Natural Language Engineering **1** (1): 1–33. Printed in the United Kingdom

New Treebank or Repurposed? On the Feasibility of Cross-Lingual Parsing of Romance Languages with Universal Dependencies†

MARCOS GARCIA,

Universidade da Coruña, LyS Group
Departamento de Letras, Facultade de Filoloxía
Campus de A Coruña, 15071, A Coruña, Galicia, Spain
e-mail: marcos.garcia.gonzalez@udc.gal

CARLOS GÓMEZ-RODRÍGUEZ and MIGUEL A. ALONSO

Universidade da Coruña, LyS Group
Departamento de Computación, Facultade de Informática
Campus de A Coruña, 15071, A Coruña, Galicia, Spain
e-mails: {carlos.gomez, miguel.alonso}@udc.es

(*Received*)

Abstract

This paper addresses the feasibility of cross-lingual parsing with Universal Dependencies (UD) between Romance languages, analyzing its performance when compared to the use of manually annotated resources of the target languages. Several experiments take into account factors such as the lexical distance between the source and target varieties, the impact of delexicalization, the combination of different source treebanks or the adaptation of resources to the target language, among others. The results of these evaluations show that the direct application of a parser from one Romance language to another reaches similar LAS values to those obtained with a manual annotation of about 3,000 tokens in the target language, and UAS results equivalent to the use of around 7,000 tokens,

† This work has been partially supported by the Spanish Ministry of Economy and Competitiveness (MICINN) through a *Juan de la Cierva formación* grant (FJCI-2014-22853), by the projects with reference FFI2014-51978-C2-1-R and FFI2014-51978-C2-2-R (MINECO), and by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 714150 - FASTPARSE).

depending on the case. These numbers can noticeably increase by performing a focused selection of the source treebanks. Furthermore, the removal of the words in the training corpus (delexicalization) is not useful in most cases of cross-lingual parsing of Romance languages. The lessons learned with the performed experiments were used to build a new UD treebank for Galician, with 1,000 sentences manually corrected after an automatic cross-lingual annotation. Several evaluations in this new resource show that a cross-lingual parser built with the best combination and adaptation of the source treebanks performs better (77% LAS and 82% UAS) than using more than 16,000 (for LAS results) and more than 20,000 (UAS) manually labeled tokens of Galician.

1 Introduction

Corpora with syntactic annotation (treebanks) are useful resources for training and evaluating statistical parsers, which in turn, can be used in different Natural Language Processing (NLP) applications, such as machine translation, information extraction or opinion mining (Gimpel and Smith 2014; Nguyen, Moschitti and Riccardi 2009; Socher, Perelygin, Wu, Chuang, Manning, Ng and Potts 2013). Furthermore, studies in corpus linguistics also benefit from the availability of treebanks, which allow researchers to extract information from real linguistic data (McEnery and Hardie 2011). However, manually labeling a new corpus is an expensive and time-consuming task, which requires a large effort by expert annotators to obtain high-quality data.

Aimed at alleviating the effort of creating a new treebank, this paper investigates the impact of different factors on cross-lingual parsing (i.e., analyzing a target language with resources from one or more source languages). Thus, we perform several experiments on cross-lingual parsing of Romance languages, and verify the practical usefulness of the lessons learned by carrying out a case study: the creation of a new treebank for Galician.

In the last few years, several strategies for projecting the syntactic annotation from a source language to a target one have been implemented, in order to automatically obtain corpora for the latter language (Hwa, Resnik, Weinberg, Cabezas and Kolak 2005; Ganchev, Gillenwater and Taskar 2009). The resulting data can then be corrected by an expert, thus reducing the effort with respect to labeling a new resource from scratch. Nevertheless, the different annotation guidelines used in each language resource make it complicated to leverage the cross-lingual resources.

A different approach consists in creating a parser for the target language without the need of a treebank. Some authors apply unsupervised methods (Klein and Manning 2004), while others rely on the direct transfer of model parameters from one language to another (Zeman and Resnik 2008; McDonald, Petrov and Hall 2011). These methods usually reduce language-specific information by ignoring the lexical features in the learning process, thus building *delexicalized* parsers.

More recently, several approaches have emerged with the aim of harmonizing the annotation of syntactic dependencies among different languages and treebanks (McDonald, Nivre, Quirmbach-Brundage, Goldberg, Das, Ganchev, Hall, Petrov, Zhang, Täckström, Bedini, Bertomeu Castelló and Lee 2013; Zeman, Dušek,

Mareček, Popel, Ramasamy, Štěpánek, Žabokrtskỳ and Hajič 2014; de Marneffe, Dozat, Silveira, Haverinen, Ginter, Nivre and Manning 2014), ending up with the creation of the Universal Dependencies (UD) initiative (Nivre, de Marneffe, Ginter, Goldberg, Hajič, Manning, McDonald, Petrov, Pyysalo, Silveira, Tsarfaty and Zeman 2016). The promoters of this project developed a set of common (*universal*) guidelines for annotating a treebank, thus facilitating the leverage of syntactic resources as well as the linguistic analysis between languages.

In this respect, some experiments using treebanks of different languages with harmonized labeling showed that cross-lingual parsing can achieve better performance than previous unsupervised approaches (McDonald *et al.* 2011). However, and even though the research on cross-lingual parsing has increased (Agić, Tiedemann, Merkler, Krek, Dobrovoljc, and Moze 2014; Rosa and Žabokrtskỳ 2015b; Tiedemann 2015), it is still difficult to answer questions such as the following:

- Is lexical distance between the source and target languages more influential for cross-lingual parsing than their structural differences?
- Is it worth it to delexicalize the models for cross-lingual parsing of Romance languages?
- To what extent can we trust cross-lingual parsing between languages from the same linguistic family?

