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).

[1] http://www.lsi.upc.edu/~nlp/freeling/

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

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.

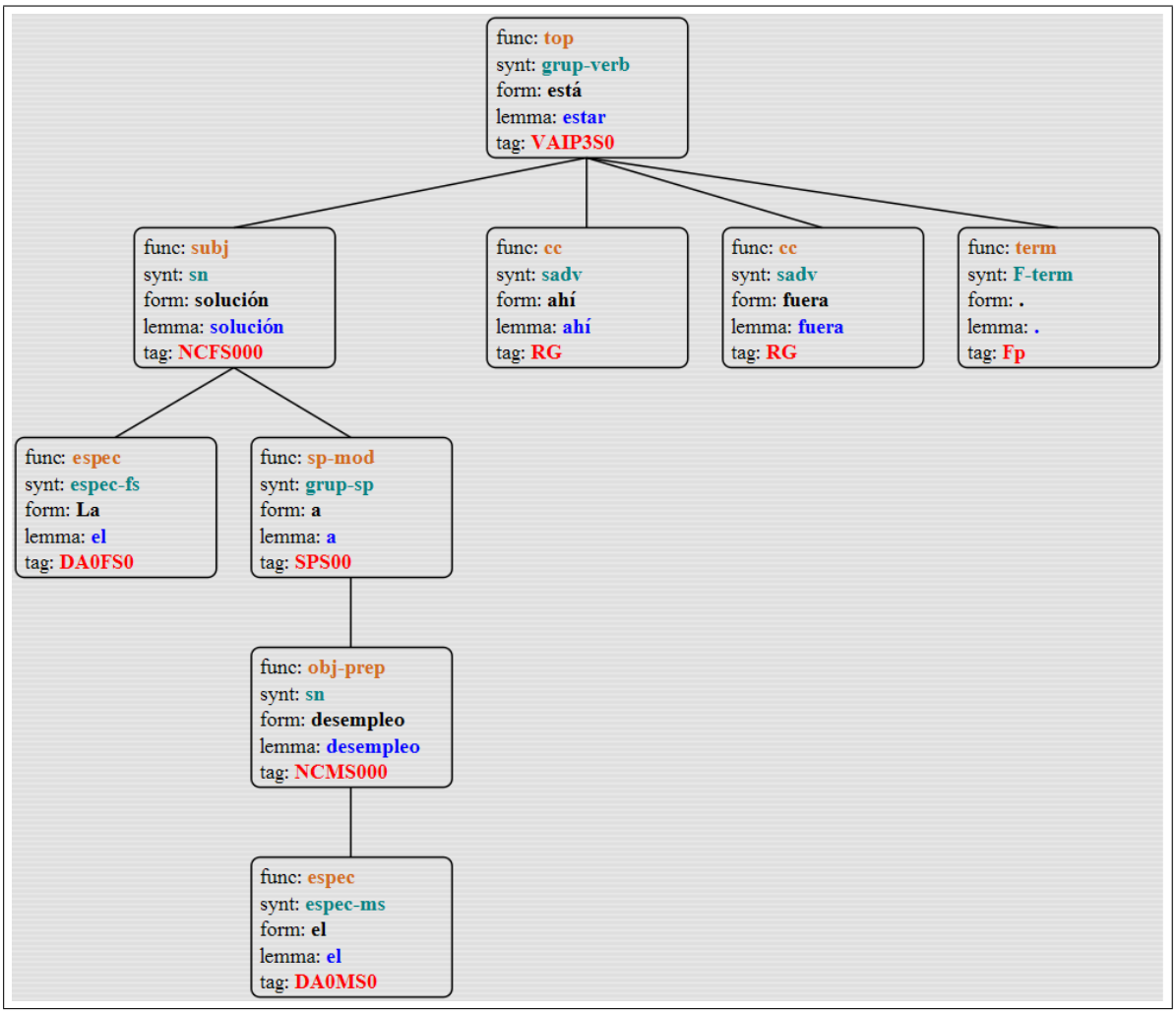

Figure 3: Preposition Phrase Attachment to Noun Phrase – *La solución al desempleo está ahí fuera* ('The solution to unemployment is out there').

