL 2006 - 4th Workshop on Information and Human Language Technology*, Ribeirão Preto, Brazil, October.

T. By. 2004. English dependency grammar. In G.M. Kruijff and D. Duchier, editors, *Proceedings of Workshop on Recent Advances in Dependency Grammar, CoLing-ACL'04*, pages 72–77, Genoa, Italy, August.

A. Ferrández, M Palomar, and L. Moreno. 2000. Slot unification grammar and anaphora resolution. In N. Nicolov and R. Mitkov, editors, *Recent Advances in Natural Language Processing II. Selected papers from RANLP 1997*, pages 155–166. John Benjamins Publishing Co., Amsterdam and Philadelphia.

A. Gelbukh, S. Torres, and H. Calvo. 2005. Transforming a constituency treebank into a dependency treebank. *Procesamiento del Lenguaje Natural*, 35:145–152, September.

T. Järvinen and P. Tapanainen. 1998. Towards an implementable dependency grammar. In Alain Polguere and Sylvain Kahane, editors, *Proceedings of Workshop on Processing of Dependence-Based Grammars, CoLing-ACL'98*, pages 1–10, Montreal, Canada, August.

D. Lin. 1998. Dependency-based evaluation of minipar. In *Workshop on the Evaluation of Parsing Systems, First International Conference on Language Resources and Evaluation, LREC'98*, Granada, Spain, May.

M. Marimón, N. Bel, S. Espeja, and N. Seghezzi. 2007. The spanish resource grammar: pre-processing strategy and lexical acquisition. In T. Baldwin, editor, *Proceedings of the Workshop on Deep Linguistic Processing, ACL'07*, pages 105–111, Prague, Czech Republic, June.

M.A. Martí, M. Taulé, M. Bertran, and L. Márquez. 2007. Ancora: Multilingual and multilevel annotated corpora. http://clic.ub.edu/ancora/ancora-corpus.pdf.

Joakim Nivre. 2006. *Inductive Dependency Parsing*, volume 34 of *Text, speech, and language technology series*. Springer, Dordrecht.

D. Sleator and D. Temperley. 1991. Parsing english with a link grammar. In *Third International Workshop on Parsing Technologies*, Tilburg, The Netherlands and Durbuy, Belgium.

ANUARI DE FILOLOGIA. ESTUDIS DE LINGÜÍSTICA (Anu.Filol. Estud. Lingüíst.)
1/2011, pp.101-134, ISSN: 2014-1408

CONSIDERACIONES SOBRE LA NATURALEZA DE LOS NÚCLEOS SINTÁCTICOS. HACIA UNA REPRESENTACIÓN SINTÁCTICA DE DEPENDENCIAS

MARINA LLOBERES
GRIAL, Universitat Oberta de Catalunya
marina.lloberes@ub.edu
IRENE CASTELLÓN
GRIAL, Universitat de Barcelona
icastellon@ub.edu

RESUMEN

En el análisis sintáctico automático, la definición de criterios lingüísticos para gramáticas basadas en conocimiento lingüístico permite de desarrollar recursos coherentes y consistentes. La construcción de EsTxala y CaTxala, dos gramáticas de dependencias del español y del catalán para FreeLing (un entorno de herramientas de Procesamiento del Lenguaje Natural), se ha llevado a cabo según el diseño previo de un repertorio de criterios ecléctico y crítico en relación con algunos de los formalismos lingüísticos implementados en el análisis automático del lenguaje, la Gramática de Dependencias y la Gramática Generativa. El objetivo de dicho repertorio es facilitar la coherencia y la consistencia de la representación sintáctica en el desarrollo de gramáticas para el análisis sintáctico automático.

PALABRAS CLAVE: análisis sintáctico automático, criterios sintácticos, Gramática de Dependencias, Gramática Generativa, español, catalán.

CONSIDERATIONS ABOUT THE NATURE OF SYNTACTIC NUCLEI. TOWARDS A SYNTACTIC REPRESENTATION OF DEPENDENCIES

ABSTRACT

The task about defining linguistic criteria for parsing linguistic-based grammars allows to build coherent resources. The development of EsTxala and CaTxala, Spanish and Catalan dependency grammars for FreeLing environment (a set of Natural Language Processing tools), was carried out regarding a set of linguistic criteria previously designed, which was developed like an eclectic and critic resource by means of the main linguistic formalisms implemented in parsing, the Dependency Grammar and the Generative Grammar. The main aim of this repertoire is to facilitate the coherence and the consistency of syntactic analysis into the development task of parsing grammars.

KEYWORDS: parsing, syntactic criteria, Dependency Grammar, Generative Grammar, Spanish, Catalan.

1. INTRODUCCIÓN

El análisis sintáctico automático es una tarea del Procesamiento del Lenguaje Natural (PLN) necesaria en los procesos de comprensión del lenguaje humano. Muchas aplicaciones de PLN (traducción automática, extracción de información, anotación automática de roles semánticos, etc.) requieren de una cierta profundidad en los análisis sintácticos para poder obtener una buena representación semántica sobre la que aplicar procesos posteriores. En consecuencia, durante los últimos años, este ámbito del PLN ha hecho grandes avances.

Dada la dificultad de asignar el árbol sintáctico adecuado a una oración, en los últimos años se han propuesto soluciones diversas a esta tarea. Una de las soluciones adoptadas más consensuadas por la comunidad investigadora ha sido dividir el proceso de análisis en dos etapas. En primer lugar, se aplica un analizador parcial que agrupa unidades léxicas de forma unívoca creando constituyentes de primer nivel, en algunas ocasiones constituyentes no completos. Este proceso es lo que llamamos análisis sintáctico parcial (Schmid 1994; Castellón et al. 1998). En segundo lugar, estos análisis parciales son conectados mediante un análisis completo que, en el área de PLN, denominamos análisis profundo. La primera etapa, el análisis parcial, parece que ha alcanzado índices de éxito suficientes, por lo que actualmente los esfuerzos están puestos en la segunda etapa, el análisis profundo.

Este segundo nivel se ha afrontado de diferentes maneras. Existen multitud de sistemas de análisis basados en conocimiento estadístico, sistemas que requieren de un corpus anotado manualmente para poder adquirir modelos de lenguaje para posteriormente aplicarlos a la resolución de las dependencias sintácticas (Magerman 1995; Collins 2000; Yamada y Matsumoto 2003; Nivre 2006; Koo et al. 2008). Aunque actualmente este tipo de sistemas es el mejor en la evaluación estándar, algunos de los problemas que presentan estos sistemas son la necesidad y dependencia de grandes corpus anotados manualmente en este nivel para poder aplicar mecanismos de aprendizaje y la poca transportabilidad, es decir, aplicar un modelo del lenguaje adquirido sobre otros corpus no muy similares al inicial produce porcentajes de error más altos (Comelles et al. 2010). Otra solución consiste en el desarrollo de gramáticas basadas en conocimiento lingüístico. Se trata de definir reglas sintácticas manualmente, basadas en corpus y en gramáticas descriptivas de las lenguas, para luego aplicarlas a corpus generales de la lengua.

En esta última línea se sitúa el trabajo que presentamos, el desarrollo de gramáticas de las lenguas española y catalana (Lloberes et al. 2010): EsTxala y

CaTxala.[1] Estas gramáticas, basadas en el formalismo de dependencias (Tèsnière 1959; Mel'čuk 1988), han sido diseñadas para el análisis sintáctico automático y, en concreto, en el entorno de la librería de herramientas de PLN FreeLing (Padró et al. 2010).[2]

En la actualidad, existen diversas gramáticas del inglés, como MaltParser (Nivre 2006), Minipar (Lin 1998), Connexor (Tapanainen y Järvinen 1998) o Link Grammar (Sleator y Temperley 1991). Sin embargo, en otras lenguas el desarrollo de este tipo de recursos es menor. En español, por ejemplo, existen algunas gramáticas como HISPAL (Bick 2006), Slot Unification Grammar (Ferrández et al. 2000) o Spanish Resource Grammar en HPSG (Marimón 2010) y, en el caso del catalán, existen gramáticas superficiales o parciales (Castellón et al. 1998) y, en algunos casos, con algún grado más de profundidad (Alsina et al. 2002).

EsTxala y CaTxala son recursos libres que proporcionan análisis sintácticos profundos y robustos. Ambas siguen el formalismo propuesto por el analizador de dependencias de FreeLing, TXALA (Atserias et al. 2005), que se basa en la escritura de reglas heurísticas y éstas, juntamente con el acceso a una serie de recursos lingüísticos externos, producen una estructura de análisis basado en la gramática de dependencias.

La gramática de dependencias tiene como origen los postulados del formalismo de dependencias propuesto en *Elements de syntaxe structurale* (Tèsnière 1959). Este formalismo (Tèsnière 1959; Mel'čuk 1988) concibe la oración como una red de conexiones entre las unidades léxicas basada en las relaciones que se establecen entre ellas. A diferencia del análisis de constituyentes (Chomsky 1957), donde los nodos terminales se proyectan en unidades más complejas o constituyentes para expresar dichas relaciones, las relaciones de dependencia se establecen entre los mismos nodos terminales, es decir, las palabras.

Así, todas las formas (palabras) o bien dependen de otras o bien otras dependen de ellas, aplicando el concepto de núcleo estructural. Además, en este formalismo, las relaciones de dependencia se etiquetan con las funciones sintácticas. Aunque el formalismo de dependencias clásico determina que para cada nodo dependiente existe un único núcleo sintáctico, hay autores (De Marneffe et al. 2006) que proponen un análisis en que los nodos dependientes pueden aceptar dos núcleos sintácticos.

[1] ExTxala y CaTxala se han desarrollado en el marco de los proyectos KNOW (Ministerio de Educación y Ciencia, TIN2006-1549-C03-02) y KNOW2 (Ministerio de Ciencia e Innovación, TIN2009-14715-C04-03, TIN2009-14715-C04-04), y también han sido utilizadas en los proyectos OpenTrad y EuroOpenTrad (Ministerio de Industria, Turismo y Comercio, Programa PROFIT, FIT-350401-2006-5), dos proyectos que tienen como objetivo desarrollar traductores basados en la transferencia para las lenguas oficiales del Estado Español (español, catalán, gallego y vasco) y para el inglés.

[2] http://nlp.lsi.upc.edu/freeling/

La representación propuesta por el formalismo de dependencias, aunque se sitúa en la sintaxis, es una representación muy próxima a la estructura semántica del predicado (Mel'čuk 1988). Es por ello por lo que el análisis de dependencias es el formalismo idóneo para hacer un análisis profundo y completo cercano a la representación de su significado. Esto ha producido que en las últimas décadas, en el área del PLN, haya crecido el interés en el desarrollo de sistemas de análisis sintáctico automático basado en las dependencias (Tapanainen y Järvinen 1998; Collins 2000; De Marneffe et al. 2006; Nivre 2006) con el objetivo de mejorar el nivel sintáctico en PLN. En español y en catalán existen muy pocos analizadores basados en este formalismo, como DILUCT (Calvo y Gelbukh 2006), DepPattern (Gamallo y González 2009), MaltParser (Nivre 2006) y Connexor (Tapanainen y Järvinen 1998).

Para la creación de una gramática computacional es fundamental disponer de un corpus representativo de la lengua para poder disponer de datos empíricos, ya que las muestras de lenguaje real contienen muchos usos no siempre tratados en las gramáticas descriptivas. Además, otro de los aspectos que creemos fundamental es la especificación de los criterios desarrollados en la elaboración de las reglas. Los criterios deben especificar tanto las condiciones de agrupación de constituyentes como la determinación del núcleo de los grupos formados. Como hemos dicho, en general asumimos la teoría expuesta en *Elements de syntaxe structurale* (Tesnière 1959) en el marco de la Sintaxis Estructural. Sin embargo, en algunas ocasiones, tomamos soluciones más eclécticas, ello es debido a la priorización de la proximidad de la representación semántica en nuestra propuesta. No pretendemos la demostración de la viabilidad de una teoría, sino más bien la construcción de un sistema que funcione y sirva de puente entre el análisis morfológico de las oraciones y su representación semántica.

En esta línea, en este artículo vamos a centrarnos en esta tarea: la elaboración de un repertorio de criterios lingüísticos para nuestras gramáticas que resuelva los puntos críticos en cuanto a la definición de núcleos sintácticos

La especificación de cómo analizar las diferentes estructuras ha implicado la creación en paralelo de un repertorio de estructuras que finalmente proponemos como corpus de evaluación cualitativa para cualquier gramática de las lenguas tratadas. Así, en este artículo presentamos los puntos críticos de los criterios elaborados, ilustrándolos con muestras de dicho corpus de evaluación.

La estructura del artículo será la siguiente. En §2, trataremos la naturaleza de los núcleos sintácticos desde dos perspectivas, el análisis de constituyentes a través de la propuesta *Goverment and Binding* (Chomsky 1981) y el análisis de dependencias, que tiene como origen los postulados propuestos por Tesnière (1959) y Mel'čuk (1988) dentro de la Sintaxis Estructural, para en último lugar explicar el punto de vista adoptado en las gramáticas de Txala. En las secciones posteriores nos centraremos en los criterios de algunos aspectos en los cuales no hay acuerdo entre estos formalismos lingüísticos (§3): el papel de la preposición

(§3.1), las oraciones subordinadas (§3.2), las estructuras comparativas (§3.3) y la representación de las oraciones coordinadas (§3.4). Por último, presentaremos las conclusiones del trabajo (§4).

2. SOBRE LA NATURALEZA DE LOS NÚCLEOS SINTÁCTICOS

En el análisis sintáctico automático, se han implementado diferentes formalismos gramaticales (*Phrase Structure Grammar*, *Dependency Grammar*, *Head-Driven Phrase Structure Grammar*, *Link Grammar*, etc.). No obstante, entre estos formalismos lingüísticos, los dos principales a partir de los cuales se han desarrollado muchos de los analizadores sintácticos automáticos son: *Phrase Structure Grammar* (Chomsky 1957) y *Dependency Grammar* (Mel'čuk 1988). El formalismo *Phrase Structure-Grammar*, que llamamos análisis de constituyentes, se basa en la Gramática Generativa (GG). Por otro lado, a partir del formalismo *Dependency Grammar* (Mel'čuk 1988), que traducimos como Gramática de Dependencias (GD), se ha desarrollado el análisis de dependencias, de base estructuralista (Tesnière 1959).

Ambos formalismos parten de la idea de que las unidades léxicas contienen información sintáctica y semántica o, en términos más específicos, subespecificaciones sintácticas y semánticas que determinan la configuración de la oración. No obstante, como se observará, la GG y la GD difieren en la noción de categoría léxica, cosa que conlleva soluciones diferentes en cuanto a las categorías léxicas que pueden funcionar como núcleo sintáctico dentro de la estructura sintáctica.

2.1. La Gramática Generativa

El análisis de constituyentes parte de la base de que el lenguaje es una estructura abstracta de signos en forma de jerarquía. A partir de las estructuras más simples, que se corresponden a las unidades del lenguaje, se forman estructuras más complejas, que tienen como último término la unidad superior de la jerarquía que reúne todos los signos de la estructura, la oración.

El hecho de que las estructuras más simples permitan crear estructuras más complejas es debido a la combinación de las unidades atómicas del lenguaje. Puesto que el lenguaje se presenta de manera inherente en un orden lineal, las operaciones de combinación no se llevan a cabo aleatoriamente, sino que están motivadas por esa linealidad. Estas operaciones son en su naturaleza reglas de reescritura que definen la estructura sintáctica (1), conocidas también como reglas de estructura sintagmática.

(1)

i.$p \rightarrow p_1\ p_2\ p_3\ \dots\ p_n$
ii. $u \rightarrow u_1\ u_2\ u_3\ \dots\ u_n$
iii. $w \rightarrow p\ \ u$

Estas reglas establecen que existe un conjunto de categorías básicas (nombre, verbo, adjetivo, etc.), las categorías terminales, que definen el lexicón (2ii). Los terminales, pues, son la unidad sintáctica básica a partir de la cual se construyen las relaciones en el marco de la oración mediante expansiones o, en términos generativistas, proyecciones (2iii). Estas proyecciones tienen como finalidad agrupar los terminales en unidades sintácticas no terminales más complejas, los constituyentes, que ponen de manifiesto la estructura sintáctica de la oración. Los elementos que determinan la organización de la oración son los núcleos sintácticos. Por esa razón, las expansiones siguen el camino de cada núcleo y especifican el tipo de constituyente (Fig. 1).

(2)

i. La noia menja
'La chica come'
ii. Terminales
DT = { La }
N = { noia }
V = { menja }
iii. Proyecciones
SN → DT N
SV → V
O → SN SV

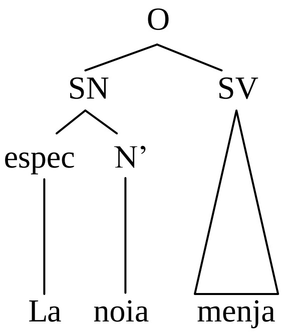


Fig. 1. Representación de la estructura sintáctica según la GG del ejemplo (2i) 'La noia menja' ('La chica come')

En la oración (2i) representada sintácticamente en Fig. 1, las tres categorías expresadas (DT, N, V) se proyectan en unidades más abstractas mediante las reglas de estructura sintagmática (2iii). De modo que el determinante 'la' y el nombre 'noia' se proyectan como *espec* y *N'* respectivamente, y, a su vez, ambas proyecciones se expanden una vez más debido a la existencia de una regla que define la anidación de *espec* y *N'* como constituyentes del sintagma nominal (SN → DT N). Por otro lado, el verbo 'menja' se proyecta como *V*, *V'*, etc., y, en último término, como *SV*. Finalmente, la regla que define la formación de la oración (O → SN SV) permite anidar el sintagma nominal *SN*, expresado léxicamente mediante 'la noia', y el sintagma verbal *SV*, representado léxicamente por 'menja'.

2.2. La Gramática de Dependencias

En el formalismo de dependencias (Tèsnière, 1959; Mel'čuk, 1988), se considera que las unidades léxicas están conectadas porque entre ellas se establecen relaciones. En otras palabras, cada unidad léxica de una oración está relacionada sintácticamente con otra unidad léxica y no hay ninguna unidad léxica que quede fuera del alcance de la oración. Por lo tanto, la estructura sintáctica es un grafo de conexiones que pone de manifiesto las relaciones entre las unidades léxicas.

Mientras que en la GG las unidades léxicas se proyectan en nodos abstractos para construir la estructura sintáctica, desde la perspectiva de la GD, las relaciones entre las unidades léxicas son directas, es decir, cada unidad léxica está gobernada directamente por otra unidad léxica (3ii). De modo que en una relación sintáctica la unidad léxica dominante es el núcleo sintáctico de la unidad léxica dependiente. Así pues, la estructura sintáctica es un grafo de conexiones directas y jerárquicas (Fig. 2).

(3)

i. La noia menja
'La chica come'

ii. menja ← noia
noia ← La

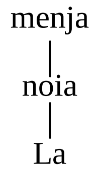


Fig. 2. Representación de la estructura sintáctica según la GD del ejemplo (3i) 'La noia menja' ('La chica come')

El hecho de que las unidades léxicas se conecten alrededor de una jerarquía permite que un núcleo sintáctico actúe, a su vez, como un nodo dependiente de otro nodo que se encuentra en un nivel superior de la jerarquía. Del mismo modo, un nodo dependiente puede ser el núcleo sintáctico de los nodos que dependen de éste. De todas formas, el nodo dependiente siempre está asociado a un solo núcleo sintáctico (Fig. 3) y, en cambio, un núcleo sintáctico puede dominar diversos nodos dependientes, a excepción de un caso. El núcleo sintáctico de la oración no está gobernado por ningún otro núcleo sintáctico. Así pues, la estructura sintáctica es un grafo acíclico de conexiones directas y jerárquicas (Fig. 3).

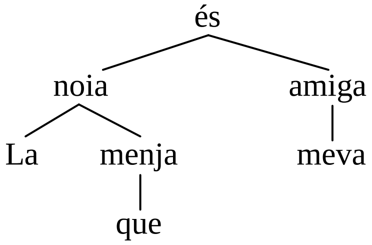


Fig. 3. Representación de la estructura sintáctica según la GD de la oración 'La noia que menja és amiga meva' ('La chica que come es amiga mía')

A pesar de que Tesnière (1959) y Mel'čuk (1988) consideran que no es posible que un nodo dependiente pueda estar gobernado por dos núcleos sintácticos, hay autores que defienden la duplicidad de núcleos sintácticos (De Marneffe et al. 2006) y, por lo tanto, consideran lícita la existencia tanto del análisis representado en Fig. 3 como del análisis ilustrado en Fig. 4.

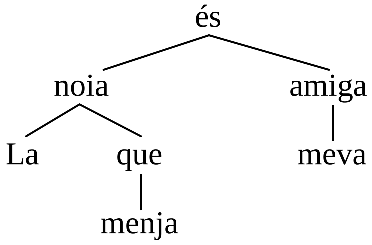


Fig. 4. Representación de la estructura sintáctica según la GD de la oración 'La noia que menja és amiga meva' ('La chica que come es amiga mía')

En cuanto a la determinación de las relaciones sintácticas, en ningún caso es aleatoria ya que está motivada por la semántica. Cada unidad léxica contiene subespecificaciones que distinguen las relaciones sintácticas entre las unidades léxicas (4).

(4)

i.La profesora lleva a los alumnos al teatro
ii. La profesora lleva a los alumnos lecturas nuevas

Por lo tanto, no es suficiente la estructuración de las unidades léxicas en el marco de la oración para explicar las relaciones de la oración. La naturaleza de las relaciones donde participan las unidades léxicas da sentido al conjunto de relaciones de la estructura sintáctica. Por ese motivo, en los ejemplos de (4), a pesar de que el verbo sea el mismo en ambos ejemplos ('llevar'), la relación establecida entre 'llevar' y 'a los alumnos' es diferente en cada caso.