In order to answer these questions, this paper presents a set of experiments concerning cross-lingual parsing between Romance languages. First, we analyze the lexical distance as well as the mutual lexical coverage between eight linguistic varieties. Then, we evaluate the performance of direct cross-lingual parsers, using both one source language and several treebank combinations.

We will show that, in most cases, delexicalization is not useful for cross-lingual parsing between Romance languages, and this phenomenon is in significant correlation with the lexical coverage of the treebanks of the source and target languages. Furthermore, the results of direct transfers using a single treebank as source achieve a performance equivalent to the use of a given amount of manually annotated tokens of the target language, and this amount can be noticeably increased if a previous selection and adaptation of the best source treebanks is performed.

As previously mentioned, the results of these experiments allowed us to reduce the effort of creating a new UD treebank for Galician. This new resource was manually corrected after an automatic annotation using cross-lingual parsing, confirming that it is possible to train a high performance UD parser for a new language with little manual effort.[1]

The remainder of this article is organized as follows. Section 2 gives an overview of the design process of the Universal Dependencies, and introduces related work on cross-lingual parsing. Then, Section 3 presents a set of parsing experiments using different treebanks of Romance languages. After that, the creation steps of the new treebank for Galician are shown in Section 4, together with several monolingual and

[1] The new Galician treebank (*Galician-TreeGal*) is freely available in the Universal Dependencies initiative since its version 1.4.

cross-lingual tests. Finally, we discuss the main results of this paper in Section 5, and present the conclusions and further work in Section 6.

2 Related Work

This section includes a brief introduction to the UD initiative, followed by a presentation of some of the most influential papers concerning cross-lingual parsing.

2.1 Universal Dependencies

The Universal Dependencies project —which stems from the Universal Treebanks promoted by Google (McDonald *et al.* 2013)— started in 2014 with the main goals of developing a cross-linguistically consistent grammatical annotation, as well as providing treebanks labeled using the same guidelines.[2] This harmonized annotation supports multilingual research in both comparative linguistics and NLP. In practice, UD unifies several attempts that had been developed, aimed at performing cross-lingual POS-tagging and dependency parsing.

The representation of UD includes, for each token, its lemma, POS-tag, morphological features and the syntactic dependency it belongs to in the sentence (all this information encoded with universal labels and tokenized using the same criteria). The UD POS-tags were started by Petrov, Das and McDonald (2012), which proposed a tagset of 12 elements, then enriched and modified to the current inventory of 17 tags.[3] As these labels only classify POS categories, a different layer encodes the morphological information. In this regard, Zeman (2008) had presented a tool for converting morphological features from different languages into a universal standard, *Interset*, later employed in various projects such as HamleDT (Zeman *et al.* 2014).

HamleDT introduced a compilation of 29 existing treebanks automatically converted to a harmonized annotation. The syntactic labeling of the first version was inspired by the Prague Dependency Treebank (Bejček, Panevová, Popelka, Straňák, Ševčíková, Štěpánek and Žabokrtskỳ 2012), being adapted to the Stanford dependencies in further versions (Rosa, Masek, Marecek, Popel, Zeman and Zabokrtskỳ 2014).

The Stanford dependencies, initially developed for English (de Marneffe, MacCartney and Manning 2006) and basis of UD, experimented an evolution towards a universal set of dependency relations (de Marneffe and Manning 2008; de Marneffe *et al.* 2014), which facilitated the combination with the mentioned universal POS-tags (e.g., in the already mentioned Google Universal Treebanks) and also with a universal set of morphological features (Tsarfaty (2013), which proposed a variation of the Standard dependencies, called U-SD) in order to develop the current version of UD.

[2] `http://universaldependencies.org/`
[3] `http://universaldependencies.org/u/pos/index.html`

Therefore, currently UD merges these different approaches for developing multilingual treebanks with consistent annotation (Nivre *et al.* 2016), using a new version of the CoNLL format called CoNLL-U. It is worth noting that although UD promotes a universal set of syntactic dependencies, it permits the use of other labels for representing language-specific phenomena. These labels, however, are subtypes of the core UD relations (named as *udrelation:subtype*), so an alignment between universal and language-specific dependencies is preserved in some way.

In sum, the unification of UD lies in the use of (i) universal tagsets for POS-tagging, morphological encoding and dependency syntax, and (ii) common guidelines for tokenizing and for labeling syntactic phenomena. In this regard, UD suggests a unified annotation of controversial structures such as coordination, verbal groups or the relation between a preposition and a noun phrase, among others.

2.2 Cross-lingual parsing

The different strategies which have been used for cross-lingual parsing can be classified in two main groups: (a) data transfer, and (b) model transfer approaches. The first one creates artificial data of the target language by projecting the linguistic information from a source treebank, sometimes with the help of machine translation. Differently, model transfer approaches use the source data for training models that can be used for analyzing one or more target languages. As pointed out in the introduction, the emergence of UD facilitates research in cross-lingual dependency parsing, but several papers had already addressed this topic from different perspectives.

2.2.1 Data transfer and annotation projection

Concerning the projection of syntactic labeling, one of the most common strategies is the use of parallel corpora, which was introduced by Yarowsky, Ngai and Wicentowski (2001) for other NLP tasks such as POS-taggers, chunkers, or lemmatizers.

Hwa *et al.* (2005) parse the English version of English-Spanish and English-Chinese parallel corpora, and then project the syntactic dependencies from the source language to Spanish and Chinese, respectively. After that, they train statistical parsers on the resulting data. Even though this strategy requires parallel corpora (which are not easy to obtain for many languages), the best results of the Spanish transfer were better than those obtained by a commercial parser ($\approx 72\%$ of unlabeled F-score). However, the Chinese results were noticeably lower ($\approx 44\%$) due to the complexity of the English-Chinese parallel corpora alignment.

