Boosting the creation of a treebank

Blanca Arias[1], Núria Bel[1], Marina Fomicheva[1], Imanol Larrea[1], Mercè Lorente[1]
Montserrat Marimón[1], Alba Milà[1], Jorge Vivaldi[1], Muntsa Padró[2]

[1]Universitat Pompeu Fabra
Roc Boronat 138, 08018-Barcelona, Spain
{blanca.arias,nuria.bel,marina.fomicheva,imanol.larrea,merce.lorente,
montserrat.marimon,alba.mila,jorge.vivaldi}@upf.edu
[2]Institute of Informatics, Federal University of Rio Grande do Sul
Porto Alegre, Brazil
muntsa.padro@inf.ufrgs.br

Abstract

We present the results of the experiment of bootstrapping a Treebank for Catalan by using a Dependency Parser trained with Spanish sentences. In order to save time and cost, our approach was to profit from the typological similarities between Catalan and Spanish to create a first Catalan data set quickly by (i) automatically annotating with a delexicalized Spanish parser, (ii) manually correcting the parses, and (iii) using the Catalan corrected sentences to train a Catalan parser. The results showed that the number of parsed sentences required to train a Catalan parser is about 1000, which were achieved in 4 months with 2 annotators.

Keywords: dependency treebank, treebank bootstrapping, less resourced languages

1. Introduction

Dependency parsing is a formalism for building syntactic representations of natural language sentences. It is based on the idea of modelling the syntactic structure of a sentence as a set of asymmetrical head-dependent relations. The resulting structure can be seen as acyclic graph where all nodes (labelled words/tokens) have a single head with an associated syntactic label (such as *subject*, *object*, etc.), only a single word in the sentence does not have a head dependency: the root node (Tesnière, 1959).

Dependency structures have proved to be useful in several areas of natural language processing such as: information extraction, machine translation, and question answering, among others.

Basically, there are two ways to obtain automatically such structures: grammar-based and data-driven. In order to parse new sentences, the former relies on formal grammars, while the latter makes use of statistics from syntactically annotated corpora (or treebanks). Several broad coverage statistical dependency parsers are available and easily portable to new languages. Probably this is the reason why they are becoming more and more popular.

The work we present here is addressed to build a dependency parser for Catalan to be used for *language for specific purposes* (LSP) texts. In order to achieve this target, first we had to compile a Catalan treebank.

Our starting point to create the treebank was the texts included in a multilingual corpus with Spanish, Catalan, and English texts. 42,000 sentences of the Spanish section have already been syntactically annotated (Marimon et al., 2012) and using its Catalan section we could also aim at building a multilingual treebank. But for already available multilingual treebanks, a common issue is that different languages do not share a common annotation schema. Therefore, the same linguistic phenomenon is analyzed differently in each language. Our secondary objective was to avoid this drawback.

We also decided that, instead of carrying out an expensive exercise of human annotation, like the one carried out for the Spanish part (Marimon et al., 2012), or the one carried out by other initiatives, such as AnCoraCAT (Taulé et al., 2008), we would experiment with the use of a Spanish Dependency parser (Padró et al., 2013) to bootstrap it.

The assumption was that, using a delexicalized version of the Spanish MaltParser (Nivre and Hall, 2005),[1] we could speed up the annotation of the number of Catalan sentences required to train a Catalan parser, given the typological similarity between the two languages. Human annotators, then, would only have to correct proposed parses and to edit them to include Catalan particular phenomena. Two further considerations supported this decision. On the one hand, the Spanish parser had a very good performance –93.16% Labelled Attachment Score (LAS)– therefore, the proposed parses were expected to be mostly good. Besides, with the MaltParser we could build a Spanish model controlling the features used to produce the parse trees, so that we could avoid using lexical items. Such a model will be usable for Catalan sentences whose structures are similar to Spanish, but with different lexical items. On the other hand, we would reduce the lack of consistency (intra-annotor and inter-annotator) that commonly happens in human annotator's first stages of the learning curve (Zeman and Resnik, 2008).

The similarity between Catalan and Spanish is well known, but, nevertheless, there are syntactic differences that were not to be covered by the Spanish parser.