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

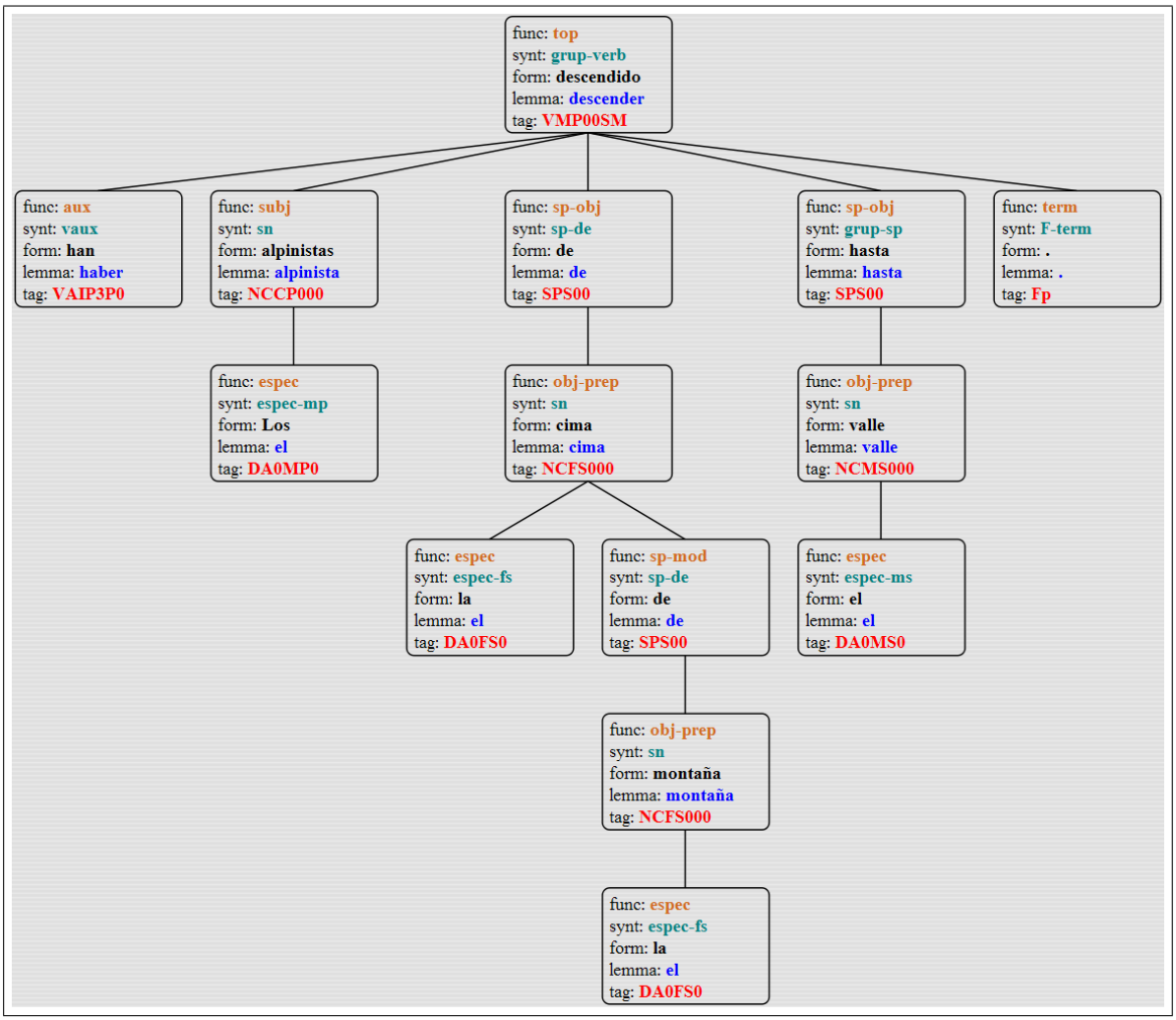

Figure 4: Preposition Phrase Attachment to Verb Phrase – *Los alpinistas han descendido de la cima de la montaña hasta el valle.* ('Climbers descended from the summit to the valley.')

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

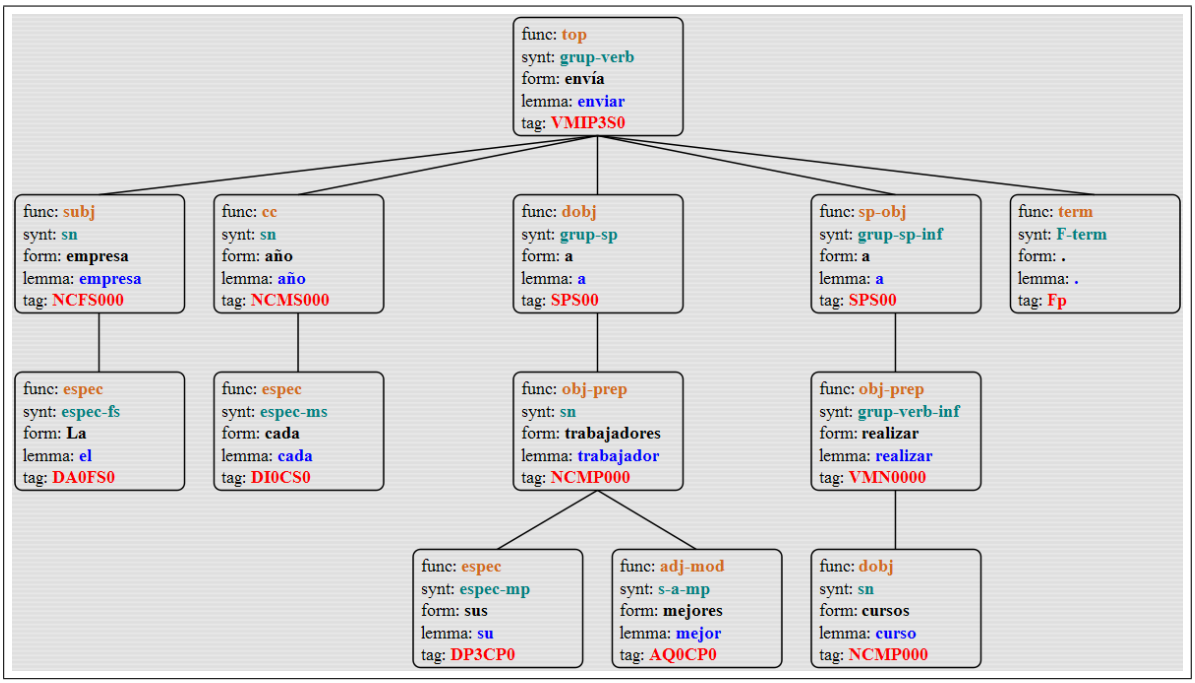

Figure 5: Argument/adjunct recognition – *La empresa envía cada año a sus mejores trabajadores a realizar cursos.* ('The company sends every year their best employees to take courses.')

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

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),

[2]Some conjunctions and relative pronouns appear in AnCoraS and SenSemS, but these occurrences are not considered clause markers.

	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

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-

	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*, vol-

ume 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.