The strategy presented by Hwa *et al.* (2005) can be improved in several ways: Smith and Eisner (2009) show that using quasi-synchronous features and some manually annotated sentences of the target language provides a boost equivalent to doubling the number of target trees. Another strategy for improving the use of parallel corpora consists in taking advantage also of target trees with partial analysis, since Hwa *et al.* (2005) only used the sentences with perfectly conserved dependencies (Ganchev *et al.* 2009). Besides, Ganchev *et al.* (2009) also add some rules for

reducing the most frequent differences between some treebanks (e.g., the selection of the main and auxiliary verb in verb groups or the status of the prepositions in noun phrases).

More recently, various approaches addressed again annotation transfer between parallel corpora, taking advantage of the emergence of resources with harmonized labeling. Thus, Tiedemann (2014) projects dependency labels using both manual translations (from Europarl) and machine translated corpora, showing that a consistent annotation between treebanks improves the performance of the transfer. Similar experiments, including data subset selection, are presented by Tiedemann (2015), which confirms that building parallel corpora with machine translation gives better results than projecting the labels in automatically annotated parallel corpora. The use of machine translation to obtain labeled data of a target language was also addressed in several works, showing the importance of the lexical features and the impact of POS-tagging in dependency parsing (Tiedemann, Agić and Nivre 2014; Tiedemann and Agić 2016).

The use of *dense projected structures* is presented by Rasooli and Collins (2015), obtaining high-quality projections that improve cross-lingual parsing performance.

Following a similar approach to the one presented by Agić, Hovy and Søgaard (2015) for POS-tagging, Agić, Johannsen, Plank, Martínez Alonso, Schluter and Søgaard (2016) perform cross-lingual parsing for languages with very low resources. Both for POS-tagging and parsing, the authors rely on multi-source strategies for projecting the labeling of widely translated texts, such as the Bible.

Finally, Lacroix, Aufrant, Wisniewski and Yvon (2016a) carry out annotation transfer using parallel corpora, ignoring unattached words and many-to-many alignments between the two resources. This paper shows that learning from high-quality (but partial) data is better than utilizing fully-annotated data with some noise. Using the same approach, Lacroix, Wisniewski and Yvon (2016b) analyze the impact of pre-processing (and post-processing) the parallel data, proving that filtering out noisy sentences improves cross-lingual parsing. Also, they address multi-source transfer of dependency annotation, achieving better results when combining treebanks from the same linguistic family, and they show that the transfer results are surpassed by supervised models trained on $\approx$ 300 sentences (depending on the languages).

2.2.2 Model transfer

As mentioned, the other main strategy for parsing a new language consists in using cross-lingual models built with resources from other linguistic varieties.

The adaptation of a parser aimed at analyzing a similar language is addressed by Zeman and Resnik (2008), who evaluate the use of Danish corpora to train a parser for analyzing Swedish. The best results are obtained when performing a delexicalization of the corpora (replacing the words with their POS-tags, previously mapped between the two languages), concluding that this strategy produces the same results as manually annotating $1,546$ sentences in the target language.

McDonald *et al.* (2011) were one of the first researchers using universal POS-

tags for dependency parsing, also introducing the multi-source approach for cross-lingual parsing. They train delexicalized models that obtain better results than unsupervised approaches, and show that multi-source parsers (built with simple concatenation of the training corpora of different languages) can be useful for cross-lingual parsing. A similar approach was presented by Cohen, Das and Smith (2011), who combine supervised models of various source languages for both POS-tagging and dependency parsing.

Søgaard (2011) trains delexicalized parsers selecting —in the source treebanks— only those sentences whose structure is similar to the target language, obtaining better performance than the previous method, also for non-related languages (Bulgarian, Portuguese, Arabic, and Danish).

Naseem, Barzilay and Globerson (2012) implement an algorithm for transferring dependency models that learns different properties from multilingual treebanks, even from non-related languages. The system first learns the *universal* distribution of each POS-tag's dependents, followed by an ordering component that determines the position (left or right) of each dependent. The results on several languages largely outperform direct delexicalized parsers as well as the concatenation of multiple source treebanks.

Täckström, McDonald and Uszkoreit (2012) and Durrett, Pauls and Klein (2012) also rely on delexicalized parsers. The former performs an enrichment of the syntactic transfer through cross-lingual word clusters used as features, while the latter adds lexical features by means of bilingual dictionaries, increasing the accuracy of the cross-lingual parsing between 1 and 2%. A similar approach performs relexicalization on multi-source parsers built by means of selective parameter sharing (Täckström, McDonald and Nivre 2013).

Several cross-lingual parsing experiments of related languages (Croatian, Serbian, and Slovene) were performed by Agić *et al.* (2014), suggesting that delexicalization is not necessary for cross-lingual parsing in these Slavic languages.

McDonald *el al.* (2013) presented the Google Universal Dependency Treebanks, the first widely-adopted set of harmonized treebanks, providing a more reliable evaluation of cross-lingual parsing. The experiments performed in that paper show that for each of the Germanic and Romance languages analyzed (German, English, Swedish, Spanish and French), the best source is from the same linguistic family.

Using the first version of the UD treebanks, Tiedemann (2015b) performs an exhaustive evaluation of cross-lingual parsing, measuring the impact of predicted POS-tags (when compared to gold ones), and also carrying out some experiments in annotation projection and treebank translation.

With the HamleDT treebanks, Rosa and Žabokrtskỳ present a metric for measuring the distance between languages (Rosa and Žabokrtskỳ 2015). This distance is then used to assign a weight to each of the source treebanks in a multi-source scenario, optimizing the training data to the target language (Rosa and Žabokrtskỳ 2015b).

