The IULA Treebank

Montserrat Marimon
Universitat de Barcelona
Gran Via de les Corts Catalanes, 585, Barcelona, Spain
E-mail: montserrat.marimon@ub.edu

**Beatriz Fisas, Núria Bel, Blanca Arias, Silvia Vázquez,
Jorge Vivaldi, Sergi Torner, Marta Villegas Mercè Lorente**
Universitat Pompeu Fabra
Roc Boronat, 138, Barcelona, Spain
E-mail: {beatriz.fisas, nuria.bel, blanca.arias, silvia.vazquez,
jorge.vivaldi, sergi.torner, marta.villegas merce.lorente}@upf.edu

Abstract

This paper describes on-going work for the construction of a new treebank for Spanish, The IULA Treebank. This new resource will contain about 60,000 richly annotated sentences as an extension of the already existing IULA Technical Corpus which is only PoS tagged. In this paper we have focused on describing the work done for defining the annotation process and the treebank design principles. We report on how the used framework, the DELPH-IN processing framework, has been crucial in the design principles and in the bootstrapping strategy followed, especially in what refers to the use of stochastic modules for reducing parsing overgeneration. We also report on the different evaluation experiments carried out to guarantee the quality of the already available results.

Keywords: treebank, Spanish, LKB

1. Introduction

We present an on-going project whose aim is to produce a rich annotated corpus for Spanish within the framework of the European project METANET4U (Enhancing the European Linguistic Infrastructure, GA 270893)[1]: the IULA treebank. The initial plan is to annotate 60,000 sentences, distributed among different domains and sentence length, in a period of two years. Planned delivery is end of 2012.

To annotate the corpus, we have pursued the strategy initiated with the LinGO Redwoods treebank for English (Oepen et al., 2004), which was the first project to use the DELPH-IN processing framework[2] for creating a rich annotated language resource in the form of a treebank. This strategy has also been followed in the development of the Hinoki treebank for Japanese (Hashimoto et al., 2007), the CINTIL treebank for Portuguese (Branco et al., 2010), and Tibidabo, a smaller treebank for Spanish based on a corpus from the press (Marimon, 2010b).

The DELPH-IN processing framework offers a range of facilities: (i) the treebanking environment is based on the selection of the correct analysis among all the analyses that are produced by a symbolic grammar instead of using human annotation only; (ii) the use of a stochastic learner that: learns the decisions taken by the annotators and applies the same in unseen parses, and reduces the outputs generated by the grammar with a reduction of the manual annotation effort, especially of long sentences; (iii) finally, the disambiguation decisions can be reused to update the treebank semi-automatically with a revised version of the grammar.

The structure of the paper is as follows. In the following section, we present our target corpus. Section 3 briefly describes the annotation schema; i.e. the DELPH-IN processing framework. Section 4 presents how we have designed the annotation process for achieving our goal. Section 5 reports on the use of the stochastic module for reducing parsing overgeneration, and the experiments carried out to explore the design of the bootstrapping strategy. Section 6 reports on evaluation of the results with the inter-annotator agreement validation exercise carried out. Finally, Section 7 presents the conclusions and future developments.

2. The target corpus

To create the treebank, we chose the Corpus Tècnic de l'IULA, a collection of written texts from the fields of Law, Economy, Genomics, Medicine, and Environment, as well as a contrastive corpus from the press (Vivaldi, 2009, Cabré *et al.*, 2006). This corpus of 1,389 documents contains 31,436,451 words distributed among 412,707 sentences. Figure 1 shows the ratio of number of sentences per sentence length for the different domains. The distribution of the sentences in the corpus is such that the total amount of sentences whose length ranges from 4

[1] http://www.meta-net.eu/projects/METANET4U/.
[2] http://www.delph-in.net.

to 30 words represent the 65,1%. We decided to choose the committed volume for our project (60,000 sentences), at random from the above mentioned range, with the same proportion as the *Corpus Tècnic de l'IULA* in terms of number of sentences per length and domain.