- Clitics. While Spanish only has accusative and dative case clitics (used as pronominal references of di-

[1]We name 'non lexicalized parser' to refer to the usage of a plain MaltParser while a 'lexicalized parser' means a parser trained using the Malt Optimizer. The latter includes the usage of word forms and POS information to optimize the parser result.

rect object and indirect object), Catalan language also has the "hi" and "en" clitics (used mainly as pronominal references of locative and modal complements and nominal complements, respectively) in addition to accusative and dative clitics.

- The auxiliary verb "anar". Spanish auxiliary verbs are "haber", for active forms, and "ser", for passive forms. Catalan main auxiliaries are the translational equivalents, but it also uses 'anar' (literally 'to go') as an auxiliary verb for the indefinite past tense: for example, the Spanish sentence "La chica compró galletas" (literally, 'the girl bought cookies') is translated in Catalan as "La noia va comprar galetes" ('the girl went to buy cookies'). This is a rather frequently used tense. The problem with this auxiliary is that it combines with the main verb in infinitive form. Such a combinantion (aux+infinitive) does not exist in Spanish, thus the Spanish parser could not annotate these structures correctly.
- Possessive determiner constructions. While in Spanish there are possessive determiners: "mi casa" ('my house'), in Catalan, there are no possessive determiners, but adjectives that require an article to build a noun phrase: "la meva casa" (literally, 'the mine house').

In addition to the linguistic differences just reported, there were other practical issues that had to be taken into account. For example, the MaltOptimizer (Ballesteros and Nivre, 2012) takes as input PoS tagged sentences, so we had to harmonize the Catalan and Spanish tagsets delivered by the FreeLing tagger (Padró and Stanilovsky, 2012). For instance, for the case of the possessives we have mentioned above, the Catalan Freeling tagger tags them as pronouns, while in Spanish equivalent words are tagged as adjectives.

2. State-of-the-art on cross lingual resources creation

The need to speed up the building of treebanks and associated parsers for different languages and particular domains has motivated different works related to what we present here. The most relevant ones are the following.

Hwa et al. (2005) projected to a different language (and automatically corrected) the parses obtained by using an English parser on the English sentences of a parallel corpus. The reported result was a dependency accuracy of 72.1% for Spanish and of 53.9% for Chinese. The results achieved for Chinese were comparable to a parser trained with approximately 2,000 sentences of the Penn Chinese Treebank (the average length of the sentences was 20.6 words).

Zeman and Resnik (2008) used a reranking parser (Charniak and Johnson, 2005) trained on Danish to parse Swedish sentences. The selection of the languages was intended to benefit of them being closely related. Because the reranking parser worked with phrase structures, the treebanks for Danish (and Swedish for the gold-standard) were converted into phrase structures. The exercise also included the representation, the normalization, and the mapping of test-sets. Their results were considered comparable with those ones reported in (Hwa et al., 2005), where it is showed that the accuracy achieved was equivalent to the one that would had been achieved by annotating about 1,500 sentences in this case. Zeman and Resnik emphasized that it may look as a little quantity, but in terms of actual effort required it was significant, since it was in the first steps of the annotation exercise, where more inconsistencies and difficulties were found.

Recently, McDonald et al. (2013) have conducted a thorough testing of the possibility of using dependency parsers trained for one language to parse typologically related languages. They converted existing treebanks for different languages into the Stanford typed dependencies for English (de Marneffe and Manning, 2008), adding the required tags for covering phenomena in the different languages addressed in the experiments. These conversion implied human annotation and also provoked an harmonization exercise for gaining consistency across different languages, i.e. that the same label was not assigned to different linguistic relations in different languages, as well as ensuring that the same relation was annotated with the same label. With a LAS score of 70.29% for parsing Spanish with the parser trained with 4,105 Spanish sentences (112,718 tokens), they achieve 63,65% LAS when parsing a French corpus made of 3,978 sentences (90,000 tokens).

Another research work concerned with the harmonization of the annotation guidelines is (Soucek et al., 2013). In this work, the objective is to review manually and iteratively a set of 15,000 sentences (in German, French, Spanish, and Brazilian Portuguese) to ensure a uniform linguistic representation across the languages of the project.

3. Methodology

For building the treebank (and therefore the dependency parser) mentioned in Section 1. we applied the methodology showed in Figure 1. As already mentioned, the idea was to bootstrap the Catalan treebank by using a MaltParser with a language model trained with the IULA Spanish LSP Treebank to start parsing Catalan sentences.