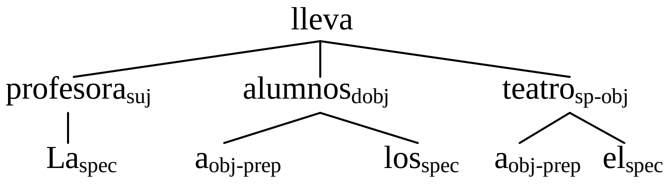


Fig. 5. Análisis de funciones sintácticas según la GD
del ejemplo (4i) 'La profesora lleva a los alumnos al teatros'

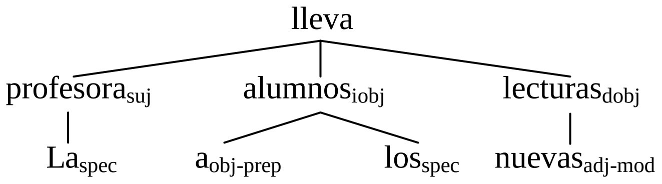


Fig. 6. Análisis de funciones sintácticas según la GD
del ejemplo (4ii) 'La profesora lleva a los alumnos lecturas nuevas'

En (4i), aparece un objeto directo y, en (4ii), la relación establecida es la de objeto indirecto. De modo que las relaciones sintácticas aportan más información a la organización de las unidades léxicas dentro de la estructura sintáctica (Fig. 5, Fig. 6). En realidad, éstas actúan como un enlace entre la prosodia y la morfología, y el significado.

2.3. Las categorías léxicas y los núcleos sintácticos

Implícitamente, en los apartados anteriores (§2.1, §2.2), se ha dejado entrever que la estructuración de las unidades léxicas dentro de la jerarquía no ocurre de forma aleatoria, ya que los constituyentes que las representan son distintos (Hernanz y Brucart 1987:31).

Los constituyentes se clasifican en distintas categorías sintácticas y dicha clasificación está determinada por la naturaleza de su núcleo sintáctico (Hernanz 2002:1019). Por ejemplo, en el sintagma *la revista cultural*, se habla de sintagma nominal ya que el núcleo de este sintagma se corresponde a un nombre; a su vez, el sintagma *leen muchos libros* es categorizado como sintagma verbal porque el núcleo sintáctico está formado por un verbo.

El hecho de que se pueda distinguir, por ejemplo, un nombre de un verbo es debido al conjunto de propiedades que definen el repertorio de categorías léxicas. Estas categorías pueden ser flexionadas, forman repertorios abiertos, son semánticamente plenas y son categorías mayores (ya que pueden seleccionar complementos y son independientes morfológicamente).

Por otro lado, existe otro conjunto de categorías, las categorías gramaticales, que Trask (1992:122) define como:

> **gramatical category** *n*. Any of various distinctions within which are expressed by variations in the form of lexical or phrasal constituents. […]

Tal como describe Hernanz (2002:1021), estas categorías se diferencian de las categorías léxicas debido a que pertenecen a repertorios cerrados, generalmente no tienen independencia morfológica ni fonológica y son de carácter relacional (como consecuencia, no contribuyen de la misma manera a la interpretación semántica de la oración).

A pesar de que intuitivamente se puedan reconocer clases distintas de palabras, no hay acuerdo alguno sobre cuáles son las categorías que forman las clases de palabras (Hernanz y Brucart 1987:32). Las diferentes clasificaciones que se han propuesto coinciden a reconocer ciertas clases, como el nombre o el verbo, pero existen divergencias para admitir como clases de palabras el resto (como el adjetivo, la preposición, el adverbio, la conjunción).

Según la perspectiva de la GG, la naturaleza de la unidad léxica determina la categoría sintáctica del núcleo sintáctico (Chomsky 1981; Haegeman 1991:30). A diferencia de los otros formalismos gramaticales, la GG considera que todas las categorías léxicas son endocéntricas (es decir, categorías que contienen un núcleo y que son de igual distribución) y, como consecuencia, tienen la capacidad para actuar como núcleo sintáctico (Hernanz y Brucart 1987: 36).

En otro estadio diferente de la GG, se sitúa la GD (Tèsnière 1959; Mel'čuk 1988). Este formalismo reconoce la existencia de unidades léxicas (o 'palabras plenas' según la GD; Tesnière, 1959: 53), que están dotadas de función semántica porque por ellas mismas representan una idea. En contraposición, otro grupo de palabras forman la clase de unidades funcionales o, tal y como Tesnière (1959: 53) las denomina, 'palabras vacías'. Se trata de unidades sin carga semántica que actúan como meros mecanismos gramaticales o de relación de unidades léxicas (Tesnière 1959: 53). Según este punto de vista, las unidades léxicas son aquellas que pueden optar a funcionar como núcleos sintácticos y, en cambio, las unidades funcionales actúan alrededor de esos núcleos (por lo tanto, no ocupan posiciones nucleares en la estructura sintáctica).

3. UNA PROPUESTA DE REPRESENTACIÓN DE LOS NÚCLEOS SINTÁCTICOS

EsTxala y CaTxala (Lloberes et al. 2010) son dos gramáticas para el análisis sintáctico profundo del español y catalán que utilizan información lingüística como fuente de conocimiento y están basadas en la teoría de la GD (Tèsnière 1959; Mel'čuk 1988). De modo que las reglas que forman ambas gramáticas se fundamentan en los postulados del formalismo de dependencias.

No obstante, estas gramáticas son críticas en la consideración de la naturaleza de algunos núcleos sintácticos, ya que el objetivo es aportar análisis sintácticos donde la representación de la estructura sintáctica sea próxima a estructura semántica del predicado (Mel'čuk 1988). Como consecuencia, proponen una alternativa de representación de algunas construcciones sintácticas. Esta propuesta combina criterios basados en la GD (Tèsnière 1959; Mel'čuk 1988) y la GG (Chomsky 1957). Por lo tanto, los criterios de representación aquí propuestos pretenden ser una aportación ecléctica y útil para otras gramáticas desarrolladas en el mismo ámbito del PLN.

A continuación, presentamos una revisión crítica de las estructuras sintácticas que comportan una representación sintáctica diferente en función de la teoría sintáctica, GG y GD. En concreto, se describen esas estructuras sintácticas y se exponen las razones que llevan a la GG y la GD a proponer análisis divergentes. Como consecuencia, el análisis de las propuestas teóricas permite optar por una de las soluciones y definir un criterio para EsTxala y CaTxala.

3.1. La preposición como unidad con significado pleno

La naturaleza sintáctica de la preposición ha sido protagonista de abundantes dialécticas teóricas. A lo largo de la historia, los estudios lingüísticos y filológicos se han centrado en dilucidar el comportamiento de la preposición respecto al resto de categorías léxicas (Fabra 1918; Badia i Margarit 1962; Alarcos 1994; Bonet y Solà 1986; Hernanz 2002). Mientras que tradicionalmente el nombre, el adjetivo y el verbo han sido consideradas categorías léxicas como tales, los teóricos discrepan a la hora de determinar si la preposición pertenece realmente o no a esas categorías.

La gramática tradicional (Fabra 1918; Badia i Margarit 1962; Alarcos 1994) propone una noción de la preposición, con frecuencia, vaga. A veces, incluso, algunos usos de la preposición se engloban en los usos de los adverbios o de expresiones adverbiales (Moreno Cabrera 1991:391). De modo que, desde un punto de vista tradicional, existe cierta confusión en la distinción entre preposición y adverbio (Tesnière 1959:52; Bonet y Solà 1986:64). Prueba de ello es la clasificación de la preposición en usos transitivos (5i), donde la preposición no puede aparecer sin la presencia de un complemento, y en usos intransitivos (5ii), en los cuales el complemento de la preposición es opcional (Sancho Cremades 2002:1693) y que la gramática tradicional denomina adverbios utilizados como preposiciones (Bonet y Solà 1986:64).

(5)

i.El pasillo conduce al comedor
ii. Encontrarás las llaves { encima de la mesa / encima }

En cambio, estudios teóricos y descriptivos más recientes apuntan que la preposición, aunque mantiene un comportamiento distinto a las categorías léxicas mayores, se pueden reconocer comportamientos afines a esas categorías (Bonet y Solà 1986; Hernanz y Brucart 1987; Hernanz 2002).

A diferencia de las categorías léxicas mayores, que, como se ha observado en §2.3, son variables, pertenecen a clases de palabras abiertas, su contenido es descriptivo, seleccionan complementos y son independientes morfológicamente (Hernanz 2002:1019), la preposición muestra un comportamiento distinto ya que es invariable, su significado es más abstracto que las categorías léxicas mayores, y no puede ocurrir sin un complemento (6).

(6)

i. Laia escucha [$_{\text{SN}}$ música]
ii. Laia parece [$_{\text{SAdj}}$ triste]
iii. Laia [$_{\text{SV}}$ ríe]
iv. Laia nada [$_{\text{SAdv}}$ bien]
v. *Laia confía [$_{\text{SP}}$ en]

De modo que el sintagma nominal, el sintagma adjetival, el sintagma verbal y el sintagma adverbial son, en realidad, sintagmas endocéntricos, es decir, sintagmas que se forman sobre la base de un núcleo (6i-iv). En contraposición, según la Sintaxis Estructural, el sintagma preposicional es de naturaleza exocéntrica ya que carece de núcleo como la oración.

No obstante, puesto que la GG considera que todas las proyecciones sintácticas son endocéntricas (Chomsky 1957; Bonet y Solà 1986; Hernanz y Brucart 1987), la preposición también presenta esa característica, pero a diferencia de las otras categorías léxicas, subcategoriza un complemento (Bonet y Solà 1986:64; Hernanz y Brucart 1987:36); de ahí que la oración de (6v) sea agramatical.

La hipótesis generativa de la endocentricidad de la preposición se confirma por el comportamiento paralelo de algunos verbos transitivos (7), que sólo aceptan la diátesis transitiva. El hecho de que (7ii) sea una estructura agramatical no quiere decir que el verbo sea una categoría exocéntrica. En tal caso, no se podría reconocer el verbo como categoría léxica (Hernanz y Brucart 1987:36). El verbo es una clase de palabra considerada categoría léxica, ya que el sintagma verbal siempre está dotado de núcleo. Entonces, el factor que determina la agramaticalidad de (7ii) reside en las propiedades sintáctico-semánticas de ese verbo: el verbo ‘saber’ prevé en su estructura un argumento que funciona como objeto directo. Por lo tanto, la ausencia de dicho argumento causa la agramaticalidad de la construcción.

(7)

i. Supe que se había ido
ii. *Supe

Debido a la caracterización de la preposición como categoría endocéntrica que propone la GG (Chomsky 1981:48; Bonet y Solà 1986:65; Hernanz y Brucart 1987:36), esta categoría admite proyecciones sintácticas de la misma forma que el resto de categorías léxicas, mediante lo cual la preposición tiene la capacidad de actuar como núcleo sintáctico (Fig. 7).

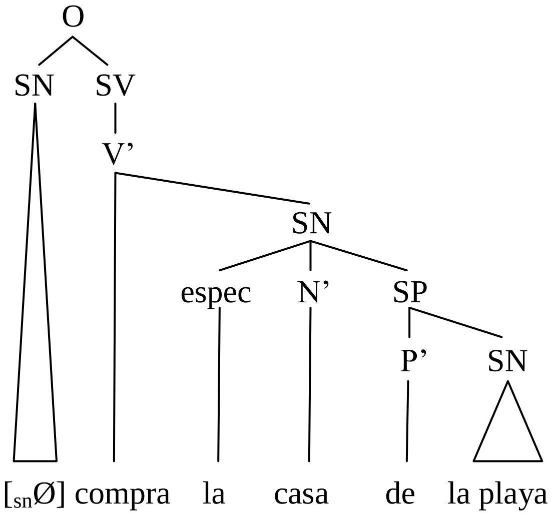


Fig. 7. Análisis del sintagma preposicional según la GG de la oración 'Compra la casa de la playa'

De manera similar a la gramática tradicional, la GG (Chomsky 1981; Bonet y Solà 1986; Hernanz y Brucart 1987) y la GD (Tesnière 1959; Mel'čuk 2003) coinciden en la consideración de la preposición como una pieza léxica relacional.

No obstante, la GD sitúa la preposición alejada de las categorías como el nombre, el adjetivo y el verbo (Tesnière 1959:53). Según la GD, la preposición es una categoría que no está cargada de ninguna función semántica y, por esa razón, se clasifica como una unidad funcional, aquellas unidades que únicamente aparecen en el discurso para indicar, precisar y transformar la categoría de las unidades léxicas y determinar las relaciones existentes entre ellas. Como consecuencia, la preposición no puede actuar como núcleo sintáctico (Fig. 8).

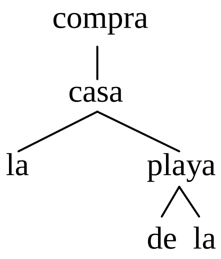


Fig. 8. Análisis del sintagma preposicional según la GD de la oración 'Compra la casa de la playa'

Puesto que EsTxala i CaTxala son gramáticas basadas en el formalismo de la GD, siguen sus postulados. No obstante, en cuanto a la naturaleza de la preposición, a diferencia de la GD (Tesnière 1959:53), se considera que es una pieza gramatical con función semántica, aunque mantiene diferencias con el resto de categorías léxicas mayores ya que funciona como un elemento relacional. La evidencia de la carga semántica de la preposición se halla en la posibilidad de distinguir preposiciones con significado propio y preposiciones con carácter semánticamente vacío (Hernanz y Brucart 1987:263). El primer tipo hace referencia a las preposiciones que introducen circunstanciales, un locativo en (8i), y el segundo se refiere a las preposiciones regidas por el verbo del predicado, un tema en (8ii).

(8)

i. Los excursionistas andan por el campo
ii. La empresa apuesta por inversiones en el extranjero

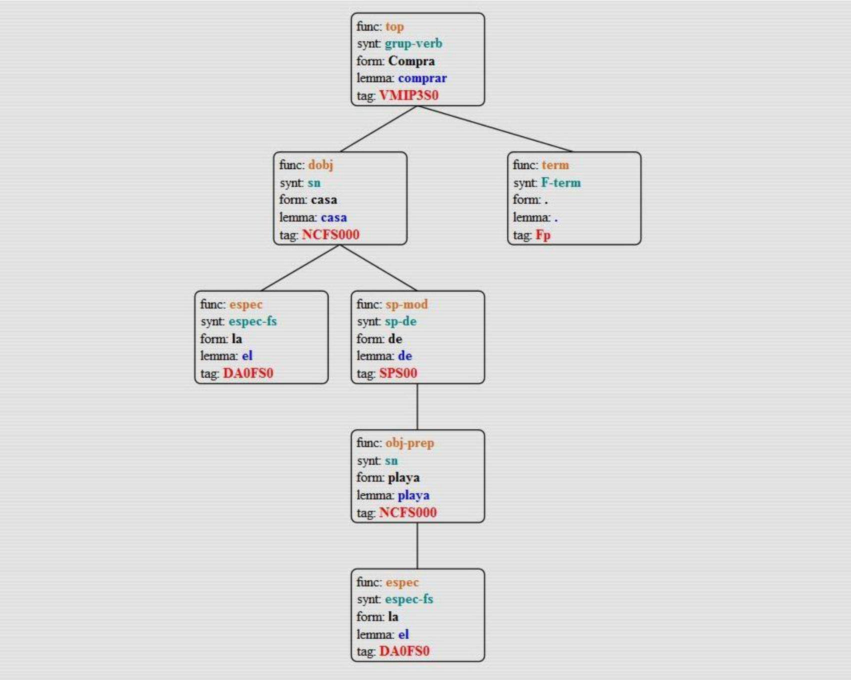


Fig. 9. Análisis del sintagma preposicional en EsTxala
de la oración 'Compra la casa de la playa'

Igualmente, la posibilidad de que el verbo del predicado subcategorice un argumento introducido mediante una determinada preposición, como se observa en el ejemplo (8ii), es una muestra más de la relevancia de la preposición como categoría léxica. Por lo tanto, en las gramáticas de Txala, la preposición, junto con el resto de categorías léxicas, funciona como una unidad léxica con capacidad de actuar como núcleo en la estructura sintáctica (Fig. 9).

3.2. La variabilidad estructural de la subordinadas

Las subordinadas con verbo en forma personal o subordinadas en tiempo finito aparecen en la oración como argumento o adjunto de otro constituyente oracional modificándolo o complementándolo (Villalba 2002:2291). De todas formas, a diferencia de los constituyentes sintagmáticos, dicha relación se lleva a cabo mediante un operador gramatical (9) que expresa la relación entre ambas oraciones, principal y subordinada (9i', 9ii').

(9)

i. Dijo que era verdad
i'. [Dijo] *que* [era verdad]
ii. Es van reunir on havien acordat
'Se reunieron donde habían acordado'
ii'. [Es van reunir] *on* [havien acordat]

Sin embargo, de la observación de los ejemplos de (9), se perciben numerosas diferencias en el tipo de modificación o complementación. En (9i), la oración subordinada toma un valor sustantivo (Moreno Cabrera 1991:671; Delbecque y Lamiroy 1999:1967; Bonet 2002:2321), de ahí que se las conozca como subordinadas sustantivas. Como consecuencia, puede ser sustituida por un nombre sin que afecte a la gramaticalidad de la construcción (*Dijo la verdad*). En cambio, en (9ii), el valor de la subordinada es de naturaleza distinta, tiene un valor adverbial (Moreno Cabrera 1991:671; Villalba 2002:2251), y, por esa razón, se las denomina como subordinadas adverbiales. Por ello, un adverbio puede aparecer en su lugar (p.e. *Es van reunir allà* 'Se reunieron allí').

La modificación o complementación puede establecerse en otros niveles diferentes de la oración, por ejemplo a nivel de nombre (p.e. *No le gusta la idea de que vengan, La possibilitat que tanquin l'empresa és real* 'La posibilidad que cierren la empresa es real') o de adjetivo (p.e. *Está segura de que ha perdido las llaves, És una dona desitjosa que els seus fills tornin aviat* 'Es una mujer deseosa de que sus hijos vuelvan pronto'). En ese contexto, existen subordinadas en tiempo finito que típicamente modifican el nombre y que funcionan de manera similar a los adjetivos, las oraciones relativas (Moreno Cabrera 1991: 671; Brucart 1999: 397; Solà 2002: 2455; Villalba 2002: 2251). Este tipo de cláusulas tienen la particularidad de que el operador gramatical que las introduce es a la vez entendido como un nexo y como pronombre relativo anafórico (10).

(10)

i. La chica que viene gruñendo es mi prima
ii. La chica enfadada es mi prima

Por otro lado, en español y en catalán, es posible otro tipo de construcciones que están a medio camino entre las interrogativas indirectas y las relativas (Solà 2002: 2534). Aparecen como complemento o modificador del verbo principal de la oración y implícitamente expresan una interrogación, cosa que las asemeja al funcionamiento de las interrogativas indirectas (11i). No obstante, el operador gramatical que las introduce se corresponde al tipo de marcador típico de las relativas, pese a que en muchas ocasiones puede coincidir con el operador gramatical que introduce una interrogativa indirecta. Por esa razón, ese tipo de construcciones se conoce con el término pseudorelativas (11ii).

(11)

i. Ha preguntado con quien habíamos hablado
ii. Telefona a qui ha trobat al cinema
'Llama a quien ha encontrado en el cine'

La necesidad de la existencia de un operador gramatical que introduzca los diversos tipos de subordinadas finitas es una característica tanto del español como del catalán, a diferencia de otras lenguas como el inglés, que, con frecuencia, en determinados contextos permite la omisión de dicho marcador (p.e. *The book I was reading was quite interesting* 'El libro que estaba leyendo era bastante interesante'). No obstante, tal y como apunta Solà (2002:2509), la elisión del pronombre relativo había estado posible en otros estadios del catalán (12i). Por otro lado, son posibles estructuras con omisión del operador gramatical. En español, se pueden dar estos casos de omisión de conjunción (12ii) cuando la subordinada es expresada en modo subjuntivo y el sujeto de la oración principal no es expresado léxicamente (Delbecque y Lamiroy 1999:2026). Por el contrario, en catalán, parece ser una estructura muy restrictiva (12iii) o, más bien, extensamente condenada por la normativa (Bonet 2002:2348).

(12)

i.Compta de la faina Ø ha feta Mestre Juan Estelrich […] Dia 16 Janer 1799[3]
ii. Se ruega no toquen los animales
iii. +Li agraïm respecti el descans dels veïns
'Le agradecemos respete el descanso de los vecinos'

Evidentemente, los trabajos de lingüística descriptiva apuntan la existencia de otras construcciones relacionadas directamente o próximas a las subordinadas (Solà 1972; Bonet y Solà 1986; Bosque y Demonte 1999; Solà et al. 2002), como, por ejemplo, las interrogativas directas (*¿Quién ha dicho eso?*), las interrogativas indirectas (*No sap com tornarà a casa* 'No sabe cómo volverá a casa'), las subordinadas causales (*Va a perder el avión porque se ha dormido*), finales (*Esta*

[3] Ejemplo del Arxiu del Regne de Mallorca citado por Solà (2002:2509).

exposición está pensada para que el visitante conozca nuevas realidades sociales), concesivas (*Anirà de vacances, encara que ha de treballar* 'Irá de vacaciones, aunque tiene que trabajar') y condicionales (*Puede saltarse clases siempre que traiga un justificante*). En nuestra propuesta, no hacemos un análisis detallado sobre la representación con fines computacionales de estas construcciones, ya que se asemejan a las construcciones que acabamos de describir. No obstante, nos referiremos a algunas de ellas a lo largo de esta sección.

Desde la perspectiva representacional, surgen divergencias a la hora de proponer una representación de las cinco construcciones descritas previamente: subordinadas sustantivas (13i), subordinadas adverbiales (13ii), relativas (13iii), pseudorelativas (13iv) y subordinadas con elisión de marcador (13v); puesto que, en función de los postulados teóricos, los formalismos gramaticales y, en nuestro caso, la GG y la GD proponen distintas soluciones a la representación de esas construcciones. La discusión gira alrededor de qué categoría gramatical actúa como núcleo de la subordinada: la conjunción subordinante o el verbo subordinado.