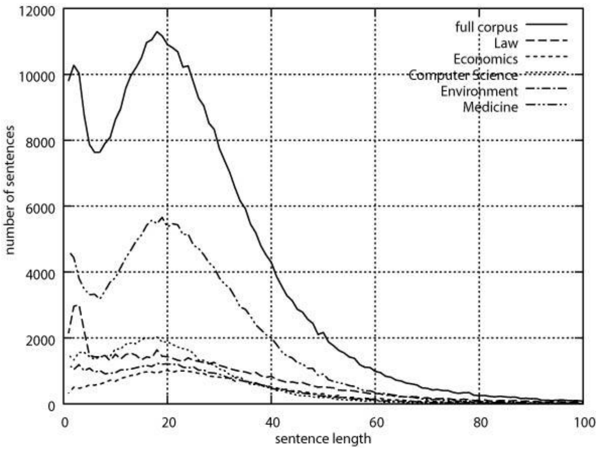


Figure 1: The *Corpus Tècnic de l'IULA.* Ratio of number of sentences per sentence length for each domain

The aim of the IULA Treebank is to contribute to the availability of parsed data in Spanish. Currently the only broadly available treebanks for Spanish are Ancora (Taulé et al., 2008) which contains 500,000 words (about 17,000 sentences) from the press, and UAM Spanish Treebank (Moreno and López, 1999), which contains 1,500 sentences only.

3. The annotation schema

The basic approach for corpus annotation of the DELPH-IN framework is a two-step based annotation. First, the corpus is parsed using a symbolic declarative grammar. In our treebanking work we use the Spanish HPSG grammar SRG (Marimon, 2010a), which consist of 230 syntactic rules, 68 lexical rules, and about 52,000 lexical entries which are defined by 500 lexical types that represent the type of words in the lexicon. As a declarative grammar produces all possible parses, the second step must be the disambiguation of the ambiguous outputs, by manually selecting the correct analysis among those produced by the system. Selection is done by rejecting (or, alternatively, selecting) the lexical items and grammar rules that originate the multiple parses to incrementally disambiguate the sentence until a single analysis is left.

Thus, the bulk of the annotation process is to select only one parse for each sentence. Because the average number of analyses produced for each sentence is typically proportional to its length, only a reduced number of top parses (500 as maximum) are shown to the annotator such that the disambiguation should not require more than 9 decisions (Kordoni and Zhang, 2009).

DELPH-IN is equipped with a stochastic parse ranking learner that learns from the decisions taken by the human annotators (Toutanova et al., 2003). Later, the system ranks unseen sentence parses according to the learned model in order to determine the selection of the 500-best parses for an input sentence.

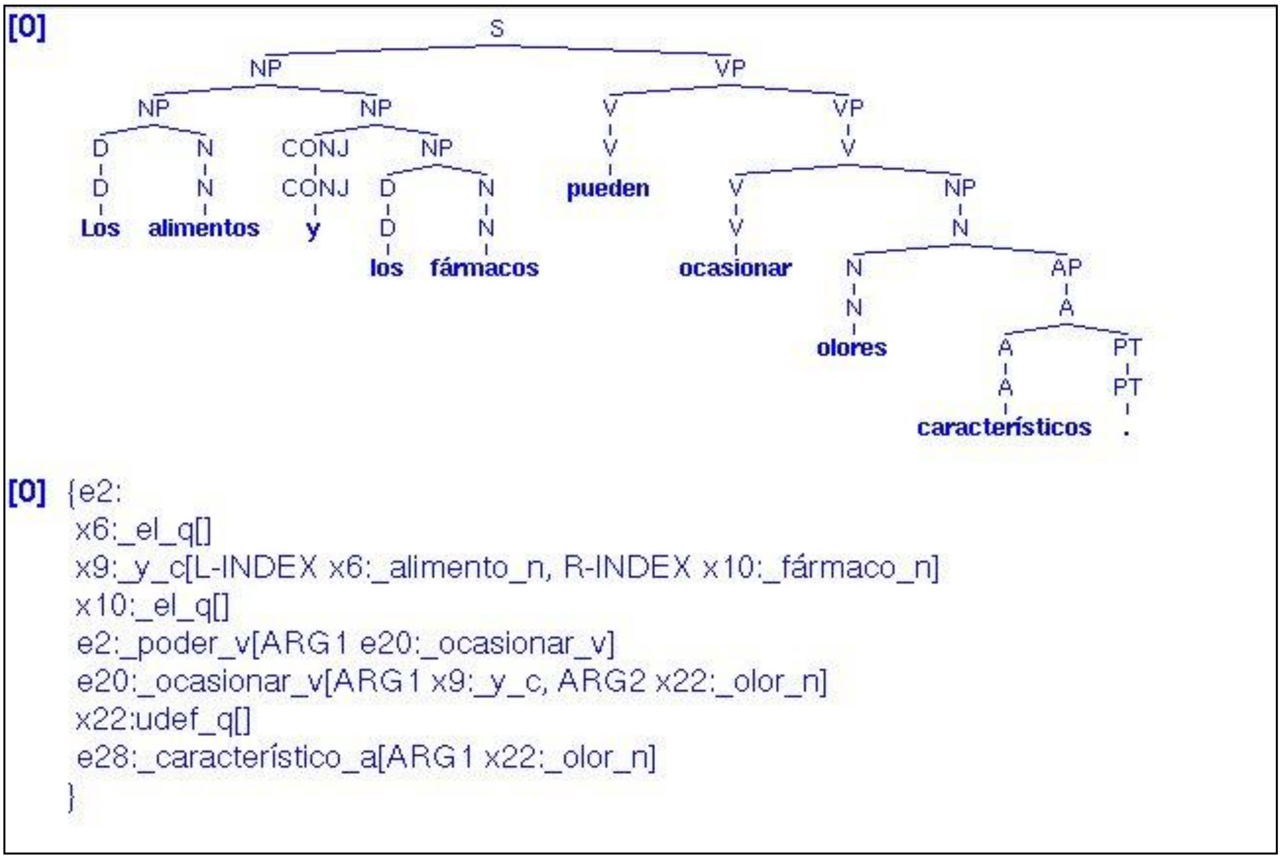
The basic annotation of each parsed sentence with the DELPH-IN processing framework simultaneously displays both a syntactic phrase structure tree and a Minimal Recursion Semantic (MRS) representation (Copestake et al., 2005).[3] A MRS representation is a syntactically flat semantic representation that offers, by means of the labelling of arguments and their co-indexation, a list of semantic relations and a set of syntactic limitations on possible scope relations among them. Figure 2 shows both the syntactic and semantic representations, for the sentence *Los alimentos y los fármacos pueden ocasionar olores característicos* (Food and drugs may produce characteristic odours).