More specifically, first, we used a delexicalized MaltParser model. Then, the Catalan parsed sentences were manually corrected. Once a number of Catalan correct parses was available, we created a Catalan language model and started using it for parsing new sentences. The hypothesis was that, at some point, it would be more convenient to use the Catalan trained model with fewer sentences instead of the larger Spanish based language model. In order to decide when to move from the Spanish parser to the Catalan one, we experimented with models trained differently: only Spanish, Spanish and Catalan, and only Catalan sentences.

The basic procedure was to add syntactic information to a Catalan text corpus. For such purpose we used the IULA's Technical Corpus, a collection of written specialized texts (Law, Economy, Genomics, Medicine, and Environment). [2] This corpus includes more than 1,300 documents in Spanish, Catalan, and English that contains about 32 millions words. The Spanish section was used to build the IULA Spanish LSP Treebank (Marimon et al., 2012). The Catalan portion of this corpus contains 976 documents with about

[2] See (Cabré et al., 2006) and (Vivaldi, 2009) for details.

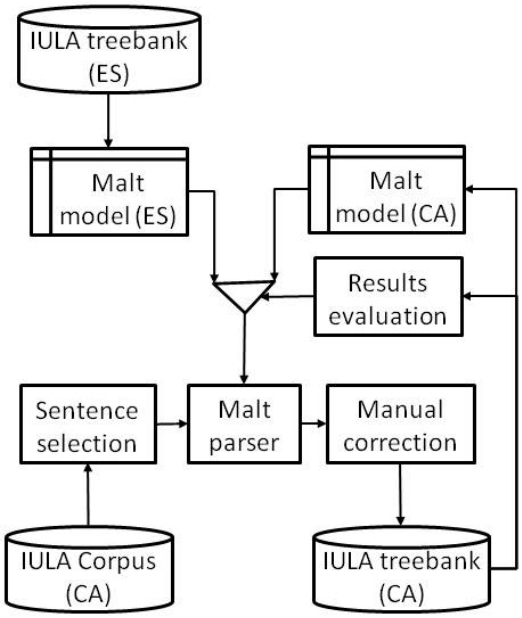

Figure 1: Methodology schematization

2.2 millions words distributed among 350,791 sentences. Figure 2 (Full corpus/50) shows the ratio of number of sentences per sentence length for the full Catalan corpus.

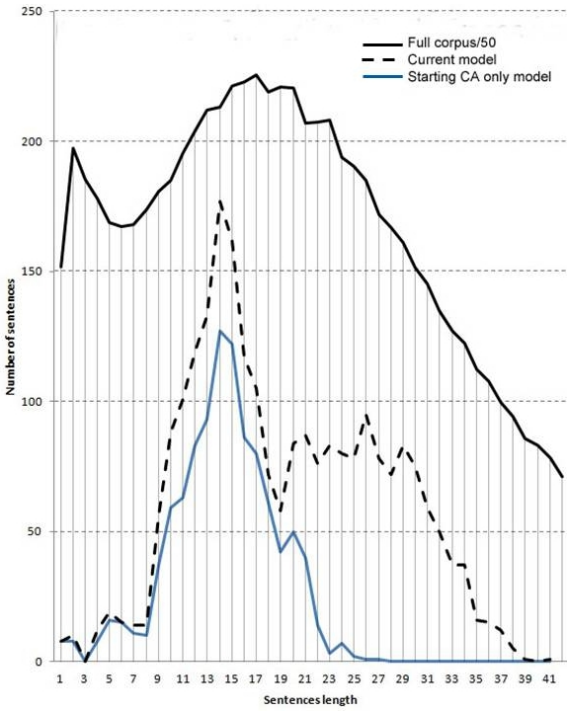

Figure 2: The IULA's Technical Corpus. Ratio of number of sentences per sentence length

The necessary morphological analysis over the chosen texts of such corpus was performed using Freeling, which morphologically annotates each word with morphological information by dictionary look-up and performs statistical PoS disambiguation (see description in (Padró and Stanilovsky, 2012)).