(13)

i.El gerente ha decidido que los sueldos se congelen
ii. Ha arribat a casa quan el partit havia acabat
'Ha llegado a casa cuando el partido había terminado'
iii. La película que vio ayer le trajo recuerdos de su pasado
iv.Es preocupa per quanta gent hi haurà
'Se preocupa por cuanta gente habrá'
v. Rogamos vigilen sus pertenencias

3.2.1. Las subordinadas sustantivas y las subordinadas adverbiales

Dada la naturaleza lexicalista de la GG (Chomsky 1957; Bonet y Solà 1986; Haegeman 1991), la categoría gramatical que define una construcción subordinada es el complementante. En otros términos, el núcleo sintáctico de las subordinadas se corresponde a la conjunción tal y como se propone en el siguiente esquema derivado de las reglas de estructura sintagmática (Bonet y Solà 1986: 21; Haegeman 1991: 107). A pesar de eso, en función de la clase de subordinada o, más concretamente, en función del complementante, el esquema puede variar como veremos.

(14)

i. O → SN SV[4]
ii. O′ → COMP O[5]

El esquema (14) explica la formación de la estructura de las subordinadas sustantivas (Bonet y Solà 1986: 21; Haegeman 1991: 107). El complementante en esta clase de subordinadas aparece como un mero operador gramatical que vincula la subordinada sustantiva y la oración principal, e insiere la subordinada sustantiva dentro de la estructura de la oración principal. De modo que la conjunción tiene como única función de marcador de inicio o encabezamiento de cláusula, como se puede observar en el ejemplo (Fig. 10).

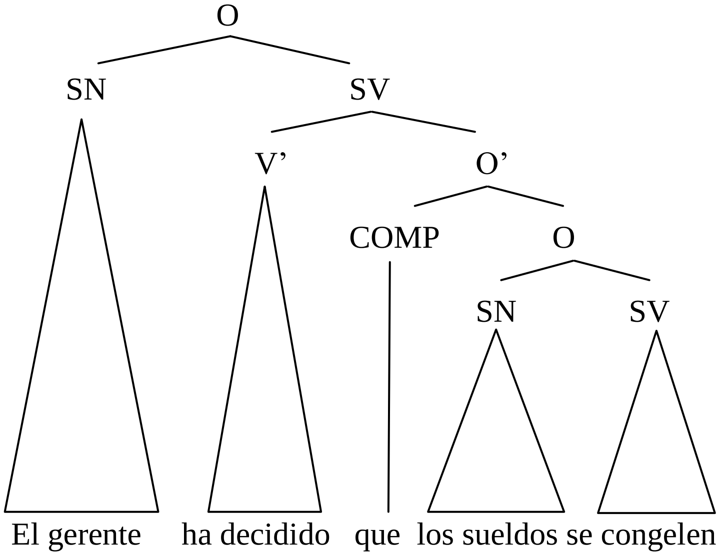


Fig. 10. Análisis de la subordinada sustantiva según la GG del ejemplo (13i), 'El gerente ha decidido que los sueldos se congelen'

[4] En realidad, el esquema original incluye la categoría terminal inflexión (INFL), que reúne aspectos como tiempo y concordancia (INFL ≡ [± Tiempo, (CONC)]), como nodo de la estructura (O → SN INFL SV), ya que la GG considera que esta categoría es exocéntrica. No obstante, en este trabajo, no la representamos gráficamente porque presentamos una representación simplificada de la GG para facilitar la lectura de los análisis.

Por otro lado, cabe la posibilidad de otra interpretación en la regla de formación de la oración relacionada con la opcionalidad del sintagma nominal sujeto, ya que, con frecuencia, hay lenguas que permiten esta opcionalidad (*Plou* 'Llueve'). Como consecuencia, Chomsky (1981) propone una regla del tipo: O → (SN) INFL SV.

[5] En realidad, la regla ilustrada en (14) es incompleta, ya que la GG considera que la categoría COMP es extensible a cualquier tipo de cláusula, ya esté o no esté inserida en otra cláusula (c). De modo que el esquema (14) debería ser representado de la siguiente manera:

O → SN SV
O′ → COMP O
O′′ → (TOP) O′

Paralelamente, las subordinadas adverbiales presentan el mismo comportamiento que las subordinadas sustantivas, es decir, el complementante, materializado léxicamente por una conjunción, opera como un marcador de la subordinada (Bonet y Solà 1986:21; Haegeman 1991:107), con lo cual es el núcleo sintáctico de la cláusula (Fig. 11).

Por otro lado, basándonos en el formalismo de la GD (Tesnière 1959; Mel'čuk 2003), la conjunción forma parte del conjunto de unidades funcionales intranucleares (Tesnière 1959: 82), es decir, categorías gramaticales sin carga semántica que actúan como simples instrumentos gramaticales. Como consecuencia, esta categoría no puede funcionar en ningún caso como núcleo sintáctico de una subordinada sustantiva (Fig. 12) o de una subordinada adverbial (Fig. 13). En realidad, la categoría que introduce estas cláusulas subordinadas es el mismo verbo subordinado, ya que el verbo sí es una unidad léxica con carga semántica capaz de actuar como núcleo sintáctico.

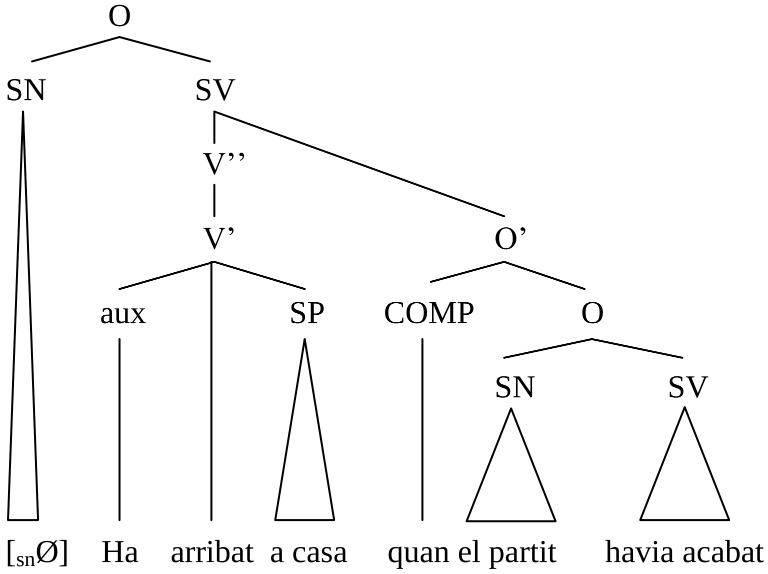


Fig. 11. Análisis de la subordinada adverbial según la GG del ejemplo (13ii), 'Ha arribat a casa quan el partit havia acabat' ('Ha llegado a casa cuando el partido había terminado)

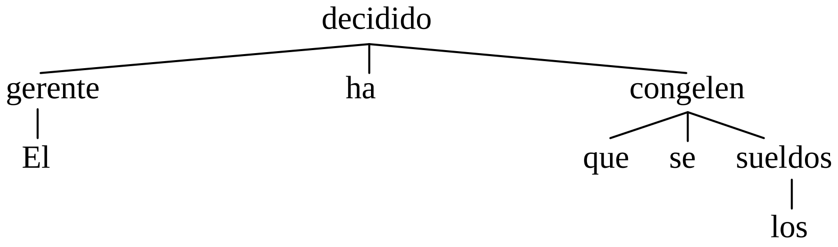


Fig. 12. Análisis de la subordinada sustantiva según la GD del ejemplo (13i), 'El gerente ha decidido que los sueldos se congelen'

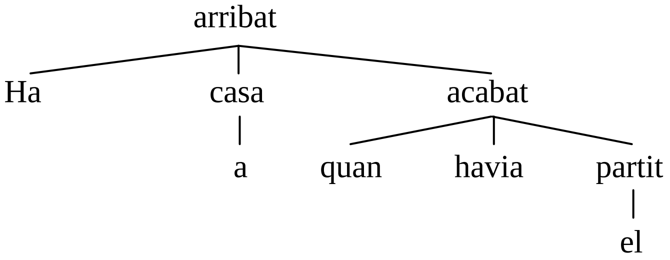


Fig. 13. Análisis de la subordinada adverbial según la GD del ejemplo (13ii), 'Ha arribat a casa quan el partit havia acabat' ('Ha llegado a casa cuando el partido había terminado)

A la hora de considerar cómo deben representarse ambas estructuras en las gramáticas de dependencias EsTxala y CaTxala, se parte de la idea de que la conjunción subordinante es un operador gramatical, de acuerdo con la GG (Chomsky 1957) y la GD (Tesnière 1959; Mel'čuk 2003). No obstante, a diferencia de la GD, en las gramáticas de Txala, se establece como la categoría sintáctica nuclear de la cláusula (Fig. 14, Fig. 15).

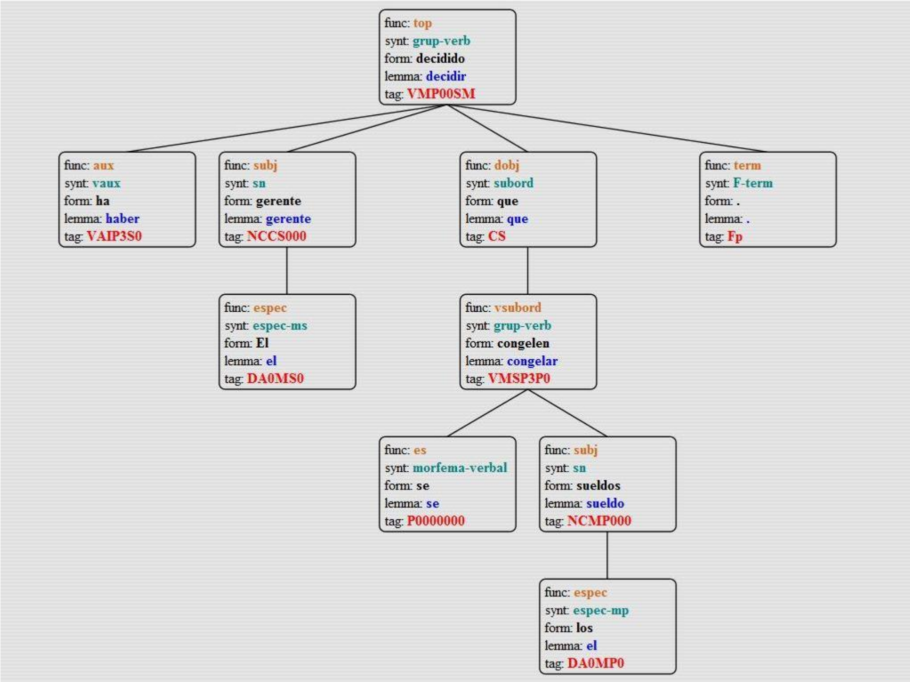


Fig. 14. Análisis de la subordinada sustantiva en EsTxala del ejemplo (13i), 'El gerente ha decidido que los sueldos se congelen'

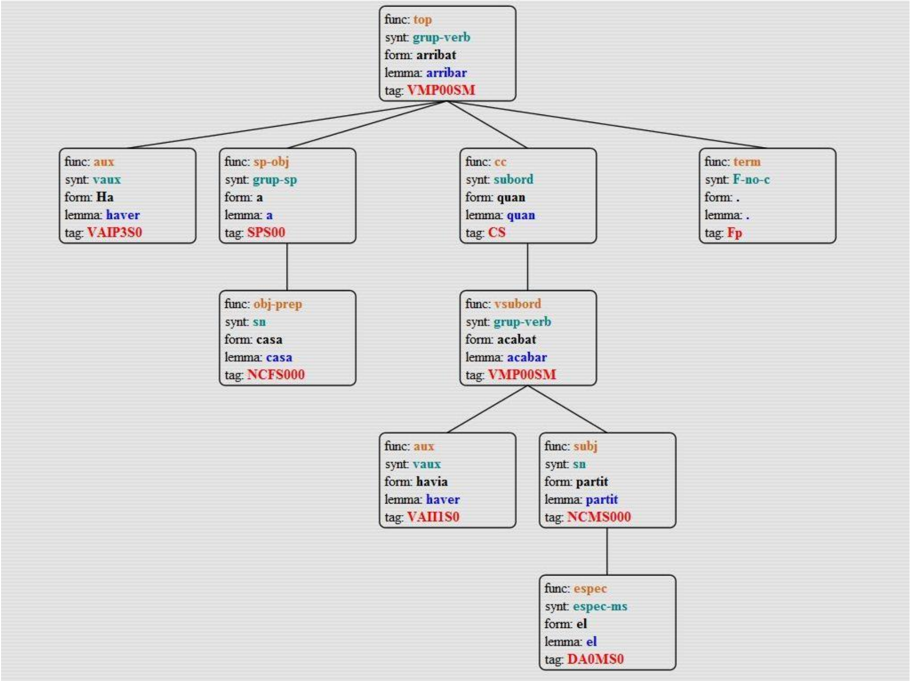


Fig. 15. Análisis de la subordinada adverbial en CaTxala del ejemplo (13ii), 'Ha arribat a casa quan el partit havia acabat' ('Ha llegado a casa cuando el partido había terminado')

El criterio establecido sigue el razonamiento generativista, ya que, precisamente, la conjunción subordinante es la categoría gramatical clave que define la construcción subordinada como tal (Moreno Cabrera 1991; Bonet y Solà 1986). Aunque cabe reconocer que este elemento nuclear mantiene numerosas diferencias con otros núcleos sintácticos (nombre, adjetivo, verbo, adverbio), es el encargado de relacionar la oración principal y la oración subordinada. Por lo tanto, la estructura que proponemos para las subordinadas substantivas y subordinadas adverbiales se puede resumir en el esquema (15).

(15) $V_{principal}$ ← CONJUNCIÓN ← $V_{subordinado}$

Este esquema explica también la representación en las gramáticas de Txala de las interrogativas indirectas y las subordinadas causales, finales, concesivas y condicionales, porque, desde una perspectiva computacional, la estructura es idéntica a ambas subordinadas descritas: la anidación de nodos implicados se establece entre el verbo principal, la conjunción subordinante y el verbo subordinado.

3.2.2. Las oraciones relativas

Como hemos avanzado, el operador gramatical característico de las subordinadas relativas (16i) es sustancialmente diferente de las subordinadas que acabamos de observar. A rasgos generales, este marcador funciona como nexo de la misma manera como ocurre en las subordinadas sustantivas y en las subordinadas adverbiales. Pero, a su vez, mantiene una relación anafórica con el antecedente (Brucart 1999: 397; Solà 2002: 2459) y, en concreto, a modo de pronombre relativo (16ii). Como consecuencia de su naturaleza pronominal, el pronombre relativo se insiere dentro de la estructura argumental de la cláusula subordinada y lleva a cabo, además, una función sintáctica dentro de la subordinada (16iii).

(16)
i. La película que vio ayer le trajo recuerdos de su pasado
ii. La película ← que
iii. Vio [OD la película] ayer

Debido a esta casuística, los formalismos gramaticales deberán tomar partido en esta cuestión y definirse para proponer una representación que diferencie las relativas de las subordinadas sustantivas y de las subordinadas adverbiales.

Desde la perspectiva generativista, se establece que el pronombre relativo introduce las relativas (Brucart 1999: 398; Solà 2002: 2459). De modo que este pronombre es el núcleo sintáctico de esta clase de subordinadas, cosa que se asemeja a las subordinadas que hemos observado (§3.2.1). No obstante, la GG reconoce que la representación de este marcador debe expresar la naturaleza pronominal de este constituyente.

Por ese motivo, se considera que los pronombres relativos son en realidad constituyentes-qu (Bonet y Solà 1986: 25; Haegeman 1991: 330),[6] es decir, categorías que ocupan la posición de COMP y que están marcadas positivamente con el rasgo [+qu].[7] Originalmente, estas categorías ocupan una posición en la estructura argumental del predicado verbal de la cláusula, pero, debido a una regla de movimiento de *qu* (Bonet y Solà 1986: 124; Haegeman 1991: 360), los constituyentes-qu se sitúan en la posición de COMP (Fig. 16).

Una concepción completamente diferente es la que propone la GD (Tesnière 1959; Mel'čuk 2003). De la misma forma que las conjunciones, como ya se ha observado, el pronombre relativo se incluye dentro del repertorio de unidades funcionales. Particularmente, según palabras de Tesnière (1959: 82), se trata de un traslativo, cuya función es "transformer la catégorie des mots pleins" (Tesnière 1959:82), es decir, operar sobre las unidades léxicas como categoría intranuclear (Tesnière 1959: 137). Por lo tanto, el pronombre relativo en ningún caso puede funcionar como núcleo sintáctico (Fig. 17).

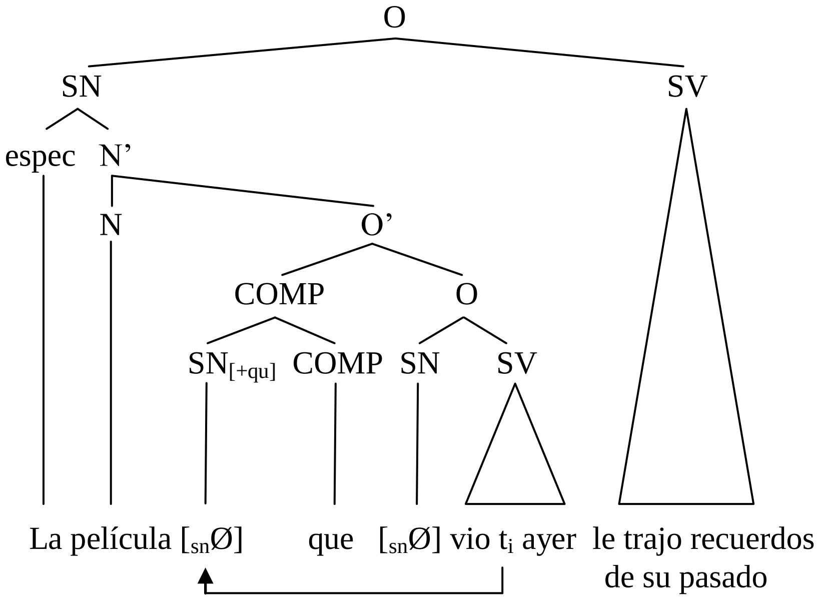


Fig. 16. Análisis de la subordinada relativa según la GG del ejemplo (13iii), 'La película que vio ayer le trajo recuerdos de su pasado'

[6] Con el término 'constituyente-qu' nos referimos efectivamente al *wh-constituent* del inglés ya que las palabras que se corresponden a este constituyente empiezan *wh-* (Haegeman, 1991:330).

[7] El rasgo [±qu] distingue, por ejemplo, los pronombres relativos de los elementos caracterizados por el rasgo [±QU] (Bonet y Solà, 1986:25). Este último rasgo es el encargado de identificar oraciones declarativas y no declarativas. En otros términos, el rasgo [±QU] se atribuye a la posición específica del complementante (COMP) y permite diferenciar los operadores gramaticales que introducen subordinadas sustantivas o de infinitivo (marcados negativamente, [-QU]) de los operadores propios de las interrogativas indirectas y directas (marcados positivamente, [+QU]).

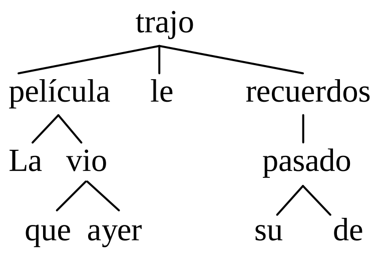


Fig. 17. Análisis de la subordinada relativa según la GD del ejemplo (13iii), 'La película que vio ayer le trajo recuerdos de su pasado'

En ambas gramáticas de Txala, el pronombre relativo es tratado como un instrumento gramatical, es decir, como una categoría gramatical que relaciona cláusulas y marca la existencia de una subordinada, de acuerdo con las hipótesis de la GG (Bonet y Solà 1986; Haegeman 1991). No obstante, la representación que se hace del pronombre relativo difiere de la representación generativista y, de la misma forma que la GD (Tesnière 1959; Mel'čuk 2003), se considera que el núcleo sintáctico de las relativas es, en realidad, el verbo subordinado.

A pesar de optar por la misma representación que la GD, los motivos que llevan a dicho análisis son alejados de este formalismo gramatical. En EsTxala y CaTxala, el pronombre relativo no es entendido como una 'palabra vacía' (en términos de *Elements de syntaxe structurale*; Tesnière, 1959); aunque se reconoce también como marcador de cláusula. La razón de este criterio viene dada por la doble funcionalidad de este marcador en las relativas, como nexo y como pronombre relativo.

El pronombre relativo no es sólo un operador gramatical sino que también desarrolla una función dentro de la cláusula subordinada, que está supeditada a la información sintáctica y semántica contenida en la pieza verbal. Por lo tanto, debe ocupar una posición dentro de la cláusula, subordinada al verbo. En resumen, el esquema de las relativas es el siguiente (17).

17) $V_{principal} \leftarrow V_{subordinado} \leftarrow PR_{rel}$

3.2.3. Las pseudorelativas

La pseudorelativa es una construcción que aparentemente es ambigua sintácticamente con la interrogativa indirecta ya que ambas subordinadas pueden aparecer con los mismos marcadores (Solà 2002: 2534). Además, esta construcción ocupa la posición típica de las subordinadas sustantivas y de las subordinadas adverbiales (Moreno Cabrera 1991: 671; Delbecque y Lamiroy 1999: 1967; Bonet 2002: 2321; Villalba 2002: 2291).

A pesar de eso, no cabe duda de que la construcción pseudorelativa está introducida por un constituyente-qu marcado positivamente (es decir, [+qu]), según la GG (Bonet y Solà 1986: 25; Haegeman 1991: 330). Por lo tanto, se diferencian de las interrogativas indirectas, cuyo operador gramatical está marcado con el rasgo [+QU] (Bonet y Solà 1986: 25). El análisis que ofrece la GG, pues, sitúa el constituyente-qu como el núcleo sintáctico de la subordinada. En este sentido, el análisis propuesto es paralelo a la estructura propia de las subordinadas relativas.