In a similar way to Naseem *et al.* (2012) or Täckström *et al.* (2013), Zhang and Barzilay (2015) address multilingual transfer parsing, taking advantage of hierarchical tensor models. In order to incorporate (partial) lexical information, they use

multilingual word-embeddings of the most frequent words. Following the distributional semantic approach, other works learn bilingual word-embeddings from parallel corpora to avoid the problems of delexicalization in multi-source cross-lingual parsing (Guo, Che, Yarowsky, Wang and Liu 2015; Guo, Che, Yarowsky, Wang and Liu 2016).

Duong, Cohn, Bird and Cook (2015) present a neural network approach for cross-lingual parsing of low-resource languages. The method uses an *interlingual* representation with some specific mappings for each language, and it also infers syntactic information from multilingual word-embeddings. Even if it is not mainly focused on syntactic analysis, Søgaard, Agić, Martínez Alonso, Plank, Bohnet and Johannsen (2015) showed how bilingual word-embeddings learned from Wikipedia (without using parallel corpora) can be useful for cross-lingual parsing.

Aufrant, Wisniewski and Yvon (2016) use linguistic information from the World Atlas of Language Structures[4] to adapt the sentences of delexicalized source treebanks to the structure of a target language (e.g., word order, use of determiners, etc.). This strategy obtains better results than using a POS-tag model of the target language.

Recent experiments also combined several source treebanks for training multilingual parsers, capable of analyzing texts in more than one language (Vilares, Alonso and Gómez-Rodríguez 2016; Ammar, Mulcaire, Ballesteros, Dyer and Smith 2016). These approaches can be implemented without performing delexicalization of the training data, so the resulting parsers effectively use lexical information from one language to analyze a different one.

In the present paper, we focus on the analysis of different syntactic properties and lexical similarity of the source and the target languages for cross-lingual parsing of Romance languages. Some of the results of this analysis are then applied in a case study, the construction of a new UD treebank of Galician.

3 Cross-lingual transfer of parsing models for Romance languages

As pointed out in the previous section, the emergence of harmonized treebanks for several languages has allowed the research community to evaluate the transfer of syntactic resources between different linguistic varieties. However, several experiments have shown that cross-lingual parsing results are not always satisfactory.

In this respect, the experiments by McDonald *et al.* (2013) suggest that Romance languages might be reasonably well analyzed using resources from other varieties from the same linguistic family (e.g., a parser for Spanish obtained $> 75\%$ LAS analyzing French data). As the cross-lingual results on Germanic languages are lower than those of the Romance ones, the authors' hypothesis is that much of the divergence in Romance languages is lexical (and not structural).

Taking the above into account, this section includes a detailed exploration of cross-lingual UD parsing in Romance languages, analyzing the impact not only of

[4] `http://wals.info/`

lexical differences but also of other divergences such as the amount of training data or the number of dependency labels utilized by the annotators. To perform the experiments we used the 1.3 version of the UD treebanks for Romance languages: Catalan (CA), Castilian Spanish (ES), French (FR), Italian (IT), Romanian (RO), European Portuguese (EP) and Brazilian Portuguese (BP).[5]

It is worth noting that, as UD is an ongoing project, some of the available treebanks present divergences in annotation, since they derive from previous corpora labeled following diverse guidelines. In this regard, differences in tokenization (e.g., the current version of the BP corpus does not split contractions, as proposed by UD) or in the use of some dependencies (e.g., ES and EP treebanks do not use the *expl* dependency) could have an effect in cross-lingual parsing tests.[6]

Concerning the experiments, we first show the results of lexical similarity and coverage tests aimed at estimating the lexical distance between the analyzed languages. Then, we carry out a set of cross-lingual parsing evaluations between all the mentioned Romance languages, using both lexicalized and delexicalized models. Finally, we calculate the learning curve for each language, and verify the amount of training data in the target language needed for outperforming cross-lingual parsing.

3.1 Lexical similarity and coverage between Romance languages

The lexical distance between two Romance languages (the source and the target) may be important in their mutual cross-lingual parsing. To find out the impact of this distance, we calculated both the lexical similarity and the treebank coverage of every language pair. The first experiment gives us an approximation of the general lexical distance between two languages, while the coverage analysis puts the focus on the frequency of co-occurrence of the words in the source and target corpora.

We used two different strategies for computing the lexical similarity between two languages and to obtain the lexical coverage of their treebanks. For the first analysis, we exploited large dictionaries of each language (namely those provided by the latest version (4.0) of FreeLing (Padró and Stanilovsky 2012),[7] together with the DELAF_PB —for Brazilian Portuguese— (Muniz, Nunes and Laporte 2005) and the MULTEXT —for Romanian— (Erjavec 2012)), to obtain a general comparison between the language pairs. The results of this analysis are shown in Table 1, where each row contains the percentage of tokens of the target dictionary (in each column) that are covered by the source one. For instance, the European Portuguese dictionary covers 20.6% of the Spanish one (i.e., 20.6% of the Spanish words appear in the EP lexicon). It is worth noting that these values refer to the

[5] In this paper, we use both *language* and *linguistic variety* as synonyms, meaning a *consistent linguistic system*. In this regard, we do not state that BP and EP are different languages even if, for clarity, they are sometimes included in expressions referring to different languages.

[6] In this respect, we did not use the Galician treebank provided by UD 1.3 in our experiments because it was not manually reviewed in its current initial stage.

[7] `https://github.com/TALP-UPC/FreeLing/blob/master/COPYING`

Table 1. Lexical similarity between Romance languages computed using large dictionaries. Each row shows the coverage percentage of a source dictionary on the target ones (in the columns). The last column is the number of entries of each dictionary.