An alternative to the use of the IULA's corpus could have been AnCoraCAT (Taulé et al., 2008), an already existent Catalan syntactic treebank. However, two main reasons motivated our choice: (i) we wanted to use the resulting parser in connection with LSP texts, therefore, we needed a model trained with similar resources while AnCoraCAT is based on general language (mainly newspapers). (ii) the linguistic criteria behind AnCoraCAT are not identical to the criteria applied in our Spanish treebank; for example, in the analysis of coordinated structures, we follow Mel'cuk (1988)'s approach (i.e. the first conjunct is the head of the other elements, which are organized in a chain), whereas in AnCoraCAT, the first conjunct is the head and all other elements, including the conjunction, are attached directly to it.

For completing the procedure showed in Figure 1, we also implemented some additional resources:

- Sentence selector: Sentences to be included in the treebank were chosen at random, but replicating the IULA's Technical Corpus sentence distribution in terms sentence length and domain[3]. We designed an interface where the annotator is allowed to choose a number of sentences following such criteria (See Figure 3). Using the same interface the annotator may define the characteristics of a set of sentences to be analyzed: (i) choose the number of phrases and its length. (ii) include only sentences of a given number of (non auxiliary) verbs. (iii) include only sentences of particular domains. (iv) delete sentences from the current selection (due to ill-formed sentences).

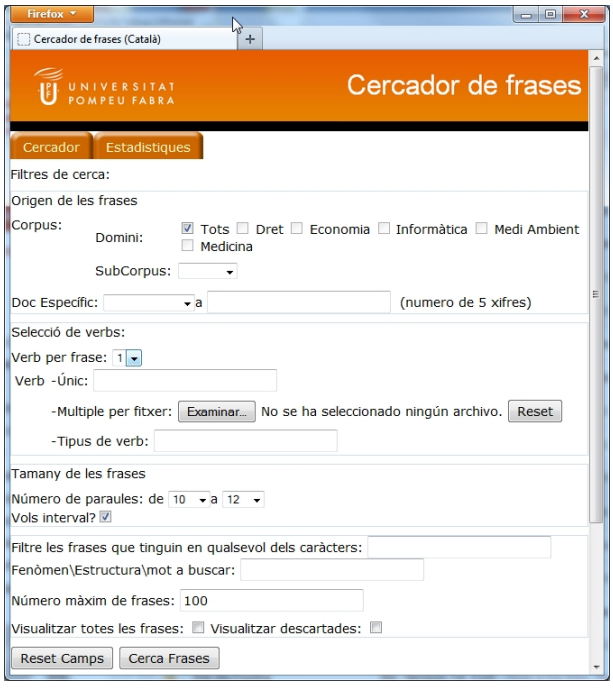

Figure 3: The IULA's Technical Corpus. Interface for choosing a sentences package

- Treebank development environment: In order to easing manual correction of parsed sentences we setup a

[3]Such strategy has been already used succesfully in the building of the Spanish treebank

working environment using using the yEd XML graph editor [4] (see Figure 4) and some supporting scripts for importing/exporting MaltParser results to/from XML. Using this environment the annotator could easily check the result proposed by the MaltParser and correct it editing nodes properties and links, if necessary. At the same time, some checking is done on corrected parses in order to keep consistency of the full treebank.

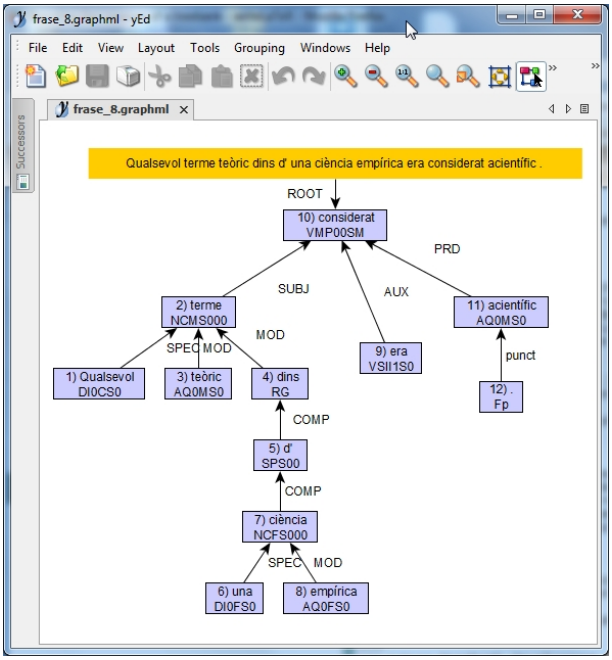

Figure 4: The IULA's Technical Corpus. Annotation environment

- Evaluation framework for assessing the results from two points of view:
  - Interannotator agreement: essential to guarantee consistency in the final treebank, that is, that a given linguistic phenomenon is always annotated in the same way. We followed the standard procedure: we periodically chose a set of sentences that was to be evaluated (accepted or corrected, if necessary) by all annotators, then, we calculated the agreement among them and discussed the differences.
  - Model evaluation: consecutive evaluation of the results was done in order to detect the appropriate time for using only the Catalan treebank for training the model to analyze more sentences.