Por otro lado, la GD (Tesnière 1959; Mel'čuk 2003) considera que la estructura de las pseudorelativas es paralela a las relativas. Como consecuencia, el análisis que propone es el mismo que el análisis de las oraciones de relativo. Es decir, el operador gramatical de esta construcción es un pronombre relativo y, por lo tanto, se incluye como una categoría propia de las unidades funcionales y no tiene opción de actuar como núcleo de la cláusula ya que el verbo subordinado es el núcleo sintáctico que insiere la cláusula relativa en el marco de la oración principal.

En EsTxala y CaTxala, admitimos que esta construcción permite dos análisis, ya que nocionalmente la pseudorelativa es una relativa, pero, a su vez, ocupa la posición propia de un argumento del predicado verbal o de un adjunto. Por lo tanto, consideramos que esta construcción admite ambos análisis: el análisis previsto para las subordinadas sustantivas y subordinadas adverbiales (como consecuencia, el operador gramatical sería el núcleo sintáctico de la cláusula), y el análisis que proponemos para las relativas (es decir, la posición del núcleo sintáctico sería ocupada por el verbo de la subordinada relativa), como podemos ver en el esquema (17).

No obstante, puesto que la arquitectura del analizador Txala (Atserias et al. 2005) no permite proponer más de un análisis para cada secuencia lingüística de entrada, nos definimos por uno de los dos análisis posibles de la pseudorelativa. Como hemos expuesto, esta construcción es 'a grosso modo' una relativa en posición de argumento o adjunto. Optamos, pues, por el análisis que define esta construcción como una relativa (Tesnière 1959: 137). Por lo tanto, el núcleo sintáctico de la construcción es el verbo de la subordinada y el operador gramatical se anida como dependiente de este núcleo.

3.2.4. Las subordinadas con elisión de marcador

Como se ha descrito al inicio de esta sección, es posible la omisión del operador gramatical en determinados situaciones tanto en español como en catalán (Delbecque y Lamiroy 1999: 2026; Bonet 2002: 2348), aunque parece que el catalán es más restrictivo en esta cuestión o, como mínimo, esta elisión ha sido condenada por la normativa.

A pesar de esa controversia, ambos formalismos gramaticales aquí considerados, GG (Bonet y Solà 1986; Haegeman 1991) y GD (Tesnière 1959; Mel'čuk 2003), están de acuerdo con la representación de esta clase de elisión. La anidación se lleva a cabo entre el verbo principal de la oración y el verbo de la subordinada, que actúa como núcleo sintáctico de la cláusula. De todas formas, la GG prevé una posición vacía de COMP ya que léxicamente no se expresa el marcador pero semánticamente está presente. Como consecuencia, en EsTxala y CaTxala, también se opta por esa representación.

Debido a la existencia de subordinadas con elisión del operador gramatical y la consiguiente representación $V_{principal} \leftarrow V_{subordinado}$, se podría contradecir el criterio de representación de las subordinadas sustantivas y subordinadas adverbiales, que establece que el núcleo sintáctico de la cláusula subordinada debe ser la conjunción. No obstante, creemos que es necesario distinguir estas subordinadas, donde el marcador es sólo un elemento relacional de cláusulas, de las relativas, donde el marcador desarrolla una función sintáctica dentro de la cláusula.

3.2.5. Resumen

En esta sección, se ha descrito y detallado el conjunto de subordinadas que se toman en consideración en las gramáticas de dependencias del español y del catalán para fines de procesamiento automático del lenguaje (Lloberes et al. 2010). Con el objetivo de establecer los criterios de representación de estas construcciones se ha llevado a cabo un análisis crítico de las propuestas de representación de la GG (Bonet y Solà 1986; Haegeman 1991) y de la GD (Tesnière 1959; Mel'čuk 2003), que ha permitido determinar cómo estas estructuras deben ser tratadas en las gramáticas de Txala (Tabla 1).

	GG	*GD*	*Txala*
Sustantivas	$V_p \leftarrow$ CONJ $\leftarrow V_s$	$V_p \leftarrow V_s \leftarrow$ CONJ	$V_p \leftarrow$ CONJ $\leftarrow V_s$
Adverbiales	$V_p \leftarrow$ CONJ $\leftarrow V_s$	$V_p \leftarrow V_s \leftarrow$ CONJ	$V_p \leftarrow$ CONJ $\leftarrow V_s$
Relativas	$V_p \leftarrow PR_{rel} \leftarrow V_s$	$V_p \leftarrow V_s \leftarrow PR_{rel}$	$V_p \leftarrow V_s \leftarrow PR_{rel}$
Pseudorelativas	$V_p \leftarrow PR_{rel} \leftarrow V_s$	$V_p \leftarrow V_s \leftarrow PR_{rel}$	$V_p \leftarrow V_s \leftarrow PR_{rel}$
S. con marcador elidido	$V_p \leftarrow V_s$	$V_p \leftarrow V_s$	$V_p \leftarrow V_s$

Tabla 1. Resumen de la distribución de núcleos sintácticos de las subordinadas según GG, GD y Txala

3.3. La estructura de las comparaciones

Algunos estudios teóricos y descriptivos profundizan en el análisis de la estructura de las comparaciones, como son los trabajos de Solà (1972), Rivara (1990) y Gutiérrez Ordóñez (1994). No obstante, es escasa la caracterización de los fundamentos teóricos de esta estructura. Además, hay divergencias de opinión en la definición de las construcciones que realmente son comparativas y de los elementos o marcadores que favorecen la lectura comparativa de una determinada construcción (Saragossà 2002: 3097).

Las construcciones comparativas tienen como finalidad comunicativa dar a conocer una información que es conocida por el receptor (Saragossà 2002: 3098). Esta definición es de carácter general y poco aporta en términos de la representación sintáctica de dicha construcción. De todas formas, puede ser desglosada y explicada más concretamente.

En la comparación, se expresa una información desconocida por el receptor contrastándola con información que ya posee. Ambas informaciones se manifiestan en la oración mediante dos términos comparados, X e Y, donde X es la incógnita que al ser comparada con la información conocida recogida en Y es resuelta (Gutiérrez Ordóñez 1994: 13). El término X aparece en primer lugar y se denomina antecedente, y el término Y se expresa a continuación del antecedente y, por ello, se le conoce como consiguiente. Los dos componentes de la comparación, X e Y, entran a funcionar dentro de la construcción comparativa gracias a que están introducidos por operadores gramaticales (p.e. *más*, *tan*, etc.) que los ponen en relación (18).

(18)

i.Aquesta noia és tan alta com tu
'Esta chica es tan alta como tú'
ii. Aquesta noia és [$_{\text{antecedente}}$ tan alta] [$_{\text{consiguiente}}$ com tu]

A partir de la observación de (18), debemos cuestionarnos cómo debe ser la representación sintáctica en EsTxala y CaTxala. En realidad, existen dos análisis posibles (Fig. 18, Fig. 19):

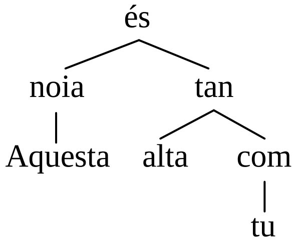


Fig. 18. Análisis de la construcción comparativa con el marcador como núcleo del ejemplo (18i) 'Aquesta noia és tan alta com tu' ('Esta chica es tan alta como tú')

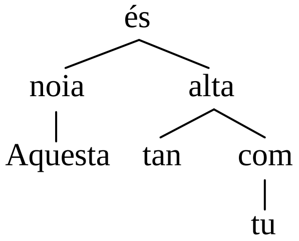


Fig. 19. Análisis de la construcción comparativa con el antecedente como núcleo

del ejemplo (18i) 'Aquesta noia és tan alta com tu' ('Esta chica es tan alta como tú')

En la Fig. 18, el análisis propuesto expresa implícitamente que el marcador funciona paralelamente como las conjunciones de las subordinadas sustantivas y adverbiales (§3.2.1) y, por lo tanto, es el introductor y el núcleo de la construcción comparativa. No obstante, al lado de ejemplos como (18i), existen otros ejemplos que plantean dificultades a esta propuesta de análisis (19).

(19) Tiene más dinero

En el ejemplo (19), se observa, por un lado, que el consiguiente no es expresado léxicamente y, por el otro, el introductor *más* actúa como un modificador del nombre *dinero*. De modo que, en realidad, el primer marcador de la comparación tiene el mismo valor que los cuantificadores (Brucart y Rigau 2002: 154) y, por lo tanto, modifica el núcleo del antecedente. En EsTxala y CaTxala, la estrategia para representar las construcciones comparativas se corresponde a este último análisis y, por lo tanto, los análisis propuestos son paralelos al análisis ilustrado en la Fig. 19, donde el antecedente es el núcleo de la construcción comparativa. A pesar de la validez de ambos análisis, desde la perspectiva del PLN, el tratamiento de las construcciones comparativas plantea graves problemas debido a la gran diversidad de estructuras y de elisiones (por ejemplo, elisión de toda la cláusula que expresa el consiguiente o parte de ella).

3.4. La conjunción coordinante como elemento nuclear

A lo largo de este artículo, se ha discutido sobre estructuras y construcciones sintácticas que tienen como rasgo común el hecho de que son relaciones estructurales jerárquicas entre un núcleo sintáctico y los nodos que dependen de él. Ahora bien, en cuanto a la coordinación, el fenómeno que se nos presenta es sustancialmente diferente, ya que generalmente en esta estructura se ha considerado que los constituyentes se encuentran alineados al mismo nivel, tal y como expone Trask (1992: 63):

> **coordinate structure** *n.* A syntactic structure in which two or more constituents are joined ('conjoined') in such a way that each of them has an equal claim to be considered a head of that structure. […]

No obstante, se ha discutido si la ordenación de los constituyentes se da al mismo nivel o, en realidad, se corresponde a una jerarquía (Bonet y Solà 1986: 321; López García 1999: 3513; Serra y Prunyonosa 2002: 2183).

Siguiendo a Bonet y Solà (1986: 349), desde una perspectiva generativista, la construcción coordinada consta de dos o más elementos que son equivalentes en cuanto a la función gramatical y están unidos al mismo nivel de la jerarquía estructural mediante un elemento de enlace (Dik 1968: 25). A partir de esa definición, Bonet y Solà (1986) hipotetizan que la coordinación puede ser explicada mediante la regla (20), que favorece el análisis de la coordinación como una ordenación al mismo nivel de los constituyentes (Fig. 20).

(20) X → X Coord X[8]

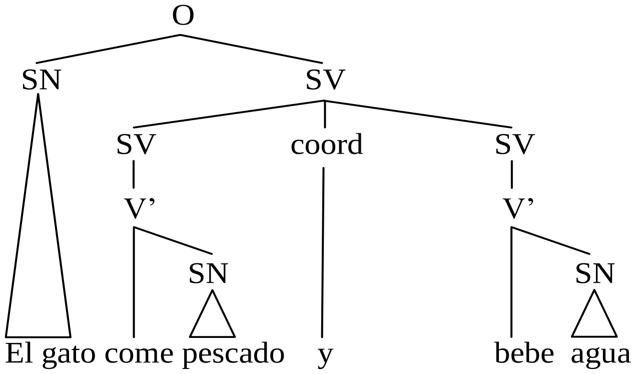


Fig. 20. Análisis de la construcción coordinada según la GG de la oración 'El gato come pescado y bebe agua'

Al lado de la preposición y de la conjunción subordinante, la GD considera la conjunción coordinante una categoría intranuclear por su naturaleza conjuntiva y, por lo tanto, una clase de unidad funcional (Tesnière 1959: 53). Al no contener ninguna carga en cuanto a la función semántica, esta categoría no funciona como núcleo sintáctico de la coordinación (Mel'čuk 2003: 50), tal como se ilustra en Fig. 21.

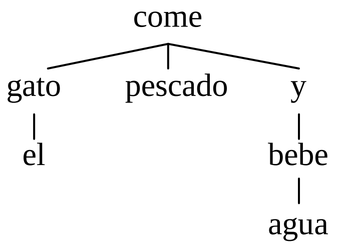


Fig. 21. Análisis de la construcción coordinada según la GD de la oración 'El gato come pescado y bebe agua'

En este punto, EsTxala y CaTxala discrepan totalmente del análisis que propone la GD (Tesnière 1959; Mel'čuk 2003), ya que, como expone Bonet y Solà (1986: 349), la construcción coordinada no responde a una estructura jerárquica y la categoría que caracteriza esta construcción es la propia conjunción coordinada.

[8] La regla expuesta a (20) prevé la posibilidad de piezas gramaticales léxicamente vacías, que pueden ser recuperadas mediante los elementos del primer coordinado (si los elementos vacíos se encuentran en el segundo coordinado), elementos externos a la construcción coordinada o elementos desplazados al margen derecho de la coordinación (en el caso que los elementos vacíos estén en el segundo coordinado).

Además, la estructura que propone la GD supone en numerosas ocasiones el nacimiento de ambigüedades estructurales (Mel'čuk 2003: 93), que a nuestro entender deben ser evitadas en el análisis automático del lenguaje. Por ejemplo, según los postulados de la GD, en la coordinación *wonderful photographs and paintings* ('fotografías y dibujos preciosos') hay implicadas dos interpretaciones. El adjetivo *wonderful* puede modificar sólo al nombre *photographs*, pero, a la vez, puede observarse que este adjetivo modifica a todo el grupo nominal coordinado, es decir, a *photohraphs* y *paintings*.

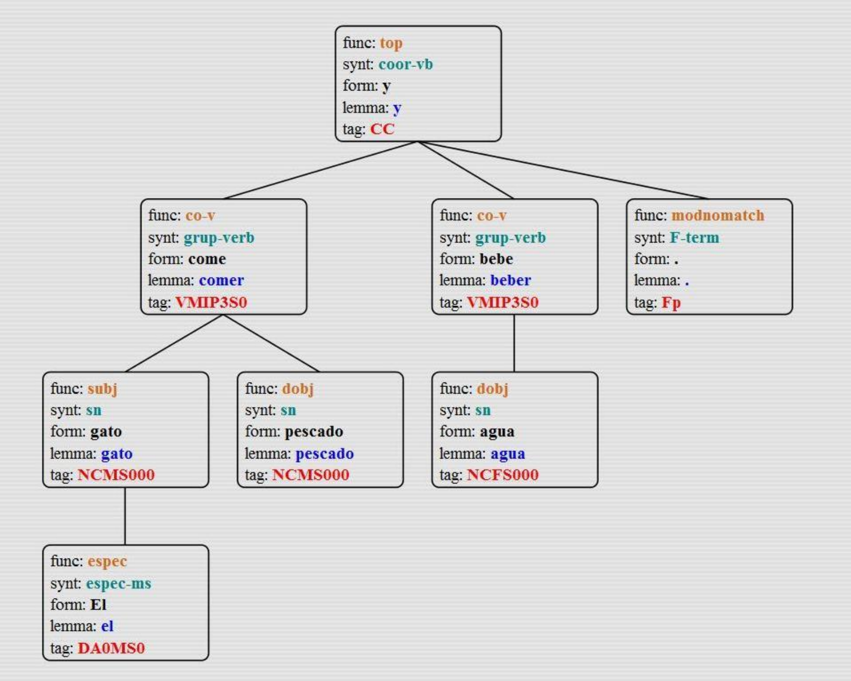


Fig. 22. Análisis de la construcción coordinada según EsTxala de la oración 'El gato come pescado y bebe agua'

De modo que el criterio establecido en ambas gramáticas para la coordinación no respalda las teorías de la GD. En otras palabras, el núcleo de toda coordinación debe ser la conjunción coordinada (Fig. 22).

4. CONCLUSIONES

En este artículo, hemos presentado los criterios adoptados para el tratamiento de diversas estructuras sintácticas en las gramáticas EsTxala y CaTxala. Hemos tratado aquellas construcciones o estructuras que desde una perspectiva teórica y computacional plantean diferentes soluciones de representación en función del formalismo lingüístico. Concretamente, hemos focalizado en la representación del sintagma preposicional y de las oraciones subordinadas (sustantivas, adverbiales, relativas, pseudorelativas y con elisión de marcador). A su vez, hemos abordado las estructuras comparativas y las coordinaciones.

Para llegar a un criterio claro en las gramáticas, hemos llevado a cabo una revisión crítica de las propuestas para este conjunto de construcciones y estructuras de los dos grandes formalismos lingüísticos implementados en análisis automático del lenguaje, la Gramática Generativa (Chomsky 1981) y la Gramática de Dependencias (Tesnière 1959; Mel'čuk 2003). Como consecuencia de la exposición de las diversas propuestas teóricas, se ha podido diseñar un repertorio de criterios lingüísticos con el fin de aportar soluciones fundamentadas sintáctica y semánticamente.

A pesar de que la tarea de elaboración de criterios es abierta, consideramos que los criterios sintácticos que hemos establecido son una versión suficientemente elaborada, ya que tratan aquellos fenómenos sintácticos que han sido interpretados de manera distinta según la Gramática Generativa y la Gramática de Dependencias. No obstante, puesto que se trata de una tarea sin límites, en las siguientes etapas del proyecto se prevé la posibilidad de incorporar o modificar criterios según las necesidades de cada etapa. Estos criterios están implementados en las gramáticas que son de acceso abierto (http://grial.uab.es/tools/download/).

En paralelo, a este trabajo, hemos desarrollado un corpus de test para gramáticas automáticas que sigue los criterios propuestos y que incluye otras estructuras incluidas en EsTxala y CaTxala además de las presentadas en este artículo. Dicho repertorio es abierto y consultable en línea (http://161.116.36.206/~publicacions/articlesDEPGRAM/repertori_avaluacio_text.pdf).

La definición del repertorio de criterios sintácticos permite plantear como línea de investigación futura la evaluación de la calidad del recurso desarrollado, las gramáticas EsTxala y CaTxala, y, en último término, la evaluación de los criterios sintácticos propuestos. Para llevar a cabo esta tarea, estamos diseñando un protocolo de evaluación para dichas gramáticas que ponga a prueba la calidad del conocimiento lingüístico contenido en las gramáticas, ya sea información sintáctica (anidación del sintagma preposicional, formación de oraciones complejas, subcategorización verbal, estructuras coordinadas, etc.) como información semántica (preferencias selectivas, rasgos de *Top Concept Ontology*).

BIBLIOGRAFÍA

ALARCOS, E. (1994), *Gramática de la lengua española*, Madrid, Espasa Calpe, Real Academia Española.

ALSINA, À., BADIA, T., BOLEDA, G., BOTT, S., GIL, À., QUIXAL, M. y VALENTÍN, O. (2002), "CATCG: Un sistema de análisis morfosintáctico para el catalán", en *Procesamiento del Lenguage Natural*, 29, 309-310.

ATSERIAS, J., COMELLES, E. y MAYOR, A. (2005), "Txala un analizador libre de dependencias para el castellano", en *Procesamiento del Lenguaje Natural*, 35:455-456.

BICK, E. (2006). "A constraint grammar-based parser for Spanish", en *Proceedings of TIL 2006, 4th Workshop on Information and Human Language Technology*.

BADIA I MARGARIT, A. M. (1962), *Gramática catalana*, Madrid, Gredos.

BONET, S. y SOLÀ, J. (1986), *Sintaxi generativa catalana*, Barcelona, Biblioteca Universitària, Enciclopèdia Catalana.

BONET, S. (2002), "Les subordinades substantives", en *Gramàtica del Català Contemporani*, Solà, J., Lloret, M.R., Mascaró, J. y Pérez Saldanya, M. (dirs.), vol. 3, Barcelona, Empúries, 2321-2387.

BOSQUE, I. y DEMONTE, V. (dirs.) (1999), *Gramática descriptiva de la lengua española*, Madrid, Espasa.

BRUCART, J. M. (1999), "La estructura del sintagma nominal: las oraciones de relativo", en *Gramática descriptiva de la lengua española*, Bosque, I. y Demonte, V. (dirs.), vol. 1., Madrid, Espasa, 395-522.

BRUCART, J. M. y RIGAU, G. (2002), "La quantificació", en *Gramàtica del Català Contemporani*, Solà, J., Lloret, M.R., Mascaró, J. y Pérez Saldanya, M. (dirs.), vol. 2, Barcelona, Empúries, 1517-1589.

CALVO, H. y GELBUKH, A. (2006), "DILUCT: An open–source Spanish dependency parsers based on rules, heuristics, and selectional preferences", en *Natural Language Processing and Information Systems. Lecture Notes in Computer Science*, Kop, C., Fliedl, G., Mayr, H.C. y Métais, E. (eds.), Berlin, Springer-Verlag, vol. 3999, 164-175.

CASTELLÓN, I., CIVIT, M. y ATSERIAS, J. (1998), "Syntactic parsing of unrestricted spanish text", en *First International Conference on Language Resources and Evaluation*, España, Granada.

CHOMSKY, N. (1957), *Syntactic Structures*, The Hague, Mouton de Gruyter.

_________ (1981^7), *Lectures on Government and Binding: The Pisa Lectures*, The Hague, Mouton de Gruyter.

COLLINS, M. (2000), "Discriminative reranking for natural language parsing", en *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000)*, 175-182.

COMELLES, E., ARRANZ, V. y CASTELLÓN, I. (2010), "Constituency and Dependency Parsers Evaluation", en *Procesamiento del Lenguaje Natural*, 45: 59-66.

DE MARNEFFE, M.C., MACCARTNEY, B. y MANNING, C. D. (2006), "Generating typed dependency parses from phrase structure parses", en *Proceedings of the Fifth Conference on International Language Resources and Evaluation (LREC'06)*, 449-454.

DELBECQUE, N. y LAMIROY, B. (1999), "La subordinación sustantiva: las subordinadas enunciativas en los complementos verbales", en *Gramática descriptiva de la lengua española*, Bosque, I. y Demonte, V. (dirs.), vol. 2., Madrid, Espasa, 1965-2081.

DIK, S. C. (1968), *Coordination. Its implications for the theory of general linguistics*, Amsterdam, North-Holland.

FABRA, P. (1918^3), *Gramàtica catalana*, Barcelona, Institut d'Estudis Catalans.