Lang.	GL	CA	ES	FR	IT	BP	EP	RO	Dict. size
GL	*100*	5.8	27.8	1.7	2.6	12.0	13.8	2.0	428,117
CA	7.1	*100*	6.9	3.1	2.3	3.6	4.1	2.1	521,978
ES	36.2	7.3	*100*	2.4	3.3	11.0	12.6	2.2	556,425
FR	1.4	2.1	1.5	*100*	1.5	1.0	1.2	1.8	350,279
IT	2.2	1.6	2.1	1.6	*100*	1.5	1.6	2.8	360,827
BP	28.1	6.9	20.0	3.0	4.2	*100*	79.0	3.0	1,001,546
EP	29.3	7.1	20.6	3.0	4.1	71.6	*100*	3.0	908,820
RO	2.0	1.7	1.7	2.2	3.3	1.3	1.4	*100*	428,194

matching of orthographic tokens (and not of *lexical entries* from a linguistic point of view), since tokens are used as features by probabilistic parsers.[8]

For computing the lexical coverage between the treebanks, we took advantage of the *train* splits of the 1.3 UD treebanks, also used for training the cross-lingual parsers.[9] The numbers in Table 2 are the percentages of word occurrences of one treebank (in each column) that are covered by the source treebank (in each row).[10] For instance, this analysis concludes that 55.4% of the Spanish word occurrences (in the treebank) can be covered by a model trained with the Catalan data. Therefore, these results are more related to cross-lingual parsing than the previous experiment, which analyzes lexical distance from a more generic point of view.

The results of Table 1 show that Spanish, both varieties of Portuguese and Galician are the languages with a closest relation in terms of lexical units, followed by Catalan. This is not strange due to their geographical closeness (all of them are Iberian Romance languages), even if they use different spelling traditions. In contrast, the dictionaries of French (and also of Italian and Romanian) have a very low coverage ($< 2\%$) of the other Romance languages.

However, these large differences are reduced if we take into account the word frequency used by the treebank coverage analysis (Table 2), which also minimizes

[8] Thus, orthographic items such as "coincidência" (in EP or BP) and "coincidencia" (in ES) are considered as different words.

[9] The Galician (GL) data was extracted from the treebank that we annotated for this article (see Section 4).

[10] In this case, we only considered two words (one in each language) as the same if both of them also have the same POS-tag. Thus, the noun "cobra" (*snake*) in Portuguese shall not be considered the same word as the verb form "cobra" (from the verb "cobrar", *to collect*) in Spanish.

Table 2. Lexical coverage (case sensitive) between Romance languages computed using the *train* splits of the UD treebanks (version 1.3). Each row shows the coverage percentage of a source treebank on the target ones (in the columns). The last columns show the average coverage (Avg), and the number of unique *token–TAG* elements (Words) of each treebank, respectively.

Lang.	GL	CA	ES	FR	IT	BP	EP	RO	Avg	Words
GL	*100*	32.2	45.9	24.3	26.0	41.1	50.8	22.4	30.3	5,590
CA	38.1	*100*	55.4	36.8	31.2	35.0	34.7	25.5	32.1	33,734
ES	62.1	56.2	*100*	43.1	43.0	49.5	52.8	25.0	41.5	50,423
FR	38.6	43.7	44.4	*100*	39.0	34.6	36.7	24.6	32.7	45,122
IT	34.2	39.4	39.9	39.7	*100*	32.4	32.1	24.3	30.5	29,149
BP	62.6	38.3	50.1	35.8	37.2	*100*	86.4	23.2	41.7	32,107
EP	63.4	39.3	48.7	31.1	32.1	83.3	*100*	23.8	40.2	29,496
RO	32.7	26.8	26.9	23.8	21.9	27.5	34.9	*100*	24.3	22,731

the impact of variations in the size of the resources.[11] In absolute terms, several Romance languages have few coincident words (token and POS-tag) between them, but some frequent function words (and also nouns and adjectives) co-occur in different languages, fact which in principle, will favour the model transfer between them. In this respect, Spanish seems to be the language with higher mutual coverage among the Romance languages (bearing in mind that there are two varieties of Portuguese in the analysis).[12] These numbers also show other common linguistic perceptions, such as the relatedness between CA and FR, GL and EP, or IT and ES. Finally, the coverage values of Romanian are lower (and more homogeneous) than those of the other varieties: the highest coverage value as target is of 25.5% (CA), and as source language it reaches 34.9% (EP). While the dictionary-based comparison classifies French as the least related language (followed by Italian and Romanian, at the same level), Romanian is the language with least lexical coverage (it covers on average $\approx$ 24% of the other languages, *versus* 31% and 33% of Italian and French, respectively) when comparing the treebank lexica.

In order to estimate the impact of the similarity and mutual coverage values, we will come back to these numbers when evaluating the cross-lingual transfer results.

3.2 Experiments on cross-lingual parsing

In the following, we show the results of several experiments concerning cross-lingual UD parsing of Romance languages. We used as training and testing data the splits

[11] In this regard, it is interesting to note that the Galician treebank, with just 5,590 unique words, has similar coverage of the other languages to some other large datasets.

[12] The inclusion of CA and GL, which are co-official with Spanish, also increases the average results of ES.

provided by the UD 1.3 treebanks. For building the models, we utilized MaltParser 1.9 (Nivre, Hall, Nilsson, Chanev, Eryigit, Kübler, Marinov and Marsi 2007) executed out-of-the-box with Nivre's arc-eager algorithm (Nivre 2004).[13] The CoNLL-X *eval.pl* script (version 1.9) was used for computing both the Labeled (LAS) and the Unlabeled (UAS) attachment scores (ignoring punctuation tokens for computing the results).