4. Treebank development design principles

By means of the DELPH-IN processing framework, the treebanking task is typically organised into iterations of: parsing with HPSG grammars, manual disambiguation, manual inspection and error analysis, and grammar/treebanking update cycles. While error analysis and treebank update guarantee the quality of the treebank and the accuracy of the parse selection ranking model, they require extra manual annotation effort.

Due to the time limitations of our project, we needed to increase the speed of the treebank development, and we decided to study strategies for gradually reducing the manual annotation workload. To achieve this we designed a bootstrapping approach. The strategy has been to subdivide the whole target corpus into six smaller sub-corpora, with the same proportion as our target corpus in terms of number of sentences per length and domain. Each sub-corpus, therefore, is a representative sample of all the elements involved: complex and simple syntactic phenomena (e.g. *se*-constructions, adverbs' ubiquity), short and long sentences (where coordination structures are frequent), single-verb and multiple verb sentences (subordinate clauses), contextual, and spurious ambiguity. Each sub-corpus, in turn, is distributed among several files according to the number of words in the sentences, thus progressively increasing the complexity of the sentences to be annotated

[3] In addition, dependencies and semantic role labels may also be derived from the sentences parsed using the DELPH-IN framework by running recently developed routines in the framework of the development of the DELPH-IN Portuguese treebank (CINTIL).



```
[0] {e2:
  x6:_el_q[]
  x9:_y_c[L-INDEX x6:_alimento_n, R-INDEX x10:_fármaco_n]
  x10:_el_q[]
  e2:_poder_v[ARG1 e20:_ocasionar_v]
  e20:_ocasionar_v[ARG1 x9:_y_c, ARG2 x22:_olor_n]
  x22:udef_q[]
  e28:_característico_a[ARG1 x22:_olor_n]
 }
```

Figure 2: Phrase structure tree
and MRS representation for a sample Spanish sentence

While we have already annotated 18,000 in an initial annotation cycle, this strategy has allowed us to get a fine-grained diagnosis of the grammar performance in parsing domain specific corpus and to update the grammar accordingly. In addition, it has allowed us to identify the areas in which difficulties may arise and the needs of annotators with regards to the theoretical framework, the DELPH-IN grammar and processing framework, as well as the complexity of the linguistic phenomena shown in the target corpus. In this way, errors (produced both by the grammar and by the annotators), and therefore annotation updates, are gradually reduced.