FERRÁNDEZ, A., PALOMAR, M. y MORENO, L. (2000), "Slot unification grammar and anaphora resolution", en *Recent Advances in Natural Language Processing II. Selected papers from RANLP 1997*, Nicolov, N. y Mitkov, R. (eds.), John Benjamins, 155-166.

GAMALLO, P. y GONZÁLEZ, I. (2009), "Una gramática de dependencias basada en patrones de etiquetas", en *Procesamiento del Lenguaje Natural*, 43, 315-323.

GUTIÉRREZ ORDÓÑEZ, S. (1994), *Estructuras comparativas*, Madrid, Arco/Libros.

HAEGEMAN, L. (1991), *Introduction to Government and Binding Theory*, Oxford, Blackwell.

HERNANZ, M. Ll. (2002), "L'oració", en *Gramàtica del Català Contemporani*, Solà, J., Lloret, M.R., Mascaró, J. y Pérez Saldanya, M. (dirs.), vol. 2, Barcelona, Empúries, 993-1073.

HERNANZ, M. Ll. y BRUCART, J. M. (1987), *La sintaxis. Principios teóricos. La oración simple*, Barcelona, Editorial Crítica.

KOO, T., CARRERAS, X. y COLLINS, M. (2008), "Simple semi-supervised dependency parsing", en *Proceedings of 46th Annual Meeting on Association for Computational Linguistics (ACL'08)*, 595-603.

LLOBERES, M., CASTELLÓN, I. y PADRÓ, Ll. (2010), "Spanish FreeLing Dependency Grammar", en *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Calzolari, N. et al. (eds.), Valletta, Malta.

LIN, D. (1998), "Dependency-based evaluation of MINIPAR", en *Proceedings of the LREC Workshop on the Evaluation of Parsing Systems*, España, Granada.

LÓPEZ GARCÍA, Á. (1999), "Relaciones paratácticas e hipotácticas", en *Gramática descriptiva de la lengua española*, Bosque, I. y Demonte, V. (dirs.), vol. 2., Madrid, Espasa, 3507-3547.

MAGERMAN, D. M. (1995), "Statistical decision-tree models for parsing", en *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics (ACL'95)*, 276-283.

MARIMÓN, M. (2010), "The Spanish Resource Grammar", en *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, Calzolari, N. et al. (eds.),Valletta, Malta.

MEL'ČUK, I. A. (1988), *Dependency Syntax: Theory and Practice*, Albany-New York, The SUNY Press.

_________ (2003), "Dependency in Natural Language", en *Dependency in Linguistic Description. Studies in Language Companion Series*, Polguère, A. y Mel'čuk (eds.), I.A., Amsterdam-Philadelphia, John Benjamins Publishing Co., 1-110.

MORENO CABRERA, J.C. (1991^2), *Curso universitario de lingüística general: Teoría de la gramática y sintaxis general*, Madrid, Síntesis.

NIVRE, J. (2006). *Inductive Dependency Parsing*, en *Text, Speech and Language Technology*, vol. 34, Dordrecht, Springer-Verlag, 2006.

PADRÓ, Ll., REESE, S., LLOBERES, M. y CASTELLÓN, I. (2010), "Freeling 2.1: Five years of open-source language processing tools, en *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Calzolari, N. et al. (eds.), Valletta, Malta.

RIVARA, R. (1990), *Le système de la comparaison. Sur la construction du sens dans les langues naturelles*, Paris, Les Éditions de Minuit.

SANCHO CREMADES, P. (2002), "La preposició i el sintagma preposicional", en *Gramàtica del Català Contemporani*, Solà, J., Lloret, M.R., Mascaró, J. y Pérez Saldanya, M. (dirs.), vol. 2, Barcelona, Empúries,1689-1796.

SARAGOSSÀ, A. (2002), "Les construccions comparatives i les oracions consecutives", en *Gramàtica del Català Contemporani*, Solà, J., Lloret, M.R., Mascaró, J. y Pérez Saldanya, M. (dirs.), vol. 3, Barcelona, Empúries, 3095-3171.

SCHMID, H. (1994), *TreeTagger - a language independent part-of-speech tagger*, http://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html.

SERRA, E. y PRUNYONOSA, T. (2002), "La coordinació", en *Gramàtica del Català Contemporani*, Solà, J., Lloret, M.R., Mascaró, J. y Pérez Saldanya, M. (dirs.), vol. 3, Barcelona, Empúries, 2181-2245.

SLEATOR, D. D.K. y TEMPERLEY, D. (1991), "Parsing english with a link grammar", en *Third International Workshop on Parsing Technologies*, Tilburg, The Netherlands y Durbuy, Belgium.

SOLÀ, J. (1972^2), *Estudis de sintaxi catalana*, Barcelona, Edicions 62.

_________ (2002), "Les subordinades de relatiu", en *Gramàtica del Català Contemporani*, Solà, J., Lloret, M.R., Mascaró, J. y Pérez Saldanya, M. (dirs.), vol. 3, Barcelona, Empúries, 2455-2565.

SOLÀ, Joan, LLORET, M. Rosa, MASCARÓ, Joan y PÉREZ SALDANYA, Manuel (dirs.) (2002), *Gramàtica del Català Contemporani*, Barcelona, Empúries.

TAPANAINEN, P. y JÄRVINEN, T. (1998), "Towards an implementable dependency grammar", en *Proceedings of the ACL Workshop on Processing of Dependency-Based Grammars*, 1-10.

TESNIÈRE, L. (1959), *Elements de syntaxe structurale*, Paris, Klincksieck.

TRASK, R. L. (1992), *A dictionary of grammatical terms in Linguistics*, London-New York, Routledge.

VILLALBA, X. (2002), "La subordinació", en *Gramàtica del Català Contemporani*, Solà, J., Lloret, M.R., Mascaró, J. y Pérez Saldanya, M. (dirs.), vol. 3, Barcelona, Empúries, 2247-2319.

YAMADA, H. y MATSUMOTO, Y. (2003), "Statistical dependency analysis with support vector machines", en *Proceedings of 8th International Workshop on Parsing Technologies*, 195-206.

ParTes. Test Suite for Parsing Evaluation [*]

ParTes: Test suite para evaluación de analizadores sintácticos

Marina Lloberes
GRIAL-UB
marina.lloberes@ub.edu

Irene Castellón
GRIAL-UB
icastellon@ub.edu

Lluís Padró
TALP-UPC
padro@lsi.upc.edu

Edgar Gonzàlez
Google Research
edgargip@google.com

Resumen: En este artículo se presenta ParTes, el primer test suite en español y catalán para la evaluación cualitativa de analizadores sintácticos automáticos. Este recurso es una jerarquía de los fenómenos representativos acerca de la estructura sintáctica y el orden de argumentos. ParTes propone una simplificación de la evaluación cualitativa contribuyendo a la automatización de esta tarea.
Palabras clave: test suite, evaluación cualitativa, analizador sintáctico, español, catalán

Abstract: This paper presents ParTes, the first test suite in Spanish and Catalan for parsing qualitative evaluation. This resource is a hierarchical test suite of the representative syntactic structure and argument order phenomena. ParTes proposes a simplification of the qualitative evaluation by contributing to the automatization of this task.
Keywords: test suite, qualitative evaluation, parsing, Spanish, Catalan

1 *Introduction*

Qualitative evaluation in Natural Language Processing (NLP) is usually excluded in evaluation tasks because it requires a human effort and time cost. Generally, NLP evaluation is performed with corpora that are built over random language samples and that correspond to real language utterances. These evaluations are based on frequencies of the syntactic phenomena and, thus, on their representativity, but they usually exclude low-frequency syntactic phenomena. Consequently, current evaluation methods tend to focus on the accuracy of the most frequent linguistic phenomena rather than the accuracy of both high-frequent and low-frequent linguistic phenomena.

This paper takes as a starting point these issues related to qualitative evaluation. It presents ParTes, the first parsing test suite in Spanish and Catalan, to allow automatic qualitative evaluation as a complementary task of quantitative evaluation. This resource is designed to simplify the issues related to qualitative analysis reducing the human effort and time cost. Furthermore, ParTes provides a set of representative linguistic utterances based on syntax. The final result is a hierarchical test suite of syntactic structure and argument order phenomena defined by means of syntactic features.

2 *Evaluation databases*

Traditionally, two analysis methods have been defined: the quantitative analysis and the qualitative analysis. Both approaches are complementary and they can contribute to a global interpretation.

The main difference is that quantitative analysis relies on statistically informative data, while qualitative analysis talks about richness and precision of the data (McEnery and Wilson, 1996).

Representativeness by means of frequency is the main feature of quantitative studies. That is, the observed data cover the most frequent phenomena of the data set. Rare phenomena are considered irrelevant for a quantitative explanation. Thus, quantitative descriptions provide a close approximation of the real spectrum.

Qualitative studies offer an in-depth description rather than a quantification of the data (McEnery and Wilson, 1996). Frequent phenomena and marginal phenomena are considered items of the same condition because the focus is on providing an exhaus-

[*] The resource presented in this paper arises from the research project SKATER: Scenario Knowledge Acquisition by Textual Reading, funded by the Ministry of Economy and Competitiveness (TIN2012-38584-C06-06 and TIN2012-38584-C06-01). Edgar Gonzàlez collaborated in the ParTes automatization process. We thank Marta Recasens for her help and suggestions on this paper.

tive description of the data.

In terms of analysis methods and databases, two resources have been widely used: corpora and test suites. Language technologies find these resources a reliable evaluation test because they are coherent and they are built over guidelines.

A corpus contains a finite collection of representative real linguistic utterances that are machine readable and that are a standard reference of the language variety represented in the resource itself (McEnery and Wilson, 1996). From this naive conceptualization, Corpus Linguistics takes the notion of representativeness as a presence in a large population of linguistic utterances, where the most frequent utterances are represented as a simulation of the reality and they are annotated according to the resource goals. That is why corpora are appropriate test data for quantitative studies.

On the other hand, test suites are structured and robust annotated databases which store an exhaustive collection of linguistic utterances according to a set of linguistic features. They are built over a delimited group of linguistic utterances where every utterance is detailed and classified according to rich linguistic and non-linguistic annotations (Lehmann et al., 1996). Thus, the control over test data and their detailed annotations make test suites a perfect guidance for qualitative studies.

Corpora have also been used in qualitative analysis, but they collect representative linguistic utterances by means of frequency rather than the representative linguistic utterances by means of exhaustiveness. Then, they are not the most appropriate tool for qualitative studies.

3 *Existing test suites*

NLP software evaluation tasks share the purpose of test suites, which originally were designed for code validation in software development. Traditional test suites were simple collections of linguistic test cases or interesting examples. However, with the success of the NLP technologies, there was a real need for developing test suites based on pre-defined guidelines, with a deep structure, richly annotated and not necessarily developed for a particular tool (Flickinger, Nerbonne, and Sag, 1987). For this reason, the new generation of test suites are databases that cover the real needs of the NLP software evaluation (Lehmann et al., 1996).

The HP test suite (Flickinger, Nerbonne, and Sag, 1987) is an English and general purpose resource developed to diagnose and monitor the progress of NLP software development. The main goal of this test suite is to evaluate the performance of heuristic-based parsers under development. The suite contains a wide-range collection of linguistic examples that refer to syntactic phenomena such as argument structure verbs and verbal subcategorization among others. It also includes some basic anaphora-related phenomena. Furthermore, these phenomena are represented by a set of artificially constructed sentences and the annotations are shallow. This resource has a minimal internal classification since the suite organizes the test data under headings and sub-headings.

In order to step further, subsequent test suites have been developed as in-depth resources with rich structure and annotations. One of the groups of the Expert Advisory Group on Language Engineering Standards (EAGLES) proposes a set of guidelines for evaluating grammar checkers based on test suites (EAGLES, 1994). The test suite is a collection of attributes that allow to validate the quality of the functions of the evaluated tool. It is derived from a taxonomy of errors, where each error class is translated into a feature which is collected in the test suite. The final result is a classification of sentences containing an error, the corresponding sentence without the error, the name of the error and the guidelines for the correction process.

The TSNLP (Lehmann et al., 1996) is a multilingual test suite (English, French and German) richly annotated with linguistic and meta-linguistic features. This test suite is a collection of test items with general, categorial and structural information. Every test item is classified according to linguistic and extra-linguistic features (e.g. number and type of arguments, word order, etc.). These test items are also included in test sets by means of positive and negative examples. Furthermore, the TSNLP includes information about frequency or relevance for a particular domain.

In Spanish, a previous test suite exists for NLP software evaluation, the SPARTE test suite (Peñas, Álvaro, and Verdejo, 2006). Specifically, it has been developed to validate Recognizing Textual Entailment systems and it is a collection of text and hypothesis pairs with true/false annotations. Although SPARTE and the presented ParTes in Spanish (ParTesEs) are resources for the same language, both test suites have been developed for different purposes which make both resources unique. With respect to the Catalan language, the version of ParTes in Catalan (ParTesCa) is the first test suite for this language.

4 The construction of ParTes

ParTes is a new test suite in Spanish and Catalan for qualitatively evaluating parsing systems. This test suite follows the main trends on test suite design, so that it shares some features with the EAGLES test suite (EAGLES, 1994) and the TSNLP (Lehmann et al., 1996).

Additionally, ParTes adds two new concepts in test suite design concerning how the data are classified and which data are encoded. The test suite is seen as a hierarchy where the phenomenon data are explicitly connected. Furthermore, representativeness is the key-concept in ParTes to select the phenomenon-testing data that configure the test suite.

The ParTes guidelines are created to ensure the coherence, the robustness and the easy implementation of this resource.

Specific purpose. While some test suites are general purpose like TSNLP, ParTes is a specific purpose test suite. Particularly, it is focused to validate the accuracy of the syntactic representations generated by parsers. For this reason, the test cases are related to syntactic phenomena and the test suite has been annotated with several syntactic features.

Test suite of syntactic phenomena. ParTes is not a simple collection of linguistic test cases nor a set of linguistic features, actually. This resource lists the syntactic phenomena that configure a language by a set of syntactic features.

For example, ParTes collects syntactic structures based on head-child relation. It also contains several features that syntactically define every phenomenon (e.g. the syntactic category of the head or the child, the syntactic relation with the node that governs it, etc.). Complementarily, every phenomenon is associated with a test case that corresponds to the linguistic utterance of the actual phenomenon described and that is used to evaluate the accuracy of the performance of the parser.

Hierarchy of syntactic phenomena. Previous test suites were a collection of test sentences, optionally structured (EAGLES and TSNLP). ParTes proposes a hierarchically-structured set of syntactic phenomena to which tests are associated.

Polyhedral hierarchy. Test suites can define linguistic phenomena from several perspectives (e.g. morphologic features, syntactic structures, semantic information, etc.). Because ParTes is built as a global test suite, it defines syntactic phenomena from two major syntactic concepts: syntactic structure and argument order (Section 5).

Exhaustive test suite. In order to evaluate NLP tools qualitatively, test suites list exhaustively a set of linguistic samples that describe in detail the language(s) of the resource, as discussed in Section 2. ParTes is not an exception and it contains an exhaustive list of the covered syntactic phenomena of the considered languages. However, some restrictions are applied to this list. Otherwise, listing the whole set of syntactic phenomena of a language is not feasible, and it is not one of the goals of the test suite's design.

Representative syntactic phenomena. As mentioned, lists of test cases need to be delimited because test suites are controlled data sets. Similarly to corpora development, the syntactic phenomena to be included in the test suite can be selected according to a certain notion of representativeness. Consequently, representative syntactic phenomena are relevant for testing purposes and they should be added in the test suite, whereas peripheral syntactic phenomena can be excluded. The next section (Section 5) details the definition of representativeness in ParTes and how it is implemented.

Rich annotations. Every syntactic phenomenon of ParTes is annotated with precise information that provides a detailed description and that allows the qualitative interpretation of the data. The annotations refer to several linguistic and extra-linguistic features that determine the syntactic phenomena.

Controlled data. As argued in Section 2, there is a direct relation between qualitative evaluation, test suites and controlled test data. Because ParTes is a test suite for qualitative evaluation, there is a strong control over the test data and, specifically, the control is applied in a double way. The number of test cases is limited to human-processing size. The sentences of the test cases are controlled to avoid ambiguities and interactions with other linguistic utterances. For this reason, test cases are artificially created.

Semi-automatically generated. Linguistic resources usually have a high cost in terms of human effort and time. For this reason, automatic methods have been implemented whenever it has been possible. Manual linguistic description of the syntactic structure has been the main method to annotate the syntactic phenomena related to the structure. On the other hand, argument order annotations have been automatically generated and manually reviewed, using the automatization process of the SenSem corpus (Fernández and Vàzquez, 2012).

Multilingual. The architecture of this resource allows it to be developed in any language. The current version of ParTes includes the Spanish version of the test suite (ParTesEs) and the Catalan version (ParTesCa).

5 *The results of ParTes*

The final result of ParTes is an XML hierarchically and richly annotated test suite of the representative syntactic phenomena of the Spanish (ParTesEs) and Catalan (ParTesCa) languages. This resource is the first test suite for the evaluation of parsing software in the considered languages. It is freely available[1] and distributed under the Creative Commons Attribution-ShareAlike 3.0 Unported License.

ParTes is built over two kinds of information: the test suite module with the syntactic phenomena to be evaluated and the test data module with the linguistic samples to evaluate over. Since it is a polyhedral test suite, it is organized according to two major concepts in Syntax: structure and order. Table 1 gives the size of the current version of ParTes.

Section	**ParTesEs**	**ParTesCa**
Structure	99	101
Order	62	46
Total	**161**	**147**

Table 1: ParTes in numbers

5.1 Syntactic structure

The structure section is a hierarchy of syntactic levels where each level receives a tag and it is associated to a set of attributes that define several aspects about the syntactic structure. This section is placed between the `<structure></structure>` tags and it is organized into the following parts:

`<level>` It can be intrachunk (i.e. any structure inside a chunk) or intraclause (i.e. any connection between a clause marker and a grammatical category, phrase or clause).

`<constituent>` Phrase or clause that determines the nature of the constituent (e.g. noun phrase, verb phrase, infinitive clause, etc.). The head of the constituent corresponds to the parent node.

`<hierarchy>` Given two connected constituents, it defines which one occurs in the parent position and which other one in the child position.

`<realization>` Definition of the attributes of the head or child:

- `id`: Numerical code that identifies every `<realization>`.
- `name`: Name of the gramatical category, phrase or clause that occurs in head or child position (e.g. noun, pronoun, etc., as heads of noun phrase).
- `class`: Specifications about the gramatical category, the phrase or the clause that occurs in head or child position (e.g. a nominal head can be a common noun or a proper noun).
- `subclass`: Sub-specifications about the gramatical category, the phrase or the clause that occur in head or child position (e.g. a nominal head can be a bare noun).
- `link`: Arch between parent and child expressed by Part of Speech tags (e.g. the link between a nominal head and a modifying adjective is ‘n-a’).

[1] http://grial.uab.es/descarregues.php

```
<constituent name="verbphrase">
   <hierarchy name="head">
      <realization id="0001" name="verb" class="finite" subclass="default" link="null"
                   parent="salir" child="null" freq="null"
                   test="Saldrán"/>
      <realization id="0002" name="verb" class="nonfinite" subclass="default" link="null"
                   parent="viajar" child="null" freq="null"
                   test="Hubiesen viajado"/>
   </hierarchy>
   <hierarchy name="child">
      <realization id="0003" name="verb" class="auxiliar" subclass="haber" link="v-v"
                   parent="vender" child="haber" freq="0.010655" test="Habrán vendido la casa"/>
      <realization id="0004" name="verb" class="auxiliar" subclass="ser" link="v-v"
                   parent="acusar" child="ser" freq="0.010655"
                   test="Es acusada de robo"/>
      ...
      <realization id="0009" name="noun" class="null" subclass="default" link="v-n"
                   parent="romper" child="taza" freq="0.131629"
                   test="La taza se rompió"/>
      <realization id="0010" name="adjective" class="null" subclass="default" link="v-a"
                   parent="considerar" child="innovador" freq="0.010373"
                   test="Se considera una propuesta innovadora"/>
            ...
   </hierarchy>
</constituent>
```

Figure 1: Syntactic structure of the verb phrase in ParTesEs

- `parent`: Lemma of the upper level between the two nodes defined in `link` (e.g. in 'casa cara' - 'expensive house', the parent is 'casa').

- `child`: Lemma of the lower level between the two nodes defined in `link` (e.g. in 'casa cara' - 'expensive house', the child is 'caro').

- `freq`: Relative frequency in the AnCora corpus of the link between the two nodes defined in `link`.

- `test`: Linguistic test data that illustrates the syntactic structure.

For example, in the definition of verb phrase as `<constituent name="verbphrase">` (Figure 1), the possible grammatical categories, phrases and clauses that can form a verb phrase are detected and classified into two categories: those pieces that can be the head of the verb phrase (`<hierarchy name="head">`) and those that occur in child position (`<hierarchy name="child">`).

Next, the set of the possible heads of the verb phrase are listed in the several instances of `<realization>`. Furthermore, all the candidates of the child position are identified.

Every realization is defined by the previous set of attributes. In the Figure 1, in the case where the realization of one of the verb phrase children is a noun (`<realization ... name="noun".../>`), the frequency of occurrence of this link (i.e. the link of a verbal head and a nominal child, `link="v-n"`) is 0.131629 (in a scale between 0 and 1) and the test case to represent this structure is 'La taza se rompió' ('The cup broke'). Furthermore, the parent of the link 'v-n' of the test case is the lemma of the finite verb form 'rompió' (`parent="romper"`, 'to break') and the child of this link is the substantive 'taza' (`child="taza"`, 'cup'). The rest of this realization's attributes are empty.

As mentioned in Section 4, the most representative syntactic structure phenomena have been manually collected. In order to determine which phenomena are relevant to be included in ParTes, linguistic descriptive grammars have been used as a resource in the decision process. Thus, the syntactic phenomena that receive a special attention in the descriptive grammars can be considered candidates in terms of representativeness. In particular, the constructions described in *Gramática Descriptiva de la Lengua Española* (Bosque and Demonte, 1999) and in *Gramàtica del Català Contemporani* (Solà et al., 2002), for Spanish and Catalan respectively, have been included.