In order to obtain realistic results (Tiedemann 2015b), we used predicted POS-tags in the test sets, obtained by UDPipe POS-tagger models previously trained on the *train* splits of the UD treebanks. Also, the language-specific relations were converted to their *universal* label in both training and test sets (e.g., *acl:relcl* → *acl*).

3.2.1 Direct model transfer

The first group of parsing experiments takes advantage of the UD harmonized treebanks for performing direct cross-lingual syntactic analysis. Thus, we trained both lexicalized and delexicalized models (replacing the tokens with "X") for each Romance language, and applied them directly to each of the other linguistic varieties. We also built several *All* models, trained with the combinations of all the source treebanks except the target one.

Tables 3 and 4 contain the MaltParser results of these evaluations. The numbers in Table 3 are LAS values, while the UAS results are shown in Table 4. In both tables, the testing treebank is represented in each column, while rows correspond to the training data. Furthermore, each language row includes the lexicalized (top) and delexicalized (bottom) variants. The rightmost columns of Tables 3 and 4 include the size and the number of dependency relations of each training corpus, respectively. The second row of Table 3 shows the precision of the predicted POS-tags.

The results of the lexicalized models in Table 3 indicate that, when parsing a new language, there are no huge differences in LAS depending on the language used as source. The largest divergence appears in the values on the EP treebank, where the Italian model achieves 65.56% while RO obtains 60.90%. However, if we momentarily ignore Romanian —whose results are very different to those obtained with the other languages—, the largest difference does not reach 4.5% (between CA and ES analyzing FR).

On average, *any* Romance language might parse a different one with LAS results between 12% and 16% lower than using the proper treebanks as training data. As mentioned, Romanian is an exception of this behaviour, since neither of the other

[13] All the parsing experiments performed in this paper were also carried out using UDPipe parser 1.0 with the swap algorithm (Straka, Hajič and Straková 2016), with very similar results than those obtained with MaltParser. On average, the results of the MaltParser models were 0.03% and 1.48% better than the UDPipe ones in monolingual and cross-lingual parsing (lexicalized), respectively. For this reason, and also because it was the system that we used for labeling the *Galician-TreeGal* treebank, the reported results are those obtained with MaltParser.

Table 3. Cross-lingual parsing results (LAS values) of Romance languages in UD 1.3. Rows correspond to source languages (train), and columns to target languages (test). Each language row contains the results of a lexicalized model (top) and a delexicalized variant (bottom). *All* models are concatenations of all the training corpora except the target one. Values in bold highlight the best source and the best monolingual source per language, underlined results are those with better performance when delexicalized, and numbers in italic are the monolingual results. The second row includes the precision of the predicted POS-tags in the target languages. The EP column has an additional value in bold (IT) since the best cross-lingual results were obtained with a variety of the same language, BP. The last column shows the size (in number of tokens) of the treebanks.

LAS	CA 98%	ES 96%	FR 96%	IT 97%	BP 97%	EP 97%	RO 95%	Deps
CA	*81.64* *74.95*	63.36 63.67	60.97 60.89	67.84 68.50	63.51 61.62	64.08 63.97	46.41 46.04	429,157
ES	67.28 65.53	*76.32* *69.71*	**65.34** 61.85	**71.41** 70.77	**65.67** 65.00	64.86 63.10	**49.62** 46.93	382,436
FR	66.75 66.43	64.69 63.17	*75.93* *68.52*	70.33 70.52	64.88 64.74	63.17 62.51	49.10 48.06	356,216
IT	66.93 66.83	**65.11** 65.25	63.81 63.15	*82.17* *75.47*	66.10 63.95	**65.56** 64.56	49.29 48.49	249,330
BP	65.15 64.41	63.09 61.87	62.16 61.31	69.30 69.70	*79.72* *72.18*	***66.84*** 63.12	46.44 46.76	239,012
EP	**68.34** 67.10	64.02 63.15	63.31 63.36	69.94 70.80	63.11 61.60	*75.95* *68.84*	48.43 46.20	214,812
RO	62.56 64.27	60.79 61.08	60.24 60.21	67.91 69.22	61.64 60.83	60.90 61.04	*70.98* *62.47*	108,618
All	69.01 67.45	**67.67** 66.00	64.17 65.71	71.17 71.77	**67.56** 65.32	**68.32** 64.56	**50.31** 50.51	—

languages achieves 50% LAS on its test data. We have to keep in mind that the Romanian treebanks showed more lexical differences with the resources of the other languages (Table 2), so this fact may have influenced the results. Also, note that

Table 4. Cross-lingual parsing results (UAS values) of Romance languages in UD 1.3. Rows correspond to source languages (train), and columns to target languages (test). Each language row contains the results of a lexicalized model (top) and a delexicalized variant (bottom). *All* models are concatenations of all the training corpora except the target one. Values in bold highlight the best source and the best monolingual source per language, underlined results are those with better performance when delexicalized, and numbers in italic are the monolingual results. The EP column has an additional value in bold (ES) since the best cross-lingual results were obtained with a variety of the same language, BP. The last column shows the number of dependency relation used in each treebank.