This strategy should have allowed us to use these manually annotated sub-corpora to incrementally update the stochastic parse ranking learner in order to get as benefit a smaller and ranked number of possible parses for human annotators to decide upon, ideally the right one only. The stochastic system delivers the requested *n*-best parses for a given sentence ranked as a prediction of the likelihood of being the right parse. Thus, the right analysis for each sentence should be in the position zero of the ranking. Figure 3 shows, for a given set of sentences, the number of right parses found from position 0 to 19; in position 20 it accumulates all the results for positions greater or equal to 20. Figure 3 show that most sentences have their right analysis in position zero, even though a number of them has a position that is greater than zero.

We initially planned to periodically generate new stochastic modules using the increasing number of human annotated results, on the assumption that, as the evidence increases, more right analyses are to be found in lower positions and thus a less number of parses would have to be shown to the human annotators.

However, as human annotation was first tackling short sentences, the question was raised whether the learner could perform the same with or without having long sentences as input. In the next section, we report on the experiment carried out to verify the need of including sentences of all lengths in the learning process to get optimized results when tackling the longer ones.

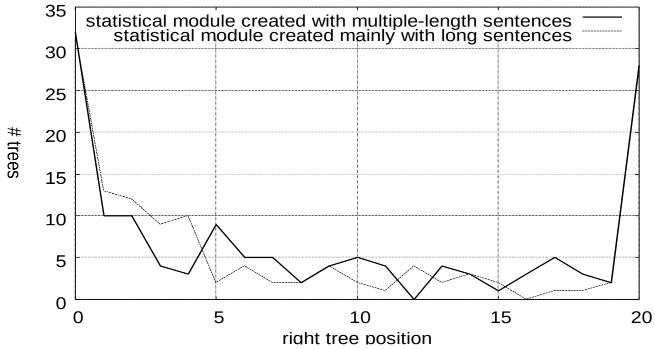


Figure 3: Analysis of the results in the test package

5. Evaluation of the stochastic module

As explained above, the experiment wanted to test whether the module mainly created with long sentences would perform well when analysing long sentences. We have created two stochastic modules: one learning mainly from short sentences and another one learning mainly from long sentences. The numerical data (number of sentences and their length) regarding both stochastic modules are shown in Table 1, while the positions of the right tree in the solution tree list are shown in Table 2.

Stochastic module created with				
multiple length sentences			long sentences mainly	
length	#sentences		length	#sentences
4-7	1334		4-7	541
8-9	4225		8-9	929
10-11	488		10-11	220
12-13	660		12-13	1712
14-15	176		14-15	1539
16-17	571		16-17	1456
18-19	0		18-19	193
Total	7454		Total	6590

Table 1: Profile of the stochastic modules

	Stochastic module created with						
	multiple length sentences				long sentences mainly		
Pos.	Num.	%	Partial		Num.	%	Partial
0	32	22,54	--		32	23,53	--
1	10	7,04	29,58		13	9,53	33,06
2	10	7,04	36,62		12	8,82	41,88
3	4	2,82	39,44		9	6,62	48,50
4	3	2,11	41,55		10	7,35	55,85
…	…	…			…	…	
Total	142				135		

Table 2: Right tree position

The results in Table 2 (and their corresponding graphical representation in Figure 3) show that, after having used the stochastic model mainly trained with long sentences, 55,85% of the sentences have their right analysis in position 5 or lower (against 41,55% reached with the stochastic module trained with short sentences). We measured the relevance of these results by means of the χ^2 test and assuming as null hypothesis that there is no difference between both stochastic packages. We found that χ^2=6,02. Thus, the null hypothesis may be refuted with a confidence greater than 95%.

As a result of this experiment, we improved the bootstrapping process creating periodically the stochastic module only when sentences of different lengths, specially long sentences were available.

6. Treebank evaluation

For guaranteeing the consistency of the resulting treebank we foresaw two main activities: (i) The compilation of annotation guidelines according to the characteristics of the process described in section 3 and (ii) running different inter-annotator agreement evaluation exercises.

Because one difficulty of the annotating process concerns the analysis of some frequent and challenging linguistic phenomena, we concentrated on them. That is, there are some phenomena which are difficult to analyse by human annotators. Some of them are classical problems affecting syntactic annotation among several languages, and others are specific to Romance languages. There are three main areas in which difficulties arise: *se*-constructions (SX), adverbs (ADV) and PP-attachment (PP-ATT). In order to ease the annotation process, guidelines making special emphasis on clarifying the annotation of these phenomena were discussed and agreed. In this section we discuss these more demanding phenomena, and in the next one we describe the interannotator agreement evaluation carried out and the analysis of the results.