In addition, the representativeness of the selected syntactic phenomena is supported by the frequencies of the syntactic head-child relations of the AnCora corpus (Taulé, Martí, and Recasens, 2008). These frequencies are automatically extracted and they are generalizations of the Part of Speech tag of both head and child given a link: all the main verb instances are grouped together, the auxiliaries are recognized into the same class, etc. Some frequencies are not extracted due to the complexity of certain constructions. For example, comparisons are excluded because it is not possible to reliable detect them by automatic means in the corpus.

The representation of the syntactic structures in ParTes follows the linguistic proposal implemented in FreeLing Dependency Grammars (Lloberes, Castellón, and Padró, 2010). This proposal states that the nature of the lexical unit determines the nature of the head and it determines the list of syntactic categories that can occur in the head position.

5.2 Argument order

Similarly to the syntactic structure section, the argument order schemas are also a hierarchy of the most representative argument structures that occur in the SenSem corpus. This section is organized in ParTes as follows:

`<class>` Number and type of arguments in which an order schema is classified. Three classes have been identified: monoargumental with subject expressed (`subj#V`), biargumental where subject and object are expressed (`subj#V#obj`), and monoargumental with object expressed (`V#obj`).

`<schema>` Sub-class of `<class>` where the argument order and the specific number of arguments are defined. For example, ditransitive verbs with an enclitic argument (e.g. '[El col·leccionista$_{subj}$] no [li$_{iobj}$] [ven$_{v}$] [el llibre$_{dobj}$]' - 'The collector to him do not sell the book') are expressed by the schema `subj#obj#V#obj` (Figure 2).

`<realization>` Specifications of the argument order schema, which are defined by the following set of attributes (Figure 2):

- `id`: Numerical code that identifies every `<realization>`.
- `func`: Syntactic functions that define every argument of the argument order schema. In Figure 2, the argument schema is composed by subject (`subj`), preverbal indirect object (`iobj`) and postverbal direct object (`dobj`).
- `cat`: Grammatical categories, phrases or clauses that define every argument of the argument order schema. For example, the three arguments of Figure 2 are realized as noun phrases (`np`).
- `parent`: Lemma of the upper level node of the argument order schema. In the case illustrated in Figure 2, the parent corresponds to the lemma of the verbal form of the test case (i.e. 'vendre'-'to sell').
- `children`: Lemmas of the lower level nodes of the argument order schema. In the test case of Figure 2, the children are the head of every argument (i.e. 'col·leccionista'-'collector', 'ell'-'him', 'llibre'-'book').
- `constr`: Construction type where a particular argument order schema occurs (active, passive, pronominal passive, impersonal, pronominal impersonal). In Figure 2, the construction is in active voice.
- `sbjtype`: Subject type of a particular argument order schema (semantically full or empty and lexically full or empty). The subject type of Figure 2 is semantically and lexically full so the value is `full`.
- `freq`: Relative frequency of the argument order schema in the SenSem corpus (Fernández and Vàzquez, 2012). The frequency of the ditransitive argument schema in Figure 2 is 0.005176, which means that the realization `subj#iobj#V#dobj` occurs 0.005176 times (in a scale between 0 and 1) in the SenSem corpus.
- `idsensem`: Three random SenSem id sentences have been linked to every ParTes argument order schema.
- `test`: Linguistic test data of the described realization of the argument order schema (in Figure 2, 'El col·leccionista no li ven el llibre'-'The collector to him do not sell the book').

The ParTes argument order schemas have been automatically generated from the syntactic patterns of the annotations of the SenSem corpus (Fernández and Vàzquez, 2012). Specifically, for every annotated verb

```
<class name="subj#V#obj">
   <schema name="subj#obj#V#obj">
      <realization id="0140" func="subj#iobj#v#dobj" cat="np#np#v#np" parent="vendre"
                   children="col·leccionista#ell#llibre" constr="active" sbjtype="full"
                   freq="0.005176" idsensem="43177#45210#52053"
                   test="El col·leccionista no li ven el llibre"/>
   </schema>
</class>
```

Figure 2: Argument order of ditransitive verbs in ParTesCa

in the corpus, the argument structure has been recognized. This information has been classified into the ParTes argument order schemas. Finally, the most frequent schemas have been filtered and manually reviewed, considering those schemas above the average. The total set of candidates is 62 argument order schemas for Spanish and 46 for Catalan.

5.3 Test data module

ParTes contains a test data set module to evaluate a syntactic tool over the phenomena included in the test suite. For the sentences in the data set, both plain text and syntactic annotations are available.

As mentioned in Section 4, the test data set is controlled in size: ParTesEs data set contains 94 sentences and ParTesCa data set is 99 sentences long. It is also controlled in terms of linguistic phenomena to prevent the interaction with other linguistic phenomena that may cause incorrect analysis. For this reason, test cases are artificially created.

A semi-automated process has been implemented to annotate ParTesEs and ParTesCa data sets. Both data sets have been automatically analyzed by the FreeLing Dependency Parser (Lloberes, Castellón, and Padró, 2010). The dependency trees have been mapped to the CoNLL format (Figure 3) proposed for the shared task on multilingual dependency parsing (Buchholz and Marsi, 2006). Finally, two annotators have reviewed and corrected the FreeLing Dependency Parser mapped outputs.

6 ParTes evaluation

To validate that ParTes is a useful evaluation parsing test suite, an evaluation task has been done. ParTes test sentences have been used to evaluate the performance of Spanish and Catalan FreeLing Dependency Grammars (Lloberes, Castellón, and Padró, 2010). The accuracy metrics have been provided by the CoNLL-X Shared Task 2007 script (Buchholz and Marsi, 2006), in which the syntactic analysis generated by the FreeLing Dependency Grammars (*system output*) are compared to ParTes data sets (*gold standard*).

The global scores of the Spanish Dependency Grammar are 82.71% for LAS[2], 88.38% for UAS and 85.39% for LAS2. Concerning to the Catalan FreeLing Dependency Grammar, the global results are 76.33% for LAS, 83.38% for UAS and 80.98% LAS2.

A detailed observation of the ParTes syntactic phenomena shows that FreeLing Dependency Grammars recognize successfuly the root of the main clause (Spanish: 96.8%; Catalan: 85.86%). On the other hand, subordinate clause recognition is not perfomed as precise as main clause recognition (Spanish: 11%; Catalan: 20%) because there are some limitations to determine the boundaries of the clause, and the node where it should be attached to.

Noun phrase is one of the most stable phrases because it is formed and attached right most of times (Spanish: 83%-100%; Catalan: 62%-100%). On the contrary, prepositional phrase is very unstable (Spanish: 66%; Catalan: 49%) because the current version of the grammars deals with this syntactic phenomenon shallowly.

This evaluation has allowed to determine which FreeLing Dependency Grammars syntactic phenomena are also covered in ParTes (*coverage*), how these syntactic phenomena are performed (*accuracy*) and why these phenomena are performed right/wrong (*qualitative analysis*).

7 Conclusions

The resource presented in this paper is the first test suite in Spanish and Catalan for

[2] *Labeled Attachment Score* (LAS): the percentage of tokens with correct head and syntactic function label; *Unlabeled Attachment Score* (UAS): the percentage of tokens with correct head; *Label Accuracy Score* (LAS2): the percentage of tokens with correct syntactic function label.

```
1    Habrán     haber     VAIF3P0    _    _    2    aux
2    vendido    vender    VMP00SM    _    _    0    top
3    la         el        DA0FS0     _    _    4    espec
4    casa       casa      NCFS000    _    _    2    dobj
5    .          .         Fp         _    _    2    term
```

Figure 3: Annotation of the sentence 'Habrán vendido la casa' ('[They] will have sold the house')

parsing evaluation. ParTes has been designed to evaluate qualitatively the accuracy of parsers.

This test suite has been built following the main trends in test suite design. However, it also adds some new functionalities. ParTes has been conceptualized as a complex structured test suite where every test case is classified in a hierarchy of syntactic phenomena. Furthermore, like the rest of test suites, it is exhaustive, but exhaustiveness of syntactic phenomena is defined in this resource as representativity in corpora and descriptive grammars.

Despite the fact that ParTes is a polyhedral test suite based on the notions of structure and order, there are more foundations in Syntax, such as syntactic functions that currently are being included to make ParTes a more robust resource and to allow more precise evaluation tasks.

In addition, the current ParTes version contains the test data set annotated with syntactic dependencies. Future versions of ParTes may be distributed with other grammatical formalisms (e.g. constituents) in order to open ParTes to more parsing evaluation tasks.

References

Bosque, I. and V. Demonte. 1999. *Gramática Descriptiva de la Lengua Española.* Espasa Calpe, Madrid.

Buchholz, S. and E. Marsi. 2006. CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164.

EAGLES. 1994. Draft Interim Report EAGLES. Technical report, Expert Advisory Group on Language Engineering Standards.

Fernández, A. and G. Vàzquez. 2012. Análisis cuantitativo del corpus SenSem. In I. Elorza, O. Carbonell i Cortés, R. Albarrán, B. García Riaza, and M. Pérez-Veneros, editors, *Empiricism and Analytical Tools For 21st Century Applied Linguistics.* Ediciones Universidad Salamanca, pages 157–170.

Flickinger, D., J. Nerbonne, and I.A. Sag. 1987. Toward Evaluation of NLP Systems. Technical report, Hewlett Packard Laboratories, Cambridge, England. Distributed at the 24th Annual Meeting of the Association for Computational Linguistics (ACL).

Lehmann, S., S. Oepen, S. Regnier-Prost, K. Netter, V. Lux, J. Klein, K. Falkedal, F. Fouvy, D. Estival, E. Dauphin, H. Compagnion, J. Baur, L. Balkan, and D. Arnold. 1996. TSNLP – Test Suites for Natural Language Processing. In *Proceedings of the 16th Conference on Computational Linguistics*, volume 2, pages 711–716.

Lloberes, M., I. Castellón, and L. Padró. 2010. Spanish FreeLing Dependency Grammar. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 693–699.

McEnery, T. and A. Wilson. 1996. *Corpus Linguistics.* Edinburgh University Press, Edinburgh.

Peñas, A., R. Álvaro, and F. Verdejo. 2006. SPARTE, a Test Suite for Recognising Textual Entailment in Spanish. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 3878 of *Lecture Notes in Computer Science.* Springer, Berlin Heidelberg, pages 275–286.

Solà, J., M.R. Lloret, J. Mascaró, and M. Pérez-Saldanya. 2002. *Gramàtica del Català Contemporani.* Empúries, Barcelona.

Taulé, M., M.A. Martí, and M. Recasens. 2008. AnCora: Multi level annotated corpora for Catalan and Spanish. In *6th International Conference on Language Resources and Evaluation, Marrakesh*, pages 96–101.

Suitability of ParTes Test Suite for Parsing Evaluation

Marina Lloberes
U. de Barcelona
Barcelona, Spain
mllobesa8@alumnes.ub.edu

Irene Castellón
U. de Barcelona
Barcelona, Spain
icastellon@ub.edu

Lluís Padró
U. Politècnica de Catalunya
Barcelona, Spain
padro@cs.upc.edu

Abstract

Parsers have evolved significantly in the last decades, but currently big and accurate improvements are needed to enhance their performance. ParTes, a test suite in Spanish and Catalan for parsing evaluation, aims to contribute to this situation by pointing to the main factors that can decisively improve the parser performance.

1 Introduction

Parsing has been a very active area, so that parsers have progressed significantly over the recent years (Klein and Manning, 2003; Collins and Koo, 2005; Nivre et al., 2006; Ballesteros and Nivre, 2012; Bohnet and Nivre, 2012; Ballesteros and Carreras, 2015). However, nowadays significant improvement in parser performance needs extra effort.

A deeper and detailed analysis of the parsers performance can provide the keys to exceed the current accuracy. Tests suites are a linguistic resource which makes it possible this kind of analysis and which can contribute to highlight the key issues to improve decisively the Natural Language Processing (NLP) tools (Flickinger et al., 1987; EAGLES, 1994; Lehmann et al., 1996).

This paper presents ParTes 15.02, a test suite of syntactic phenomena for parsing evaluation. This resource contains an exhaustive and representative set of structure and word order phenomena for Spanish and Catalan languages (Lloberes et al., 2014). The new version adds a development data set and a test data set.

The rest of the paper describes the main contributions in test suite development (Section 2). Section 3 shows the characteristics and the specifications of ParTes. The results of an evaluation task of the FreeLing Dependency Grammars (FDGs) with verb subcategorization information added (Lloberes et al., 2010) using ParTes are discussed in Section 4. Finally, the main conclusions and future work are exposed (Section 5).

Features	HP	EAGLES	TSNLP
Domain	general	specific	general
Goal	parsing	grammar checkers	NLP software
Languages	English	English	English, German, French
Annotation	minimal	minimal	robust
Content	syntax	taxonomy of errors	(extra-)linguistic

Table 1: HP, EAGLES & TSNLP features

2 Test suite development

The main aim of qualitative studies is to offer empirical evidence about the richness and precision of the data, in comparison with quantitative studies which provide a view of the actual spectrum (McEnery and Wilson, 1996). For this reason, qualitative analysis are deep and detail-oriented, while quantitative analysis focus on statistically informative data. In the qualitative approach, representativeness of the studied phenomena focuses on exhaustiveness rather than frequency, which is the base of the quantitative approach. Both approaches are not exclusive because they contribute to build a global interpretation.

While corpora are a large databases of the most frequent linguistic utterances (McEnery and Wilson, 1996), test suites are controlled and exhaustive databases of linguistic utterances classified by linguistic features. These collections of cases are internally organized and richly annotated (Lehmann et al., 1996). Controlledness, exhaustiveness and detailedness properties allow these databases to provide qualitatively analyzed data.

They were developed in parallel with the NLP technologies. The more sophisticated the software became, the more complex the test suites evolved to be (Lehmann et al., 1996). From a collection of interesting examples, they transformed into deeply structured and richly annotated databases (Table 1), such as the HP test suite (Flickinger et al., 1987), the test suite developed by one of the groups of EAGLES (EAGLES, 1994), the TSNLP (Lehmann et al., 1996) and the corpus of unbounded depdendencies (Rimell et al., 2009).

Concerning the languages of this study, a test suite for Spanish was developed by Marimon et al. (2007). The goal of this test suite is to assess the development of a Spanish Head-Driven Phrase Structure Grammar and it offers grammatical and agrammatical test cases.

```
<level name="intrachunk">
   <constituent name="nounphrase">
      <hierarchy name="child">
         <realization id="0037"
            name="prepositionalphrase"
            class="noun" subclass="prepobj"
            link="n-s" freq="0.084357"
            parent_devel="recurso"
            child_devel="para"
            parent_test="libro"
            child_test="para"
            devel="Es un recurso para los
            alumnos"
            test="Los alumnos tienen un
            libro para la lectura"/>
      </hierarchy>
   </constituent>
</level>
```

Figure 1: Structure in ParTes. Example of the PP-attachment in the noun phrase

3 The ParTes test suite

This test suite is a hierarchically structured and richly annotated set of of syntactic phenomena for qualitative parsing evaluation available in Spanish (ParTesEs) and Catalan (ParTesCa) and freely distributed under the Creative Commons Attribution-ShareAlike 3.0 Unported License.[1]

The new release of ParTes (15.02) consists in the improvement of the linguistic data sets. Initially, ParTes included a test data module formed by sentences illustrating the syntactic phenomena of the test suite (Lloberes et al., 2014). The current version incorporates a set of linguistic data for development purposes that extends the capabilities of the test suite by allowing the parser development monitoring and a second iteration of the evaluation task.

This resource has been created following the main contributions in test suite design (Flickinger et al., 1987; EAGLES, 1994; Lehmann et al., 1996). The main feature shared with the existent test suites is the control over the data, which makes it possible to work as a qualitative evaluation tool. Furthermore, ParTes adds the concepts of complexity of the resource organization, exhaustiveness of the phenomena descriptions and representativity of the phenomena included.

ParTes is a test suite of syntactic phenomena annotated with syntactic and meta-linguistic information. The content has been hierarchically structured by means of syntactic features and over two major syntactic concepts (Figures 1 and 2): structure and word order.

It provides an exhaustive description of the syntactic phenomena, offering a detailed view of their features and their behavior. A selection of the representative phenomena has been performed, which allowed to delimit the number of cases preserving the control over the data.

The test suite has been semi-automatically generated, extracting automatically data from computational resources when available. Otherwise, written linguistic resources have been used to populate manually the resource. Its architecture makes it possible to extend the test suite to new languages, although the current version is available in two languages.

3.1 Test suite specifications

The current version contains a total of 161 syntactic phenomena in ParTesEs (99 relate to syntactic structure and 62 to word order) and a total of 145 syntactic phenomena in ParTesCa (99 concern to syntactic structure and 46 to word order).

The structure phenomena have been manually collected from descriptive grammars (Bosque and Demonte, 1999; Solà et al., 2002) and represented following the criteria of the FDGs (Lloberes et al., 2010). The selection of phenomena has been validated by the dependency links frequency of the AnCora Corpus (Taulé et al., 2008).

As Figure 1 shows, the first level of the hierarchy determines the *level* of the syntactic phenomenon (inside a chunk or between a marker and the subordinate verb). The second level expresses the phrase or the clause involved in the syntactic phenomenon (*constituent*) and the third level describes the position (*head* or *child*) in the *hierarchy*. Finally, a set of syntactic features describes the type of constituent observed (*realization*).

Specifically, the syntactic features of the *realization* concern to the grammatical category, the

[1] http://grial.uab.es/recursos.php

```
<class name="subj#V">
   <schema name="subj#V">
      <realization id="0104"
         func="subj#v"
         cat="pron#v"
         parent="perdre"
         children="tot"
         constr="passive-pron"
         sbjtype="full"
         freq="0.001875"
         idsensem="45074#45239#48770"
         test="Tot s'ha perdut"/>
   </schema>
</class>
```

Figure 2: Word order in ParTes. Example of pronominal passive with particle 'se'

phrase or the clause that defines the structure phenomenon (*name*), its syntactic specifications (*class*, *subclass*), the arch between the parent and the child (*link*), the occurrence frequency of the link (*freq*) in the AnCora Corpus. Additionally, every phenomenon is identified with a numeric *id*.

For every syntactic structure phenomenon, two linguistic examples have been manually defined, one of them to be used for development purposes (*devel*) and the other one for testing purposes (*test*). The lemmas of the parent and the child of the exemplified phenomenon are also provided (*parent_devel*, *parent_test*, *child_devel*, *child_test*).

Word order in ParTes is semi-automatically built from the most frequent argument structure frames of the SenSem Corpus (Fernández and Vàzquez, 2014).

The hierarchy about the word order is structured firstly by the number and the type of arguments of the word order schema (*class*), as Figure 2 illustrates. Every class is defined by a set of *schemas* about the number of arguments and their order. The most concrete level (*realization*) describes the properties of the schema.

These properties refer to the syntactic function (*func*)[2] and the grammatical category (*cat*) of every argument of the schema. Furthermore, the type of construction (*constr*) where the schema occurs in and the type of subject (*sbjtype*) are provided. The occurrence frequency of the schema in the SenSem Corpus is associated (*freq*). In addition, a numeric *id* is assigned to every schema and a link to SenSem Corpus sentences with the same schema is created (*idsensem*).

Every schema recorded is exemplified with a sentence for testing purposes (*test*). For every test sentence, the lemmas of the *parent* and the *children* corresponding to the head of the arguments of the schema are added.

3.2 Description of the data sets

The development and the test data are built over the manually defined linguistic examples of the syntactic phenomena of ParTes.

The sentences have been automatically annotated by using the FDGs, so that a complete dependency analysis of the whole sentence is offered. The output has been reviewed manually by two annotators: a native in Spanish responsible for the annotation of ParTesEs and a native in Catalan who annotated the ParTesCa. A second manual revision has been performed: the Catalan annotator reviewed the ParTesEs annotated and the Spanish annotator reviewed the ParTesCa annotated guaranteeing the agreement between the annotations in both languages and preserving the quality of the annotation according to the criteria.

Up to the current version, the number of sentences referring to the syntactic structure are: 95 sentences in the ParTesEs development data set, 99 sentences in the ParTesEs test data set, 98 sentences in the ParTesCa development data set and 99 sentences in the ParTesCa test data set. The data sets are distributed in plain text format and in the CoNLL annotation format (Nivre et al., 2007).

4 Evaluation task

In order to test the usability of ParTes for parsing evaluation, it has been applied as a gold standard in an evaluation task of the FDGs. Particularly, the capabilities of the test suite have been tested for explaining the performance of FDG as regards the argument recognition since it still remains to be solved successfully (Carroll et al., 1998; Zeman, 2002; Mirroshandel et al., 2013).

The FDGs are the core part of the rule-based FreeLing Dependency Parser (Padró and Stanilovsky, 2012). They provide a deep and complete syntactic analysis in the form of dependencies. The grammars are a set of manually-defined rules that comple the structure of the tree (*linking rules*) and assign a syntactic function to every link of the tree (*labelling rules*) by means of a system of priorities and a set of conditions.

Two FDGs versions for both languages have been evaluated: a version without verb subcategorization classes (*Bare*) and a version with verb sub-

[2] Tagset: *adjt* - adjunct; *attr* - attribute; *dobj* - direct object; *iobj* - indirect object; *pobj* - prepositional object; *pred* - predicative; *subj* - subject.

	ParTesEs		ParTesCa	
Metric	**Bare**	**Subcat**	**Bare**	**Subcat**
LAS	77.57	79.66	79.41	81.80
UAS	88.21	88.21	88.24	88.24
LA	78.90	81.94	80.88	83.64

Table 2: Label Accuracy of FDG on ParTes

categorization classes (*Subcat*) extracted from the verbal frames of the SenSem Corpus (Fernández and Vàzquez, 2014). The system analysis built for every version of the grammars is compared to the ParTes analysis using the evaluation metrics of the CoNLL-X Shared Task (Nivre et al., 2007).[3]

According to the accuracy results (Table 2), the evaluation with ParTes shows that FDGs performance is medium-accuracy (near or above 80% in LAS). Both versions of the grammar in both languages perform in high-accuracy in terms of attachment (UAS), whereas they obtain medium accuracy on syntactic function labelling (LA). ParTes data highlight that the *Subcat* grammar scores better than the *Bare* grammar in LA, which is directly related to the addition of subcategorization classes, as stated in the following discussion.