UAS	CA	ES	FR	IT	BP	EP	RO	Deps
CA	*85.92* *81.10*	72.88 73.40	72.26 72.33	78.13 78.02	74.21 72.05	72.88 72.99	62.88 62.74	29
ES	**77.88** 76.96	*81.47* *76.95*	**75.85** 75.38	**80.42** 80.15	**74.82** 73.59	**74.45** 73.23	**66.53** 63.68	31
FR	76.34 75.66	73.94 72.36	*81.43* *77.06*	79.13 79.23	73.15 72.11	72.69 71.45	64.37 63.02	38
IT	76.57 76.61	**74.27** 74.37	73.76 73.33	*86.56* *82.26*	74.39 72.70	73.30 73.13	62.40 61.55	35
BP	75.64 75.87	73.78 72.28	72.52 72.73	78.56 78.71	*83.62* *78.33*	***76.43*** 72.52	61.75 64.59	31
EP	76.84 75.13	73.98 72.46	74.19 74.19	78.67 79.32	72.48 71.24	*81.67* *76.06*	64.03 60.85	31
RO	73.84 75.85	71.46 71.97	73.12 73.77	77.32 78.64	71.33 71.43	70.54 71.80	*79.21* *72.82*	38
All	77.87 69.57	**75.58** 74.73	74.24 75.28	79.88 80.35	74.41 73.21	**76.89** 72.99	65.74 66.33	

the Romanian treebank has, together with FR, the highest number of dependency relations (38).

The UAS cross-lingual results follow similar tendencies to the LAS values, even though the difference between the monolingual and cross-lingual values decreases to an average of $\approx 9\%$. Furthermore, the divergences among different source languages (parsing the same target) are even smaller when computed using UAS ($\approx 2\%$).

Again, Romanian has the worst results, both as source and as target language (except in one case: it works slightly better than Catalan and Brazilian Portuguese analyzing French).[14] Apart from the large tagset of Romanian (which may have an influence on low LAS results in cross-lingual parsing), other possible factors causing poor UAS results, when compared to the other Romance languages, are structural divergences as well as the importance of lexical features. Related to this, the highest difference between the lexicalized and delexicalized models (both monolingual) occurs again in Romanian ($> 8\%$ LAS and $> 6\%$ UAS).

If we look at the size of the training corpora for cross-lingual parsing, the results suggest that, once we have a certain amount of data (which is the case of the analyzed languages), the quantity might not be critical.[15] Even though Romanian (which has the smallest training dataset) had the worst results among Romance languages, the models of Catalan (with the largest corpus) do not surpass the results of Spanish, and only had better values than European Portuguese —whose training corpus has half the size— in one case (interestingly, parsing BP).

Concerning the performance of cross-lingual transfer of delexicalized models, Table 5 shows the differences between full and delexicalized variants for each language. In each language row, the top values are LAS, while UAS values are at the bottom. If we analyze the LAS results, only in 15 out of 48 cases delexicalization produced better results (in UAS, this value increases to 20). Most of these improvements, again, occur in Romanian, whose delexicalized model works better in every Romance language (except in the LAS results of BP and FR).

Furthermore, the drops of the delexicalized models are noticeably higher than the benefits produced by the referred 15/20 parsers. The last column of Table 5 includes the average impact of the delexicalization process (ignoring the monolingual results). On average, all the delexicalized parsers behave worse than the full models in a cross-lingual setting, except for those trained on Romanian, which show average improvements of 2.61% (LAS) and 5.85% (UAS). Consequently, the results suggest that, at least for Romance languages, removing word features does not seem the best strategy for cross-lingual parsing.

This latter remark, together with the relatively lower coverage values of Romanian, took us to explore a possible correlation among lexical coverage between two languages and the impact of delexicalization. So we applied Kendall's tau coefficient (Kendall 1938) in each cross-lingual model, concluding that the coverage values (Table 2) between a source and a target language are significantly correlated with the impact of delexicalization (with $p = 0.0023$ in LAS, and $p = 0.0006$ in UAS): the higher the coverage, the lower the benefits of delexicalization, and vice-versa. In this regard, and even if this is not a rigid rule, the results indicate that some of the language pairs with less than $\approx 35\%$ of lexical coverage (European Portuguese–

[14] This only occurs in the MaltParser results. The UDPipe models of Romanian obtain the worst results in all the cross-lingual parsing experiments.

[15] In this respect, Table 8 will show the cross-lingual results using a smaller training corpus (of Galician) as source.

Table 5. Differences between lexicalized and delexicalized models for each language. Rows correspond to source languages (train), and columns to target languages (test). Each language row contains the LAS (top) and the UAS (bottom) results. Underlined numbers denote models with better results when delexicalized. The last column shows the average impact of delexicalization in the cross-lingual experiments.

Lang.	CA	ES	FR	IT	BP	EP	RO	Average
CA	-6.69	0.31	-0.08	0.66	-1.89	-0.11	-0.37	-1.48
	-4.82	0.52	0.07	-0.11	-2.16	0.11	-0.14	-1.71
ES	-1.75	-6.61	-3.49	-0.64	-0.67	-1.76	-2.69	-11.00
	-0.92	-4.52	-0.47	-0.27	-1.23	-1.22	-2.85	-6.96
FR	-0.32	-1.52	-7.41	0.19	-0.14	-0.66	-1.04	-3.49
	-0.68	-1.58	-4.37	0.10	-1.04	-1.24	-1.35	-5.79
IT	-0.10	0.14	-0.66	-6.70	-2.15	-1.00	-0.80	-4.57
	0.04	0.10	-0.43	-4.30	-1.69	-0.17	-0.85	-3.00
BP	-0.74	-1.22	-0.85	0.40	-7.54	-3.72	0.32	-5.81
	0.23	-1.50	0.21	0.15	-5.29	-3.91	2.84	-1.98
EP	-1.24	-0.87	0.05	0.86	-1.51	-7.11	-2.23	-4.94
	-1.71	-1.52	0	0.65	-1.24	-5.61	-3.18	-7.00
RO	1.71	0.29	-0.03	1.31	-0.81	0.14	-8.51	2.61
	2.01	0.51	0.65	1.32	0.10	1.26	-6.39	5.85
All	-1.56	-1.67	1.54	0.60	-2.24	-3.76	0.20	-6.89
	-0.86	-0.85	1.04	0.47	-1.20	-3.90	0.59	-4.71

Italian, Italian–Spanish, or different pairs including Romanian) are better analyzed when delexicalized.