6.1 Most frequent and challenging phenomena

Se-constructions are characteristic of Romance languages. The main difficulty which arises from the analysis of such structures is that one form may have many different functions, corresponding to different structures and meanings (Mendikoetxea 1999a, 1999b; Sánchez, 2002; RAE 2009, among others). For instance, (1) may have three readings: unnacusative, passive and impersonal sentence.

(1) El producto resultante se destruye.
(The product waste *se*-pron. destroys)

Out of context, there is no way to choose among these three readings. To overcome this problem, we have taken decisions in the design process about which analysis must be prioritized in case of ambiguity, regardless of other possible meanings. Importantly, as a result of this strategy there are no significant differences among annotators in the analysis of these structures, which contributes to the system robustness.

On the other hand, adverbs are difficult to analyse because they have a complex syntax, as described by many authors (Cinque, 1999; Ernst, 2002, among others). In Spanish, the difficulties increase because there is no strict correlation between word order and function, so the use of an adverb in a sentence usually produces structural ambiguity. Despite the extensive literature on adverbs available in Spanish (Kovacci, 1999; Rodríguez, 2003 and Torner 2007, among others), adverbs cause analysis differences among annotators, especially in relation to scope ambiguities of focus adverbs.

The difficulties in adverb analysis are quite alike to a major problem that has been previously observed in many projects of treebank building —the ambiguities caused by PP-attachment (Toutanova et al., 2003). As in the case of adverbs, prepositional phrases may be attached to many different structural positions, each one carrying a different meaning. In some occasions, ambiguities cannot be solved because of the lack of context, but in some other occasions they would not be solved even if more context was accessible. In our project, PP-attachment ambiguities are the third cause of inter-annotator disagreement.

6.2 Inter-annotator agreement evaluation

In order to evaluate the consistency of the annotation, we have begun a series of inter-annotator agreement tests. Our aim is to detect which is the level of understanding of the annotation guidelines, which are the most frequently disagreed linguistic phenomena, as well as estimating a baseline of the maximum results we can expect from our machine-learning process.

The tests have been done on a series of 100 sentences, all of which have a length of 8-9 words and one verb. Two annotators (A1-with a longer training period, and A2-after a short training period) have annotated independently the same set of sentences.

For the time being, we have done two tests in a two-months time period: IA-test1 took place in December 2011 and IA-test2 in February 2012. The analysis of the results in terms of agreement (same syntactic tree selected) and in terms of the reasons of the disagreement is the following:

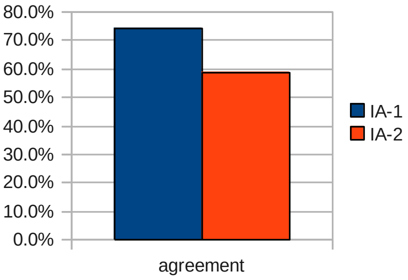


Figure 4: Inter-annotator Agreement

Surprisingly, the agreement percentage in the second test was lower than the first time, and this could only lead us to conclude that our agreement will be in this order of figures (60-75%). In any case, we are aware of the small size of our sample and these percentages must be confirmed in the following inter-annotator tests with series of more sentences. The results of the different types of disagreement between annotators are more clarifying:

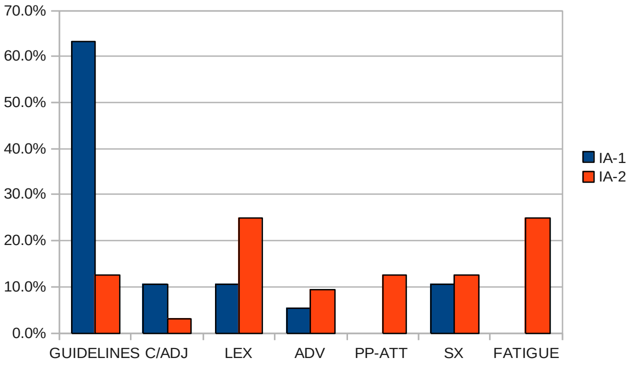


Figure 5: Disagreement Analysis

The fine-grained analysis of which were the reasons of the disagreement allowed the classification of these in seven categories:

a. Misunderstanding of the guidelines of annotation, the meaning of the syntactic rules to select or the disambiguation decisions agreed to apply. (GUIDELINES)
b. The distinction between complements and adjuncts. (C/ADJ)
c. The choice of the appropriate lexical entry of some words in certain contexts. (LEX)
d. The structural position and type of adverbs. (ADV)
e. The inherent linguistic ambiguity of PP-attachment. (PP-ATT)
f. Errors due to complex syntactical issues. (SX)
g. Error due to annotators' fatigue. Annotators easily agree about which would be the correct choice when they comment on these sentences. (FATIGUE)

As it is shown on the previous graph, the errors due to misunderstanding of the guidelines were many more in the first IA test (IA-1), whereas in the second test (IA-2) these have nearly disappeared. In contrast, in IA-2 the fatigue errors appear to be one of the most common, together with those caused by different lexical choices.
Our conclusions to these inter-annotator agreement controls are three-fold:

Which is the information we now have?

a. The inter-annotator agreement percentage up to the moment is around 70%.
b. The understanding of annotation guidelines has improved considerably and the disagreement caused by the inexperience has been reduced. Linguistic ambiguities (PP-attachments), questionable matters (Complement vs Adjuncts), lexical choices and the ubiquity of adverbs in Spanish are the classical and expected reasons of disagreement between annotators. Errors due to fatigue

should not be disregarded.

What can we do to improve our results?

a. Fatigue errors must be reduced by scheduling shorter periods of annotation.

b. Disagreement due to categories b-f are difficult to minimize, but syntax forums must be encouraged in order to improve the common knowledge of annotators.

c. The number of sentences in the test-sets must be increased to obtain more reliable results.

What can we expect from our automatic process?

If human annotators agree in a 70%, as our results seem to indicate, we can consider this as a baseline for our expected results from the machine-learning process with a stochastic module selecting the correct syntactical tree.

7. Conclusions and future work

We have presented on-going work for the creation of a new Spanish treebank based on a technical corpus. The initial plan is to annotate 60,000 sentences distributed among different domains and sentence length before the end of 2012. This ambitious plan is only achievable because we are using the DELPH-IN framework and the previously developed resources for Spanish, and we can work upon the experience of other treebanks that used the same environment. In our opinion it is worth giving the notice of this project in an early stage as to promote the creation of resources and tools that, like in this case, may allow a quick and accurate production of new, demanded, language resources.

Further lines of research are to be developed in order to improve the final results. The actions to be taken refer to different factors that intervene in our process, such as:

- GRAMMAR and LEXICON. We intend to upgrade the coverage of the HPSG grammar for Spanish (HPSG) so as to include some syntactical phenomena which are still not implemented. We will also test the results, in terms of percentage of annotated sentences, if we include specialized terminology in one specific domain.
- STOCHASTIC MODULE EFFICIENCY. We will invest resources in optimizing the efficiency of our stochastic module, by modifying the mix of sentence types included in it. In other words, we want to confirm the results that suggest that long sentences in the stochastic module positively influence the performance of the grammar.
- CORPUS SPECIALISATION. Another area on which we shall experiment is the influence of the corpus specialization degree. We plan to compare the performance of our grammar when annotating a series of sentences from our specialized economy corpus with another series extracted from economic texts published on general press.
- WORKING PROCESS. Our strategy will be to progress simultaneously in depth and breadth of our annotation process. We will therefore build two different teams of annotators: the first one will annotate long sentences and will disambiguate among 300 syntactical trees; in contrast, the second team of inexperienced annotators will annotate shorter or medium sized sentences and will choose among the first five trees offered by the grammar. In this way, the first team will collaborate to include new and infrequent syntax rules in our treebank and stochastic module, while the second team will annotate quicker and contribute to increase the volume of our treebank.
- INTERFACES. We are already working on the interfaces of our output with other software solutions. In such sense we plan to use our Treebank to train a MaltParser.

Acknowledgments

This work was co-funded by the Ramón y Cajal program of the Spanish Ministerio de Ciencia e Innovación, the EU UNER - Competitiveness and Innovation Framework Program, METANET (CIP-PSP-270893), and the UPF-IULA PhD grant program. We wish to thank also to Stephan Oepen for his assistance in using the DELPH-IN environment and Lluís Padró for his assistance in the significance calculation.