A detailed observation reveals that ParTes sentences related to subcategorization are performed better in precision by *Subcat* rather than *Bare* (Table 3). Furthermore, the test data allows to show that subcategorization has more impact in the recognition of the majority of arguments (*dobj*, *pobj*, *pred*) and the subject (*subj*) than in the adjuncts (*adjt*) because the precision scores increment is higher. Subcategorization do not have an effect on the attribute (*attr*) because it can be solved lexically. The indirect objects (*iobj*) correspond to cases of dative clitic, which are solved by morphological information.

The integration of subcategorization information bounds the rules to the verbs included in the classes. Consequently, some cases may be not captured if the verb is not expected by the subcategorization classes as it happens in the prepositional object (*pobj*). For example, the prepositional argument of the sentence ‘Ha creido en sí mismo’ (‘He has believed in himself’) should be labelled as *pobj*, but the *adjt* tag is assigned because the verb ‘creer’ is not in any of the prepositional argument classes of the grammar. However, in the majority of types of arguments and the adjuncts the recall is maintained or increased (Table 4).

		ParTesEs			ParTesCa	
Tag	**#**	**Bare**	**Subcat**	**#**	**Bare**	**Subcat**
adjt	39	53.85	65.96	30	60.00	61.90
attr	28	88.89	83.87	20	90.00	78.26
dobj	39	65.31	73.81	42	74.51	86.96
iobj	7	100.00	100.00	3	100.00	75.00
pobj	11	23.68	37.50	13	45.83	60.00
pred	2	25.00	100.00	2	22.22	100.00
subj	51	93.02	93.02	43	87.88	90.62

Table 3: Precision scores of FDG on ParTes

		ParTesEs			ParTesCa	
Tag	**#**	**Bare**	**Subcat**	**#**	**Bare**	**Subcat**
adjt	39	35.90	79.49	30	50.00	86.67
attr	28	85.71	92.86	20	90.00	90.00
dobj	39	82.05	79.49	42	90.48	95.24
iobj	7	28.57	28.57	3	66.67	100.00
pobj	11	81.82	54.55	13	84.62	69.23
pred	2	50.00	50.00	2	100.00	50.00
subj	51	78.43	78.43	43	67.44	67.44

Table 4: Recall scores of FDG on ParTes

5 Conclusions

The new version of the ParTes test suite for parsing evaluation has been presented. The main features and the data sets have been described. In addition, the results of an evaluation task of the FDGs with ParTes data have been exposed.

The characteristics of the test suite made it possible to analyze in detail the causes of the performance improvement on the argument recognition of the FDGs including subcategorization information. Therefore, these results show that ParTes is an appropriate resource for parsing evaluation.

Currently, ParTes is extended to English following the methodology explained in this paper. In the upcoming releases, test and development sentences belonging to the word order will be incorporated in the ParTes data sets. Furthermore, we are exploring a systematic methodology to generate agrammatical variants of the existent sentences.

Acknowledgments

This work has been funded by the SKATeR project (Spanish Ministry of Economy and Competitiveness, TIN2012-38584-C06-06 and TIN2012-38584-C06-01).

[3] *Labeled Attachment Score* (LAS): the percentage of tokens with correct head and syntactic function label; *Unlabeled Attachment Score* (UAS): the percentage of tokens with correct head; *Label Accuracy* (LA): the percentage of tokens with correct syntactic function label; *Precision* (P): the ratio between the system correct tokens and the system tokens; *Recall* (R): the ratio between the system correct tokens and the gold standard tokens.

References

M. Ballesteros and X. Carreras. 2015. Transition-based Spinal Parsing. In *Proceedings of CoNLL-2015*.

M. Ballesteros and J. Nivre. 2012. MaltOptimizer: A System for MaltParser Optimization. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*.

B. Bohnet and J. Nivre. 2012. A Transition-based System for Joint Part-of-speech Tagging and Labeled Non-projective Dependency Parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

I. Bosque and V. Demonte. 1999. *Gramática Descriptiva de la Lengua Española*. Espasa Calpe.

J. Carroll, G. Minnen, and T. Briscoe. 1998. Can Subcategorisation Probabilities Help a Statistical Parser? In *Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora*.

M. Collins and T. Koo. 2005. Discriminative Reranking for Natural Language Parsing. *Computational Linguistics*, 31(1).

EAGLES. 1994. Draft Interim Report EAGLES. Technical report, Expert Advisory Group on Language Engineering Standards.

A. Fernández and G. Vàzquez. 2014. The SenSem Corpus: an annotated corpus for Spanish and Catalan with information about aspectuality, modality, polarity and factuality. *Corpus Linguistics and Linguistic Theory*, 10(2).

D. Flickinger, J. Nerbonne, and I.A. Sag. 1987. Toward Evaluation of NLP Systems. Technical report, Hewlett Packard Laboratories. Distributed at the 24th Annual Meeting of the Association for Computational Linguistics (ACL).

D. Klein and C.D. Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*.

S. Lehmann, S. Oepen, S. Regnier-Prost, K. Netter, V. Lux, J. Klein, K. Falkedal, F. Fouvy, D. Estival, E. Dauphin, H. Compagnion, J. Baur, L. Balkan, and D. Arnold. 1996. TSNLP - Test Suites for Natural Language Processing. In *Proceedings of the 16th Conference on Computational Linguistics*, volume 2.

M. Lloberes, I. Castellón, and L. Padró. 2010. Spanish FreeLing Dependency Grammar. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*.

M. Lloberes, I. Castellón, L. Padró, and E. Gonzàlez. 2014. ParTes. Test Suite for Parsing Evaluation. *Procesamiento del Lenguaje Natural*, 53.

M. Marimon, N. Bel, and N. Seghezzi. 2007. Test-suite Construction for a Spanish Grammar. In *Proceedings of the GEAF 2007 Workshop*.

T. McEnery and A. Wilson. 1996. *Corpus Linguistics*. Edinburgh University Press, Edinburgh.

S.A. Mirroshandel, A. Nasr, and B. Sagot. 2013. Enforcing Subcategorization Constraints in a Parser Using Sub-parses Recombining. In *NAACL 2013 - Conference of the North American Chapter of the Association for Computational Linguistics*.

J. Nivre, J. Hall, J. Nilsson, G. Eryiğit, and S. Marinov. 2006. Labeled Pseudo-projective Dependency Parsing with Support Vector Machines. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*.

J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret. 2007. The CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*.

L. Padró and E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*.

L. Rimell, S. Clark, and M. Steedman. 2009. Unbounded Dependency Recovery for Parser Evaluation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*.

J. Solà, M.R. Lloret, J. Mascaró, and M. Pérez-Saldanya. 2002. *Gramàtica del Català Contemporani*. Empúries.

M. Taulé, M.A. Martí, and M. Recasens. 2008. AnCora: Multi level annotated corpora for Catalan and Spanish. In *6th International Conference on Language Resources and Evaluation, Marrakesh*.

D. Zeman. 2002. Can Subcategorization Help a Statistical Dependency Parser? In *19th International Conference on Computational Linguistics*.

Enhancing FreeLing Rule-Based Dependency Grammars with Subcategorization Frames

Marina Lloberes
U. de Barcelona
Barcelona, Spain
mllobesa8@alumnes.ub.edu

Irene Castellón
U. de Barcelona
Barcelona, Spain
icastellon@ub.edu

Lluís Padró
U. Politècnica de Catalunya
Barcelona, Spain
padro@cs.upc.edu

Abstract

Despite the recent advances in parsing, significant efforts are needed to improve the current parsers performance, such as the enhancement of the argument/adjunct recognition. There is evidence that verb subcategorization frames can contribute to parser accuracy, but a number of issues remain open. The main aim of this paper is to show how subcategorization frames acquired from a syntactically annotated corpus and organized into fine-grained classes can improve the performance of two rule-based dependency grammars.

1 Introduction

Statistical parsers and rule-based parsers have advanced over recent years. However, significant efforts are required to increase the performance of current parsers (Klein and Manning, 2003; Nivre et al., 2006; Ballesteros and Nivre, 2012; Marimon et al., 2014).

One of the linguistic phenomena which parsers often fail to handle correctly is the argument/adjunct distinction (Carroll et al., 1998). For this reason, the main goal of this paper is to test empirically the accuracy of rule-based dependency grammars working exclusively with syntactic rules or adding subcategorization frames to the rules.

A number of studies shows that subcategorization frames can contribute to improve parser performance (Carroll et al., 1998; Zeman, 2002; Mirroshandel et al., 2013). Particularly, these studies are mainly concerned with the integration of subcategorization information into statistical parsers.

The list of studies about rule-based parsers integrating subcategorization information is also extensive (Lin, 1998; Alsina et al., 2002; Bick, 2006; Calvo and Gelbukh, 2011). However, they do not explicitly relate the improvements in parser performance to the addition of subcategorization.

This paper analyses in detail how subcategorization frames acquired from an annotated corpus and distributed among fine-grained classes increase accuracy in rule-based dependency grammars.

The framework used is that of the FreeLing Dependency Grammars (FDGs) for Spanish and Catalan, using enriched lexical-syntactic information about the argument structure of the verb. FreeLing (Padró and Stanilovsky, 2012) is an open-source library of multilingual Natural Language Processing (NLP) tools that provide linguistic analysis for written texts. The FDGs are the core of the FreeLing dependency parser, the Txala Parser (Atserias et al., 2005).

The remainder of this paper is organized as follows. Section 2 contains an overview of previous work related to this research. Section 3 presents the rule-based dependency parser used and the Spanish and Catalan grammars. Section 4 describes the strategy followed initially to integrate subcategorization into the grammars and how this information has been redesigned. Section 5 focuses on the evaluation and the analysis of several experiments testing versions of the grammars including or discarding subcategorization frames. Finally, the main conclusions and the further research goals arisen from the results of the experiments are exposed in Section 6.

2 Related Work

There has been an extensive research on parser development, and most approaches can be classified as *statistical* or *rule-based*. In the former, a statistical model learnt from annotated or unannotated texts is applied to build the syntactic tree (Klein and Manning, 2003; Collins and Koo, 2005; Nivre et al., 2006; Ballesteros and Nivre, 2012), whereas the latter uses hand-built grammars to guide the parser in the construction of the tree (Sleator and Temperley, 1991; Järvinen and Tapanainen, 1998; Lin, 1998).

Concerning the languages this study is based on, some research on Spanish has been performed from the perspective of Constraint Grammar (Bick, 2006), Unification Grammar (Ferrández and Moreno, 2000), Head-Driven Phrase Structure Grammar (Marimon et al., 2014), and Dependency Grammar for statistical parsing, both supervised (Carreras et al., 2006) and semi-supervised (Calvo and Gelbukh, 2011). For Catalan, a rule-based parser based on Constraint Grammar (Alsina et al., 2002) and a statistical dependency parser (Carreras, 2007) are available.

Despite the huge achievements in the area of parsing, argument/adjunct recognition is still a linguistic problem in which parsers still show low accuracy and in which there is still no generalized consensus in Theoretical Linguistics (Tesnière, 1959; Chomsky, 1965). This phenomenon refers to the subcategorization notion, which corresponds to the definition of the type and the number of arguments of a syntactic head.

The acquisition of subcategorization frames from corpora is one of the strategies for integrating information about the argument structure into a parser. Depending on the level of language analysis of the annotated corpus, two main strategies are used in automatic acquisition.

If the acquisition is performed over a *morphosyntactically annotated text*, the subcategorization frames are inferred by applying statistical techniques on morphosyntactically annotated data (Brent, 1993; Manning, 1993; Korhonen et al., 2003).

Alternatively, acquisition can be performed with *syntactically annotated texts* (Sarkar and Zeman, 2000; O'Donovan et al., 2005; Aparicio et al., 2008). Subcategorization acquisition can be performed straightforwardly because the information about the argument structure is available in the corpus. Therefore, this approach generally focuses on the methods for subcategorization frames classification.

The final classification in a lexicon of frames is a computational resource for several NLP tools. In the framework which this research focuses on, the integration of the acquired subcategorization is orientated to the contribution towards building the syntactic tree when the parser has incomplete information to make a decision (Carroll et al., 1998).

Depending on the characteristics of the parser, subcategorization assists in this task in a different way. Subcategorization information can be used to assign a probability to every possible syntactic tree and to rank them in parsers that perform the whole set of possible syntactic analysis of a particular sentence (Carroll et al., 1998; Zeman, 2002; Mirroshandel et al., 2013).

In contrast, subcategorization may help to restrict the application of certain rules. Then, when the parser detects the subcategorization frame in the input sentence, it labels the syntactic tree according to the frame discarding any other possible analysis (Lin, 1998; Calvo and Gelbukh, 2011).

3 Dependency Parsing in FreeLing

The rule-based dependency grammars presented in this article are the core of the Txala Parser (Atserias et al., 2005), the NLP module in charge of Dependency Parsing in the FreeLing library (Padró and Stanilovsky, 2012).[1]

FreeLing is an open-source project that has been developed for more than ten years. It is a complete NLP pipeline built on a chain of modules that provide a general and robust linguistic analysis. Among the available tools, FreeLing offers sentence recognition, tokenization, named entity recognition, tagging, chunking, dependency parsing, word sense disambiguation, and coreference resolution.

3.1 Txala Parser

The Txala Parser is one of the dependency parsing modules available in FreeLing. It is a rule-based, non-projective and multilingual dependency parser that provides robust syntactic analysis in three steps.

Txala receives the partial syntactic trees produced by the chunker (Civit, 2003) as input. Firstly, the head-child relations are identified using a set of heuristic rules that iteratively decide whether two adjacent trees must be merged, and in which way, until there is only one tree left. Secondly, it is converted into syntactic dependencies according to Mel'čuk (1988). Finally, each dependency arch of the tree is labelled with a syntactic function tag.

[1] `http://nlp.cs.upc.edu/freeling/`

		Rules	
Language	**Total**	**Linking**	**Labelling**
English	2961	2239	722
Spanish	4042	3310	732
Catalan	2879	2099	780
Galician	178	87	91
Asturian	4438	3842	596

Table 1: Sizes of the FDGs

3.2 FreeLing Dependency Grammars

The current version of FreeLing includes rule-based dependency grammars for English, Spanish, Catalan, Galician and Asturian (see Table 1 for a brief overview of their sizes). In this paper, the Spanish and Catalan dependency grammars are described.

The FDGs follow the linguistic basis of syntactic dependencies (Tesnière, 1959; Mel’čuk, 1988). However, we propose a different analysis for prepositional phrases (preposition-headed), subordinate clauses (conjunction-headed) and coordinating structures (conjunction-headed).

A FDG is structured as a set of manually defined rules which link two adjacent syntactic partial trees (*linking rules*) and assign a syntactic function to every link of the tree (*labelling rules*), according to certain conditions and priority. They are applied based on this priority: at every step, two adjacent partial trees will be attached or will be labelled with a syntactic function tag if their rule is the highest ranked for which all the conditions are met.

Linking rules can contain four kind of conditions, regarding morphological (part-of-speech tag), lexical (word form, lemma), syntactic (syntactic context, syntactic features of lemmas) and semantic features (semantic properties predefined by the user).

For instance, the rule shown in Figure 1 has priority `911`, and states that a sub-tree marked as a subordinate clause (`subord`) whose head is a relative pronoun (`PR`) attached as a child to the noun phrase (`sn`) to its left (`top_left`) when these two consecutive sub-trees are not located to the right of a verb phrase (`!grup-verb_$$`).

```
911 !grup-verb_$$ - (sn,subord{^PR})
    top_left RELABEL -
```

Figure 1: Linking rule for relative clauses

```
grup-verb   dobj
            d.label=grup-sp
            p.class=trans
            d.side=right
            d.lemma=a|al
            d:sn.tonto=Human
            d:sn.tonto!=Building|Place
```

Figure 2: Labelling rule for human direct objects

Concerning the *labelling rules*, the set of conditions that the parent or the child of the dependency must meet may refer to morphological (part-of-speech tag), lexical (word form, lemma), syntactic (lower/upper sub-tree nodes, syntactic features of lemmas) and semantic properties (EuroWordNet Top Concept Ontology -TCO- features, WordNet Semantic File, WordNet Synonyms and Hypernyms and other semantic features predefined by the user).

In the rule illustrated in Figure 2, the direct object label (`dobj`) is assigned to the link between a verbal head (`grup-verb`) and a prepositional phrase (`grup-sp`) child when the head belongs to the transitive verbs class (`trans`) and the child is post-verbal (`right`), the preposition is `a` (or the contraction `al`), and the nominal head inside the prepositional phrase has the TCO feature `Human` but not (`!=`) the features `Building` or `Place` (to prevent organizations from being identified as a direct object).

4 CompLex-VS lexicon for Parsing

Following the hypothesis that subcategorization frames improve the parsing performance (Carroll et al., 1998), the first version of FDGs included verbal and nominal frames in order to improve argument/adjunct recognition and prepositional attachment (Lloberes et al., 2010). In this paper, only the verbal lexicon is presented because it is the resource used for the argument/adjunct recognition task in the grammars.

4.1 Initial CompLex-VS lexicon in FDGs

The initial Computational Lexicon of Verb Subcategorization (CompLex-VS) was automatically extracted from the subcategorization frames of the SenSem Corpus (Fernández and Vàzquez, 2014), which contains 30231 syntactically and semantically annotated sentences per language, and of the Volem Multilingual Lexicon (Fernández et al., 2002), which has 1700 syntactically and semantically annotated verbal lemmas per language. The patterns extracted from both resources are organized according to the linguistic-motivated classification proposed by Alonso et al. (2007).

The final lexicon applied to the FDGs has 11 subcategorization classes containing a total of 1314 Spanish verbal lemmas and 847 Catalan verbal lemmas with a different subcategorization frame.

A first experimental evaluation of the Spanish Grammar with the initial subcategorization lexicon (Lloberes et al., 2010) showed that incorporating subcategorization information is promising.

4.2 Redesign of the CompLex-VS lexicon

According to the evaluation results of the grammars with the initial CompLex-VS included, the lexicon has been redesigned, proposing a set of more fine-grained subcategorization frame classes in order to represent verb subcategorization in the dependency rules in a controlled and detailed way.

New syntactic-semantic patterns have been extracted automatically from the SenSem Corpus according to the idea that every verbal lemma with a different subcategorization frame expresses a different meaning. Therefore, a new lexicon entry is created every time an annotated verbal lemma with a different frame is detected.

The CompLex-VS contains 3102 syntactic patterns in the Spanish lexicon and 2630 patterns in the Catalan lexicon (see Section 4.3 for detailed numbers). They are organized into 15 subcategorization frames as well as into 4 subcategoriztion classes. The lexicon is distributed in XML format under the Creative Commons Attribution-ShareAlike 3.0 Unported License.[2]

Certain patterns have been discarded because they are non-prototypical in the corpus (e.g. clitic left dislocations), they alter the sentence order (e.g. relative clauses), or they involve controversial argument classes (e.g. prepositional phrases seen as arguments or adjuncts depending on the context).

As Figure 3 shows, the extracted patterns (`<verb>`) have been classified into `<frame>` classes according to the whole set of argument structures occurring in the corpus (`subj` for intransitive verbs, `subj,dobj` for transitive verbs, etc.). Simultaneously, frames have been organized in `<subcategorization>` classes (monoargumental, biargumental, triargumental and quatriargumental).

```
<subcategorization
    class="monoargumental"
    ref="1" freq="0.188480">
    <frame class="subj" ref="1"
        freq="0.188480">
        <verb lemma="pensar"
            id="2531"
            ref="1:1" fs="subj"
            cat="np" rs="exp"
            head="null"
            construction="active"
            se="no" freq="0.000070"/>
    </frame>
</subcategorization>
<subcategorization
    class="biargumental"
    ref="2" freq="0.733349">
    <frame class="subj,dobj" ref="2"
        freq="0.617452">
        <verb lemma="agradecer"
            id="454" ref="2:2" fs="subj,dobj"
            cat="np,complsc" rs="ag_exp,t"
            head="null,null"
            construction="active"
            se="no" freq="0.000140"/>
    </frame>
</subcategorization>
```

Figure 3: Example of the CompLex-VS

Every lexicon entry contains the syntactic function of every argument (`fs`), the grammatical category of the head of the argument (`cat`) and the thematic role (`rs`). The type of `construction` (e.g. active, passive, impersonal, etc.) has been inferred from the predicate and aspect annotations available in the SenSem Corpus.

Two non-annotated lexical items of the sentence have also been inserted into the subcategorization frame because the information that they provide is crucial for the argument structure configuration (e.g. the particle '`se`' and the lexical value of the prepositional phrase `head`).

In addition, meta-linguistic information has been added to every entry: a unique `id` and the relative frequency of the pattern in the corpus (`freq`). A threshold frequency has been established at $7 \cdot 10^{-5}$ (Spanish) and at $8.5 \cdot 10^{-5}$ (Catalan). Patterns below this threshold have been considered marginal in the corpus and they have been discarded.

Every pattern contains a link to the frame and subcategorization class that they belong to (`ref`). For example, if an entry has the reference `1:1`, it means that the pattern corresponds to a monoargumental verb whose unique argument is a subject.

4.3 Integration of CompLex-VS in the FDGs

From the CompLex-VS, two derived lexicons per language containing the verbal lemmas for every recorded pattern have been created to be integrated into the FDGs. The CompLex-SynF lexicon con-

[2] `http://grial.uab.es/descarregues.php`

Frames	Spanish	Catalan
subj	203	386
subj,att	3	7
subj,dobj	440	230
subj,iobj	37	61
subj,pobj	126	93
subj,pred	45	31
subj,attr,iobj	2	1
subj,dobj,iobj	113	72
subj,dobj,pobj	42	34
subj,dobj,pred	21	18
subj,pobj,iobj	2	1
subj,pobj,pobj	14	9
subj,pobj,pred	1	0
subj,pred,iobj	4	5
subj,dobj,pobj,iobj	1	0

Table 2: CompLex-SynF lexicon in numbers

tains the subcategorization patterns generalized by the syntactic function (Table 2). The CompLex-SynF+Cat lexicon collects the syntactic patterns combining syntactic function and grammatical category (adjective/noun/prepositional phrase, infinitive/interrogative/completive clause).