4. Annotation task

As explained before, the Catalan Treebank annotation task was, in fact, a matter of deciding whether the parse provided by the parser was correct or not. In case it was not, the annotators had to manually correct it.[5] The most common phenomena that annotators had to correct were: (i) Parser mistakes in the distinction of nominal complements (COMP) and modifiers (MOD) and (ii) the correct parse for the clitic 'se' that can take different values (impersonal, reflexive, passive) and which is a difficult decision even for humans.

The annotation guidelines for Catalan mostly followed the decisions already taken for the IULA Spanish Treebank (see Marimon et al. 2014) in order to maximize an homogeneous treatment of syntactic phenomena both in Spanish and Catalan Treebanks, but some Catalan specific phenomena had to be introduced. For instance, for some verbs, infinitive objects take the form of a prepositional phrase headed by the preposition 'de', as in "les preguntes li permeten de formular algunes característiques" ('the questions allowed him to formulate some characteristics'). In this case, the preposition is annotated as the head of the relation. Another example is the annotation of the 'hi' and 'en' clitics, nonexistent in Spanish, where our guidelines followed the annotation made by the French Dependency treebank (Candito et al.2010).[6]

5. Experiments and Results

Based on the schema shown in Figure 1, we started our work with two annotators amounting to a total of 4 person/month. As usual, there was a starting stage for them to gain experience with the environment and to set up the annotation guidelines.

In order to evaluate the consistency of the annotation process, we carried out a series of inter-annotator agreement tests. Our aim was: (i) to detect which was the level of understanding of the annotation guidelines, (ii) to detect which were the most frequently disagreed linguistic phenomena, (iii) to estimate a baseline of the maximum results we can expect from our machine-learning process. The agreement has been calculated at three levels:

- Full sentence;
- Full CoNLL line;
- Elementary decisions: target node and dependency name independently.

We perform such evaluations using two packages of sentences of different lengths (one with sentences whose length is between 7 and 8 tokens and another with length 12-13). Table 1 shows the Kappa value for each of the above mentioned evaluations, together with some associated figures.

The main reasons for the disagreement were: (i) distinction between complements and adjunct modifiers; (ii) structural position and type of adverbs; (iii) the inherent linguistic ambiguity of PP-attachment; (iv) complex syntactical issues; (v) misunderstanding of the annotation guidelines; (vi) annotators' fatigue.

At the same time, we measured the LAS score as the Catalan treebank was becoming larger. For this purpose, we evaluated two sentence packages (lengths 12-13 and 18-19) using four models; each one trained using different combinations of Spanish and Catalan sentences (see details in

[4] http://www.yworks.com/en/products_yed_about.html

[5] An annotator manual with a set of guidelines was compiled in order to ensure a unified annotation schema.

[6] http://alpage.inria.fr/statgram/frdep/Publications/FTB-GuideDepSurface.pdf (consulted in January 2014).

Table 1: Inter-annotator agreement results

Sentence length [tokens]		7-8	12-13
Total number of phrases in the test package		75	68
Evaluating full phrases	Number of sentences	74	63
	Identical sentences	54	37
	Kappa	0.73	0.58
Evaluating elementary decisions	Number of decisions	1498	1862
	Differences in the dependency name	24	22
	Kappa	0.98	0.98
Evaluating full CoNLL line	Number of lines	749	931
	Differences	26	30
	Kappa	0.98	0.98

Table 2): (i) original Spanish Malt model; (ii) original Spanish Malt model plus some Catalan sentences; (iii) using a model trained exclusively with the Catalan corrected sentences (de-lexicalized); (iv) using a model trained exclusively with the Catalan corrected sentences, but including lexical information (using the MaltOptimizer).