References

Branco, A. and F. Costa, and J. Silva, and S. Silveira, and S. Castro, and M. Avelãs, and C. Pinto, and J. Graça (2010) Developing a Deep Linguistic Databank Supporting a Collection of Treebanks: the CINTIL DeepGramBank. In *Proceedings, LREC2010* - The 7th international conference on Language Resources and Evaluation, La Valleta, Malta.

Cabré, M. T., C. Bach, & J. Vivaldi. (2006). 10 anys del Corpus de l'IULA. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra.

Cinque, G. (1999) *Adverbs and Functional Heads. A Crosslinguistic Perspective*. Oxford, Oxford University Press.

Copestake, A. and D. Flickinger, and I. Sag, and C. Pollard (2005). Minimal Recursion Semantics: An Introduction. In *Journal of Research on Language and Computation,* 3(2-3). 281-332.

Ernst, T. (2002) *The Syntax of Adjuncts*. Cambridge, Cambridge University Press.

Hashimoto, C. and F. Bond, and M. Siegel (2007) Semi-automatic documentation of an implemented linguistic grammar augmented with a treebank. In *Language Resources and Evaluation*. (Special issue on Asian language technology).

Kordoni V. and Yi Zhang. 2009. Annotating Wall Street Journal Texts Using a Hand-Crafted Deep Linguistic

Grammar. In *Proceedings of the ACL-IJCNLP 2009 Workshop LAW III* (The Third Linguistic Annotation Workshop), Suntec, Singapore.

Kovacci, O. (1999) «El adverbio». In I. Bosque & V. Demonte (eds.) *Gramática descriptiva de la lengua española.* Madrid, Espasa Calpe. 705-786.

Marimon, M. (2010a). The Spanish Rosurce Grammar. In *Proceedings of the 7th international conference on Language Resources and Evaluation*, La Valleta, Malta.

Marimon, M. (2010b). The Tibidabo Treebank. Procesamiento del Lenguaje Natural, ISSN 1135-5948.

Mendikoetxea, A. (1999a). *Construcciones inacusativas y pasivas*. In I. Bosque & V. Demonte (dirs.) *Gramática descriptiva de la lengua española*. Madrid, Espasa-Calpe. 1575-1630.

Mendikoetxea, A. (1999b). *Construcciones con se: medias, pasivas e impersonales*. In I. Bosque & V. Demonte (dirs.) *Gramática descriptiva de la lengua española. 1.* Madrid, Espasa-Calpe. 1631-1722.

Moreno, A. and S. López (1999). Developing a Spanish Tree Bank. In *Proc. Journées ATALA, Corpus annotés pour la syntaxe*. Paris, 18-19 June 1999.

Oepen, S. and D. Flickinger, and K. Toutanova, and C.D. Manning (2004) LinGo Redwoods. In *Research on Language and Computation*, 2(4), Hinrichs, E.W. and K. Simov (ed).

Pollard, C.J. and I.A. Sag (1994). *Head-driven Phrase Structure Grammar*. The University of Chicago Press and CSLI Publications, Chicago.

Real Academia Española y Asociación de Academias de la Lengua (2009). *Nueva gramática de la lengua española.* Madrid, Espasa-Calpe.

Rodríguez (2003). *La gramática de los adverbios en* mente *o cómo expresar maneras, opiniones y actitudes a través de la lengua.* Madrid, Ediciones de la Universidad Autónoma de Madrid.

Sánchez, C. (ed.) (2002). *Las construcciones con* se. Madrid, Visor.

Taulé, M.; M.A. Martí and M. Recasens (2008). Ancora: Multilevel Annotated Corpora for Catalan and Spanish. *Proceedings of 6th International Conference on Language Resources and Evaluation*. Marrakesh.

Torner, S. (2007) *De los adjetivos calificativos a los adverbios en -mente: semántica y gramática*. Madrid, Visor Libros.

Toutanova, K., Manning, C. D., and Oepen, S. (2003), Stochastic HPSG Parse Selection using the Redwoods Corpus, in *Journal of Logic and Computation*, 2005.

Vivaldi, J. (2009). "Corpus and exploitation tool: IULACT and bwanaNet" in Cantos Gómez, Pascual; Sánchez Pérez, Aquilino (ed.) *A survey on corpus-based research (CICL-09)*, Asociación Española de Lingüística del Corpus. 224-239.