The addition of grammatical categories makes it possible to restrict the grammar rules. For example, a class of verbs containing the verb *quedarse* ('to get') whose argument is a predicative and a prepositional phrase allows the rules to identify that the prepositional phrase of the sentence *Se ha quedado de piedra* ('[He/She] got shocked') is a predicative argument. Furthermore, it allows for discarding the prepositional phrase of the sentence *Aparece de madrugada* ('[He/She] shows up at late night') being a predicative argument, although *aparecer* belongs to the class of predicative verbs but conveying a noun phrase as argument.

While in the CompLex-SynF lexicon the information is more compacted (1054 syntactic patterns classified in 15 frames), in the CompLex-SynF+Cat lexicon the classes are more granular (1356 syntactic patterns organized in 77 frames).

Only subcategorization patterns corresponding to lexicon entries referring to the active voice have been integrated in the FDGs, since they involve non-marked word order. Both lexicons also exclude information about the thematic role, although they take into account the value of the head (if the frame contains a prepositional argument) and the pronominal verbs (lexical entries that accept 'se' particle whose value neither is reflexive nor reciprocal).

Two versions of the Spanish dependency grammar and two versions of the Catalan dependency grammar have been created. One version contains the CompLex-SynF lexicon and the other one the CompLex-Synf+Cat.

Grammar	Spanish	Catalan
Bare	450	508
Baseline	732	780
SynF	872	917
SynF+Cat	869	917

Table 3: Labelling rules in the evaluated grammars

The old CompLex-VS lexicon classes have been replaced with the new ones. Specifically, this information has been inserted in the part of the labelling rules about the syntactic properties of the parent node (observe `p.class` in Figure 2).

Finally, new rules have been added for frames of CompLex-SynF and CompLex-SynF+Cat that are not present in the old CompLex-VS lexicon. Furthermore, some rules have been disabled for frames of the old CompLex-VS lexicon that do not exist in the CompLex-SynF and CompLex-SynF+Cat lexicons (see Table 3 for the detailed size of the grammars).

5 Evaluation

An evaluation task has been carried out to test empirically how the FDGs performance changes when subcategorization information is added or subtracted. Several versions of the grammars have been tested using a controlled annotated linguistic data set.

This evaluation specifically focuses on analysing the results of the experiments qualitatively. This kind of analysis makes it possible to track the decisions that the parser has made, so that it is possible to provide an explanation about the accuracy of the FDGs running with different linguistic information.

5.1 Experiments

Four versions of both Spanish and Catalan grammars are tested in order to assess the differences of the performance depending on the linguistic information added.

- Bare FDG. A version of the FDGs running without subcategorization frames.
- Baseline FDG. A version of the FDGs running with the old CompLex-VS lexicon.
- SynF FDG. A version of the FDGs running with the CompLex-SynF lexicon.

Tag	SenSem Spanish	ParTes Spanish	SenSem Catalan	ParTes Catalan
subj	42.23	34.03	43.03	28.08
dobj	35.77	29.86	34.64	34.25
pobj	16.73	13.89	16.56	17.12
iobj	4.64	6.25	4.70	2.05
pred	0.49	2.08	0.51	0.68
attr	0.14	13.89	0.56	17.81

Table 4: Comparison of the labelling tags distribution in SenSem and ParTes (%)

- SynF+Cat FDG. A version of the FDGs running with the CompLex-Synf+Cat lexicon.

Since this research is focused on the implementation of subcategorization information for argument/adjunct recognition, only the *labelling rules* are discussed in this paper (Table 3). However, metrics related to *linking rules* are also mentioned to provide a general description of the FDGs.

5.2 Evaluation data

To perform a qualitative evaluation, the ParTes test suite has been used (Lloberes et al., 2014). This resource is a multilingual hierarchical test suite of a representative and controlled set of syntactic phenomena which has been developed for evaluating the parsing performance as regards syntactic structure and word order.

It contains 161 syntactic phenomena in Spanish (99 referring to structure and 62 to word order) and 147 syntactic phenomena in Catalan (101 corresponding to structure phenomena and 46 to word order).

The current version of ParTes is distributed with an annotated data set in the CoNLL format. Although this data set is not initially developed for evaluating the argument/adjunct recognition, the number of arguments and adjuncts contained in ParTes is proportional to the number of arguments and adjuncts of the SenSem Corpus (Table 4). Therefore, the ParTes data set is a reduced sample of the linguistic phenomena that occur in a larger corpus, which makes ParTes an appropriate resource for this task.

5.3 Evaluation metrics

The metrics have been computed using the CoNLL-X Shared Task 2007 script (Nivre et al., 2007). The output of the FDGs (*system output*) has been compared to the ParTes annotated data set (*gold standard*).

Tag	Description
adjt	Adjunct
attr	Attribute
dobj	Direct Object
iobj	Indirect Object
pobj	Prepositional Object
pred	Predicative
subj	Subject

Table 5: Tagset of syntactic functions related to the subcategorization

The metrics used to evaluate the performance of the several FDGs versions are the following ones:

Accuracy[3]

$$\text{LAS} = \frac{\text{correct attachments and labellings}}{\text{total tokens}}$$

$$\text{UAS} = \frac{\text{correct attachments}}{\text{total tokens}}$$

$$\text{LAS2} = \frac{\text{correct labellings}}{\text{total tokens}}$$

Precision

$$\text{P} = \frac{\text{system correct tokens}}{\text{system tokens}}$$

Recall

$$\text{R} = \frac{\text{system correct tokens}}{\text{gold tokens}}$$

Both quantitative and qualitative analysis detailed in Section 5.4 pay special attention to the metric LAS2, which informs about the number of heads with the correct syntactic function tag.

Precision and recall metrics of the labelling rules provide information about how the addition of verbal subcategorization information contributes to the grammar performance. For this reason, in the qualitative analysis, only labelling syntactic function tags directly related to verbal subcategorization are considered (Table 5).

5.4 Accuracy results

The global results of the FDGs evaluation (LAS) show that the whole set of evaluated grammars score over 80% accuracy in Spanish (Table 6) and around 80% in Catalan (Table 7).

In the four Spanish grammar versions (Table 6), the correct head (UAS) has been identified in 90.01% of the cases. On the other hand, the tendency changes in syntactic function labelling (LAS2). The *Baseline* establishes that 85.54% of tokens have the correct syntactic function tag.

[3] LAS: Labeled Attachment Score; UAS: Unlabeled Attachment Score; LAS2: Label Accuracy

Grammar	LAS	UAS	LAS2
Bare	81.37	90.01	82.86
Baseline	83.76	90.01	85.54
SynF	84.50	90.01	86.29
SynF+Cat	84.50	90.01	86.29

Table 6: Accuracy scores (%) in Spanish

Grammar	LAS	UAS	LAS2
Bare	78.99	86.84	81.91
Baseline	79.52	86.84	82.85
SynF	81.78	86.84	85.24
SynF+Cat	81.78	86.84	85.24

Table 7: Accuracy scores (%) in Catalan

However, *Bare* drops 2.68 scores and *SynF* and *SynF+Cat* improve 0.75 scores with respect to the baseline.

A parallel behaviour is observed in Catalan, although the scores are slightly lower than in Spanish (Table 7). The four Catalan grammars score 86.84% in attachment (UAS). The *Baseline* scores 82.85% in syntactic function assignment (LAS2). Once again FDGs perform worse without subcategorization information (0.94 points less in *Bare* grammar) and better with subcategorization information (2.39 points more in *SynF* and *SynF+Cat*).

From a general point of view, accuracy metrics show a medium-high accuracy performance of all versions of FDGs in both languages. Specifically, these first results highlight that subcategorization information helps with the syntactic function labelling. However, qualitative results will reveal how subcategorization influences the grammar performance (Sections 5.5 and 5.6).

5.5 Precision results

As observed in the quantitative analysis (Section 5.4), in both languages most of the syntactic function assignments drop in precision when subcategorization classes are blocked in the grammar (Tables 8 and 9), whereas syntactic function labelling tends to improve when subcategorization is available.

For example, the precision of the prepositional object (*pobj*) in both languages drops drastically when subcategorization is disabled (*Bare*). On the contrary, the precision improves significantly when the rules include subcategorization information (*Baseline*). Furthermore, the introduction of more fine-grained frames helps the grammars reach a precision of 94.74% in Spanish and 94.12% in Catalan (*SynF* and *SynF+Cat*). Figure 4 shows this dichotomy.

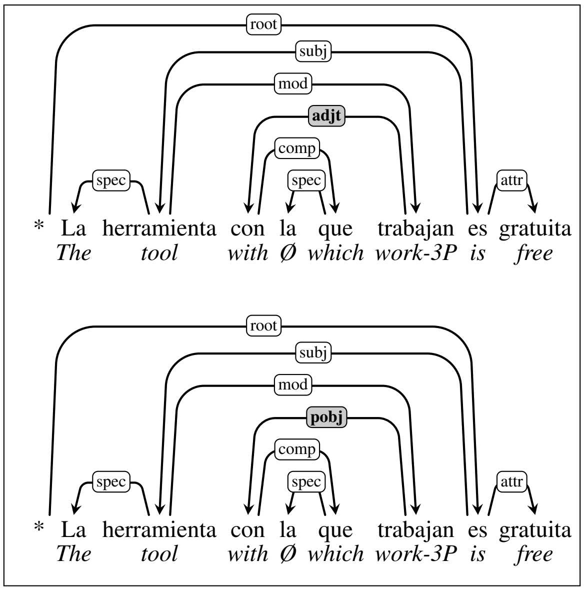

root
subj
mod
adjt
comp
spec
spec
attr
* La herramienta con la que trabajan es gratuita
The tool with Ø which work-3P is free
root
subj
mod
pobj
comp
spec
spec
attr
* La herramienta con la que trabajan es gratuita
The tool with Ø which work-3P is free

Figure 4: Example of *bare* FDGs wrongly labelling a *pobj* as *adjt* (above) and of *SynF* FDGs correctly labelling it (below)

Tag	Bare	Baseline	SynF	SynF+Cat
adjt	59.26	70.27	61.54	61.54
attr	84.21	71.43	71.43	71.43
dobj	78.26	85.71	87.80	87.80
iobj	100.00	100.00	100.00	100.00
pobj	42.50	77.27	94.74	94.74
pred	12.50	0.00	33.33	33.33
subj	90.24	90.91	91.11	91.11

Table 8: Labelling precision scores (%) in Spanish

Despite these improvements, some items differ from the general tendency.

In Spanish, the improvement of the copulative verbs (*attr*) is due to lexical information in the *Bare* FDG, while they keep stable in *SynF* and *SynF+Cat*. Precision remains the same in the indirect object (*iobj*) because morphological information is enough to detect dative clitics in singular.

The performance of predicative (*pred*) in all the grammars is related to the lack or addition of subcategorization. The *Baseline* FDG subcategorization classes do not include the same set of verbs as in the evaluation data. For this reason, a generic rule for capturing predicatives (*Bare* FDG) covers the lack of verbs in a few cases. The improvement of the coverage with new verbs (*SynF* and *SynF+Cat*) shows an increment of the precision.

Adjunct (*adjt*) recognition drops for mislabellings with predicative because of the ambiguity between the participle clause expressing time and a true predicative complement.

Tag	Bare	Baseline	SynF	SynF+Cat
adjt	60.71	61.76	62.50	62.50
attr	95.65	82.14	95.83	95.83
dobj	75.00	83.33	84.78	82.98
iobj	100.00	100.00	100.00	100.00
pobj	61.29	66.67	94.12	94.12
pred	50.00	0.00	100.00	100.00
subj	72.50	71.43	73.81	73.81

Table 9: Labelling precision scores (%) in Catalan

FDGs in Catalan show a parallel behaviour to that in Spanish, but they follow the general tendency in more cases. *SynF* and *SynF+Cat* increase the precision in all the cases, except for the direct object (*dobj*) in *SynF+Cat*. Once more the prepositional object (*pobj*) performance raises when subcategorization frames are available.

Although a drop in all the cases in the *Bare* FDG is expected, the attribute (*attr*) and the predicative (*pred*) increase the precision because of the same reasons as the Spanish grammars.

The results of *SynF* and *SynF+Cat* are almost identical. The analysis of their outputs shows that more fine-grained subcategorization classes including grammatical categories do not have a contribution to the precision improvement.

5.6 Recall results

The addition of subcategorization information in the FDGs also contributes to the improvement, almost in all the cases, in Spanish as well as in Catalan (Tables 10 and 11). The use of FDGs without subcategorization involves a decrease in the recall most of times.

In Spanish, the *Baseline* grammar contains very generic rules to capture adjuncts and more fine-grained subcategorization classes restrict these rules. For this reason, the recall slightly drops in *SynF* and *SynF+Cat*. As observed in the precision metric (Section 5.5), small populated classes related to predicative arguments make recall drop in the baseline. Consequently, generic rules for predicative labelling in the *Bare* grammar and better populated predicative classes in *SynF* and *SynF+Cat* allows a recovery in recall.

FDGs in Catalan show a similar tendency. In the *Bare* grammar, prepositional objects and predicatives are better captured than in the baseline because the lack of subcategorization information allows rules to apply in a more irrestrictive way. On the other hand, the addition of subcategorization information does not seem to help with capturing more direct objects. Lower results are due to some verbs missing.

Tag	Bare	Baseline	SynF	SynF+Cat
adjt	57.14	92.86	85.71	85.71
attr	80.00	100.00	100.00	100.00
dobj	83.72	83.72	83.72	83.72
iobj	33.33	33.33	44.44	44.44
pobj	85.00	85.00	90.00	90.00
pred	33.33	0.00	33.33	33.33
subj	75.51	81.63	83.67	83.67

Table 10: Labelling recall scores (%) in Spanish

Tag	Bare	Baseline	SynF	SynF+Cat
adjt	50.00	61.76	73.53	73.53
attr	84.62	88.46	88.46	88.46
dobj	72.00	80.00	78.00	78.00
iobj	33.33	33.33	33.33	33.33
pobj	76.00	56.00	64.00	64.00
pred	100.00	0.00	100.00	100.00
subj	70.73	73.17	75.61	75.61

Table 11: Labelling recall scores (%) in Catalan

Once again there are no significant differences between *SynF* and *SynF+Cat*, which reinforces the idea that grammatical categories do not provide new information for capturing new argument and adjuncts.

5.7 Analysis of the results

The whole set of experiments demonstrate that subcategorization improves significantly the performance of the rule-based FDGs.

However, some arguments, such as the prepositional object and the predicative, are difficult to capture without subcategorization information. Meanwhile, there are others, such as the attribute, that do not need to be handled with subcategorization classes.

Proper subcategorization information also contributes to capture more arguments and adjuncts. The recall scores are stable among the grammars that use subcategorization information. Secondly, most of these scores are medium-high precision.

Overall, the results show that the new CompLex-VS is a suitable resource to improve the performance of rule-based dependency grammars.

The classification of frames proposed is coherent with the methodology. Furhtermore, it is an essential resource for the grammars tested since it ensures medium-high precision results (compared to medium precision results in the FDGs using the old CompLex-VS). It is important to consider the kind of information to define the subcategorization classes because it can be redundant, such as the combination of syntactic function and grammatical category.

The CompLex-VS lexicon still needs the inclusion of new verbs, since some arguments for verbs missing in the lexicon are not captured properly.

6 Conclusions

This paper presented two rule-based dependency grammars in Spanish and Catalan for the FreeLing NLP library.

Besides the grammars, a new subcategorization lexicon, CompLex-VS, has been designed using frames acquired from the SenSem Corpus. The new frames have been integrated in the argument/adjunct recognition rules of the FDGs.

A set of experiments has been carried out to test how the subcategorization information improves the performance of these grammars.

The results show that subcategorization frames ensure a high accuracy performance. In most cases, the old CompLex-VS frames and the new CompLex-VS frames show an improvement.

However, the increment is more evident in some arguments –such as the prepositional object and the predicative– than others, like the complement in attributive verbs. These results indicate that some arguments necessarily need subcategorization information to be disambiguated, while others can be disambiguated just with syntactic information.

Furthermore, the new frames of CompLex-VS provide better results than the initial ones. Therefore, more fine-grained frames (CompLex-SynF) contribute to raise the accuracy. Despite this evidence, fine-grained classes do not necessarily mean improvement of the parser performance. The most fine-grained lexicon (CompLex-SynF+Cat), which combines syntactic function and grammatical category information, neither improves nor worsens the results of the FDGs.

These conclusions are built on a small set of test data. Although it is a controlled and representative evaluation data set, these results need to be contrasted with a larger evaluation data set.

It would be interesting to evaluate how the parsing performance improves while subcategorization information is added incrementally.

Acknowledgments

This research arises from the research project SKATeR (Spanish Ministry of Economy and Competitiveness, TIN2012-38584-C06-06 and TIN2012-38584-C06-01).

References

L. Alonso, I. Castellón, and N. Tincheva. 2007. Obtaining coarse-grained classes of subcategorization patterns for Spanish. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*.

À. Alsina, T. Badia, G. Boleda, S. Bott, À. Gil, M. Quixal, and O. Valentn. 2002. CATCG: Un sistema de análisis morfosintáctico para el catalán. *Procesamiento del Lenguaje Natural*, 29.

J. Aparicio, M. Taulé, and M.A. Martí. 2008. AnCora-Verb: A Lexical Resource for the Semantic Annotation of Corpora. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*.

J. Atserias, E. Comelles, and A. Mayor. 2005. TXALA un analizador libre de dependencias para el castellano. *Procesamiento del Lenguaje Natural*, 35.

M. Ballesteros and J. Nivre. 2012. MaltOptimizer: A System for MaltParser Optimization. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*.

E. Bick. 2006. A Constraint Grammar-Based Parser for Spanish. In *Proceedings of TIL 2006 - 4th Workshop on Information and Human Language Technology*.

M.R. Brent. 1993. From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax. *Computational Linguistics*, 19(2).

H. Calvo and A. Gelbukh. 2011. *DILUCT: Análisis Sintáctico Semisupervisado Para El Español*. Editorial Academica Espanola.

X. Carreras, M. Surdeanu, and L. Màrquez. 2006. Projective Dependency Parsing with Perceptron. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*.

X. Carreras. 2007. Experiments with a Higher-Order Projective Dependency Parser. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*.

J. Carroll, G. Minnen, and T. Briscoe. 1998. Can Subcategorisation Probabilities Help a Statistical Parser? In *Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora*.

N. Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press.

M. Civit. 2003. Criterios de etiquetación y desambiguación morfosintáctica de corpus en español. In *Colección de Monografías de la Sociedad Española para el Procesamiento del Lenguaje Natural: 8*. Sociedad Española para el Procesamiento del Lenguaje Natural.

M. Collins and T. Koo. 2005. Discriminative Reranking for Natural Language Parsing. *Computational Linguistics*, 31(1).

A. Fernández and G. Vàzquez. 2014. The SenSem Corpus: an annotated corpus for Spanish and Catalan with information about aspectuality, modality, polarity and factuality. *Corpus Linguistics and Linguistic Theory*, 10(2).

A. Fernández, G. Vazquez, P. Saint-Dizier, F. Benamara, and M. Kamel. 2002. The VOLEM Project: A Framework for the Construction of Advanced Multilingual Lexicons. In *Proceedings of the Language Engineering Conference*.

A. Ferrández and L. Moreno. 2000. Slot Unification Grammar and Anaphora Resolution. In N. Nicolov and R. Mitkov, editors, *Recent Advances in Natural Language Processing II. Selected papers from RANLP 1997*. John Benjamins Publishing Co.

T. Järvinen and P. Tapanainen. 1998. Towards an implementable dependency grammar. In *Proceedings of Workshop on Processing of Dependence-Based Grammars, CoLing-ACL'98*.

D. Klein and C.D. Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*.

A. Korhonen, Y. Krymolowski, and Z. Marx. 2003. Clustering Polysemic Subcategorization Frame Distributions Semantically. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.

D. Lin. 1998. Dependency-Based Evaluation of MINIPAR. In *Workshop on the Evaluation of Parsing Systems, First International Conference on Language Resources and Evaluation*.

M. Lloberes, I. Castellón, and L. Padró. 2010. Spanish FreeLing Dependency Grammar. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*.

M. Lloberes, I. Castellón, L. Padró, and E. Gonzàlez. 2014. ParTes. Test Suite for Parsing Evaluation. *Procesamiento del Lenguaje Natural*, 53.

C.D. Manning. 1993. Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*.

M. Marimon, N. Bel, and L. Padró. 2014. Automatic Selection of HPSG-parsed Sentences for Treebank Construction. *Computational Linguistics*, 40(3).

I.A. Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. State U. Press of NY.

S.A. Mirroshandel, A. Nasr, and B. Sagot. 2013. Enforcing Subcategorization Constraints in a Parser Using Sub-parses Recombining. In *NAACL 2013 - Conference of the North American Chapter of the Association for Computational Linguistics*.

J. Nivre, J. Hall, J. Nilsson, G. Eryiğit, and S. Marinov. 2006. Labeled Pseudo-projective Dependency Parsing with Support Vector Machines. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*.

J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret. 2007. The CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*.

R. O'Donovan, M. Burke, A. Cahill, J. Van Genabith, and A. Way. 2005. Large-Scale Induction and Evaluation of Lexical Resources from the Penn-II and Penn-III Treebanks. *Computational Linguistics*, 31(3).

L. Padró and E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*.

A. Sarkar and D. Zeman. 2000. Automatic Extraction of Subcategorization Frames for Czech. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 2*.

D. Sleator and D. Temperley. 1991. Parsing English with a Link Grammar. In *Third International Workshop on Parsing Technologies*.

L. Tesnière. 1959. *Eléments de syntaxe structurale*. Klincksieck.

D. Zeman. 2002. Can Subcategorization Help a Statistical Dependency Parser? In *19th International Conference on Computational Linguistics*.