Table 2 shows the LAS figures obtained using the above mentioned models in the sentence packages. Obviously, the sentence packages used for this evaluation were not included in the training model. The results showed that, when the MaltParser model was trained with the delexicalized mode with about 1,000 Catalan sentences, it performed better that including Spanish sentences in the training model, and when the Catalan model was trained in lexicalized mode, the parser performed even better. Therefore, we decided to proceed in the treebank enlargement using this model to tag new sentences. Figure 2 ('Starting CA only model' curve) shows the sentence distribution corresponding to such model.

We performed some experiments for checking the behaviour of the LAS score as the model became larger. A corpus of 2,476 sentences was divided in five subcorpora of increasing sizes and approximately reflecting evolution along the project life. We also defined a package of 139 sentences with one main verb and length between 13 and 23 tokens (it covers the most frequent sentence lengths in our corpus, see Figure 2 -current model-). Figure 5 illustrates the evolution of LAS using MaltOptimizer as the size of the training corpus became larger. At the time of writing this paper, the training corpus contains 2,400 sentences.

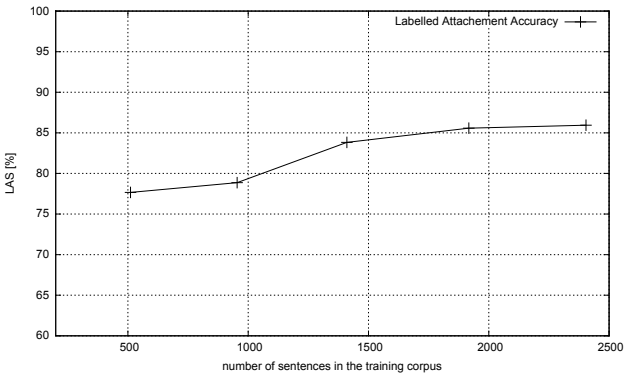

Figure 5: Evolution of LAS against training corpus size.

6. Discussion

Our task proceeds by enlarging the Catalan treebank taking care that sentences distribution follows the same profile as that of the IULA's corpus. Figure 3 shows both the original distribution of the corpus (Full corpus/50) and the linguistic model described here (Current model). Sentence packages are selected in order to make both curves (ideally) proportional.

Our approach followed the same general policy of the experiments described in (McDonald et al., 2013). More concretely, it uses the same annotation schema for both Catalan and Spanish languages. Considering that both French and Catalan are relatively comparable Latin languages, Table 3 shows that the results may be considered as comparable. The results for Catalan are slightly better, but it is necessary to take into consideration that our starting point was a ten times larger model. Later, after reaching a good behaviour using a pure Catalan model, we further improved results (reaching a LAS of 86%) using a lexicalized model. Such improvement together with the use of a lexicalized model fully justifies that with only one thousand sentences were enough for continuing with only the Catalan model.

Table 3: Results comparison

		Source	Target
(McDonald et al., 2013)	Language	Spanish	French
	Sentences	4,105	3,578
	LAS	70.29	63.65
ours	Language	Spanish	Catalan
	Sentences	42,000	1,000
	LAS	93.16	79.00

For achieving the results showed in Table 2 there were necessary two annotators during 4 months. In order to evaluate the improvement in terms of resources necessary to build the treebank, we may compare this figure with the resource

Table 2: LAS score using different MaltParser models

Sentence length (tokens)			12-13	18-19
MaltParser model trained with ...	Training model size			
	ES	CA		
.. Spanish sentences only	42000	0	78.97	79.04
.. Spanish sentences enriched with Catalan sentences	42000	1000	82.94	82.08
.. only with Catalan sentences (without lexical information)	0	1000	83.64	85.39
.. only with Catalan sentences (with lexical information)	0	1000	85.98	86.49

necessary to build the Spanish treebank. At that time, the resources necessary were twice in relation to this project. Although, it has to be taken into account that tasks are similar but no directly comparable. For Spanish, we used the DELPH-IN environment, [7] were the task was to choose the correct parse among a number of sorted parses. For Catalan, the parser proposes a single parse that the annotator must check and modify if necessary. In the latter case, it may be necessary to modify several attributes of one or more nodes. In any case, these results confirm the general hypothesis that using a model from a comparable language it is possible to boost the creation of a treebank for a new language.

7. Conclusions and future work

This paper describes an ongoing work for the creation of the IULA Catalan LSP Treebank, a dependency treebank. We have described the methodology that we have used to create the resource. Such methodology is innovative as it takes profit of an already existent resource for another linguistically close language. We also describe how the new resource has been continuously evaluated regarding both the agreement among the manual annotators as well as the performance of the language model that is being created.

8. Acknowledgments

This work was partially supported by the SKATER project (Ministerio de Economía y Competitividad, TIN2012-38584-C06-05).

9. References

Ballesteros, Miguel and Nivre, Joakim. (2012). Maltoptimizer: An optimization tool for maltparser. In *Proceedings of the 13th Conference of the European Chapter of the Association for computational Linguistics (EACL 2012). Demo Session.*

Cabré, María Teresa, Bach, Carme, and Vivaldi, Jorge. (2006). 10 anys del corpus de l'IULA.

Charniak, Eugene and Johnson, Mark. (2005). Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 173–180. Association for Computational Linguistics.

[7] http:///www.delph-in.net

de Marneffe, Marie-Catherine and Manning, Christopher D. (2008). The stanford typed dependencies representation. In *COLING Workshop on Cross-framework and Cross-domain Parser Evaluation*.

Hwa, Rebecca, Resnik, Philip, Weinberg, Amy, Cabezas, Clara, and Kolak, Okan. (2005). Bootstrapping parsers via syntactic projection across parallel texts. (11):311–325.

Marimon, Montse, Fisas, Beatriz, Bel, Núria, Arias, Blanca, Vázquez, Silvia, Vivaldi, Jorge, Torner, Sergi, Villegas, Marta, and Lorente, Mercè. (2012). The IULA treebank. In Calzolari, Nicoletta, Choukri, Khalid, Declerck, Thierry, Doğan, Mehmet Uğur, Maegaard, Bente, Mariani, Joseph, Odijk, Jan, and Piperidis, Stelios, editors, *Proceedings of the 8th international conference on Language Resources and Evaluation*. European Language Resources Association (ELRA).

McDonald, Ryan, Nivre, Joakim, Quirmbach-Brundage, Yvonne, Goldberg, Yoav, Das, Dipanjan, Ganchev, Kuzman, Hall, Keith, Petrov, Slav, Zhang, Hao, Tackstrom, Oscar, Bedini, Claudia, Castell, Núuria Bertomeu, and Lee, Jungmee. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97. Association for Computational Linguistics.

Mel'cuk, Igor. (1988). *Dependency Syntax: Theory and Practice*. State University of New York Press, New York.

Nivre, Joakim and Hall, Johan. (2005). Maltparser: A language-independent system for data-driven dependency parsing. In *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories (TLT)*, pages 137—-148.

Padró, Lluís and Stanilovsky, Evgeny. (2012). Freeling 3.0: Towards wider multilinguality. In Calzolari, Nicoletta, Choukri, Khalid, Declerck, Thierry, Doğan, Mehmet Uğur, Maegaard, Bente, Mariani, Joseph, Odijk, Jan, and Piperidis, Stelios, editors, *Proceedings of the 8th international conference on Language Resources and Evaluation*. European Language Resources Association (ELRA).

Padró, Muntsa, Ballesteros, Miguel, Martínez, Hector, and Bohnet, Bernd. (2013). Finding dependency parsing limits over a large spanish corpus. In *Proceeding of the 6th International Joint Conference on Natural Language Processing (IJCNLP)*.

Soucek, Milan, Jarvinen, Timo, and LaMontagne, Adam.

(2013). Managing a multilingual treebank project. In *Proceedings of the 2nd International Conference on Dependency Linguistics*, pages 292–297.

Taulé, Mariona, Martí, María Antònia, and Recasens, Marta. (2008). Ancora: Multilevel annotated corpora for catalan and spanish. In *Proceedings of the 6th international conference on Language Resources and Evaluation*. European Language Resources Association (ELRA).

Tesnière, L. (1959). *Elèments de syntaxe structurale*. Editions Klincksieck.

Vivaldi, Jorge. (2009). Corpus and exploitation tool: IULACT and bwanaNet. In de Lingüística del Corpus, Asociación Española, editor, *A survey on corpus-based research (CICL-09)*.

Zeman, Daniel and Resnik, Philip. (2008). Cross-language parser adaptation between related languages. In *Proceedings of IJCNLP 2008 Workshop on NLP for Less Privileged Languages*.