Note, however, that although lexical coverage has an impact on delexicalization, it is just one among other factors that influence cross-lingual parsing. As an example, Romanian has better coverage values of European Portuguese than Italian or Catalan, but these latter models produce better results on EP than the RO parser.

3.2.2 Monolingual learning curve versus *cross-lingual model transfer*

The previous evaluations showed that direct cross-lingual parsing obtains reasonably good results between Romance languages. Therefore, the following set of tests

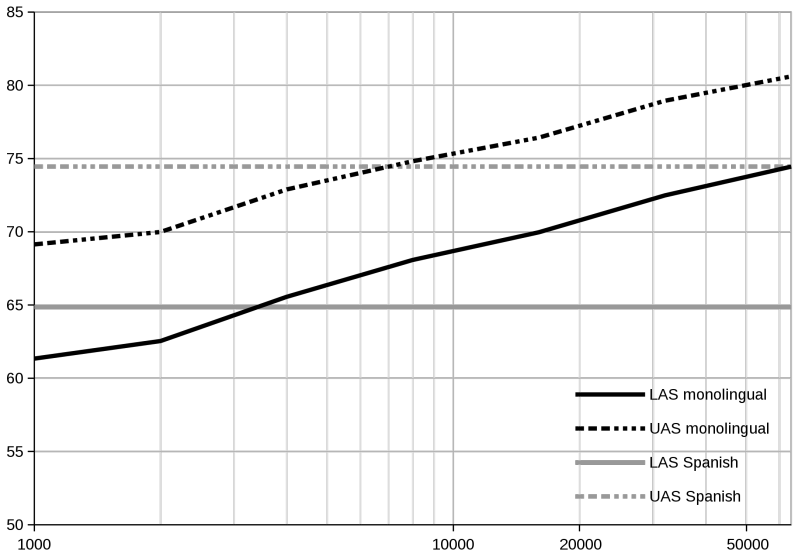

Fig. 1. LAS and UAS monolingual learning curves of European Portuguese ($0 - 64,000$ tokens) *versus* LAS and UAS results of the Spanish model in the same EP *test* data.

is aimed at knowing to what extent it is needed to manually annotate a treebank for a new language in order to train a statistical parser. Obviously, high-quality manually annotated data for the target language achieves better results than direct cross-lingual transfer, but the labeling process may be very expensive. Taking the above into account, we obtained LAS and UAS learning curves for each language and compared them with the previously explained cross-lingual results.

In order to create the learning curves we built different monolingual models by starting with just $1,000$ tokens of training data, and then adding $1,000$ more in 9 iterations, until achieving $10,000$ tokens. After that, we incrementally enlarged the amount of new data, with additions of about $2,000$, $5,000$, $25,000$ and $50,000$ tokens, before using the whole training set.[16]

Figure 1 represents the LAS and UAS learning curves of European Portuguese, together with the performance of the full Spanish model parsing the same EP data (horizontal lines). Similarly, the learning curves of Spanish can be seen in Figure 2, which also contains the results of the Italian parser on the Spanish gold standard.

As can be seen in these two example figures (and in all the other curves that have been created, omitted here due to space reasons), the best single models obtain similar LAS results in cross-lingual parsing to $> 3,000$ (EP) and $> 2,000$ (ES) manually annotated tokens of the target language. Interestingly, this value is

[16] The size of each addition is approximate, since it depends on sentence boundaries.

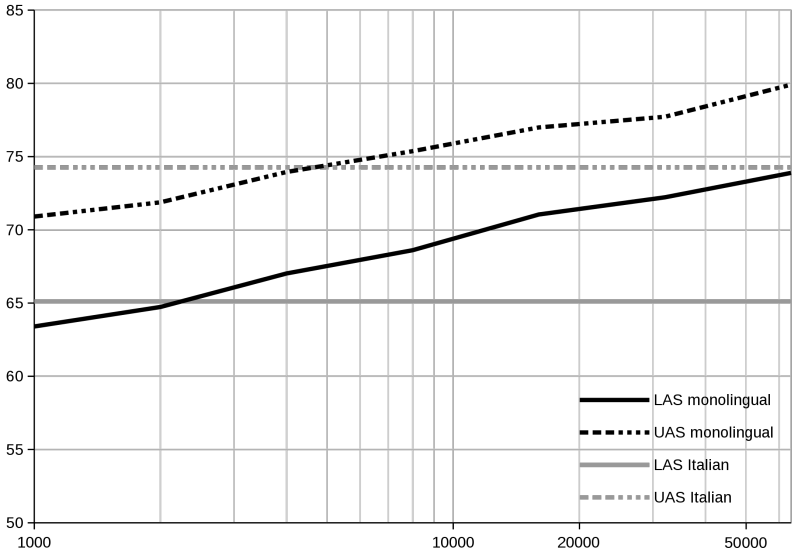

Fig. 2. LAS and UAS monolingual learning curves of Spanish ($0 - 64,000$ tokens) *versus* LAS and UAS results of the Italian model in the same ES *test* data.

very similar in all the analyzed languages (once again, except for Romanian, which outperforms both LAS and UAS cross-lingual parsing with just $1,000$ tokens).

Concerning UAS, the direct transfer of single models varies depending on the language, achieving performance values comparable to $4,000 - 8,000$ tokens of the target language.

Therefore, this evaluation indicates that one could parse, without labeled data of the target language, a new linguistic variety with similar results than those obtained with $\approx 3,000$ or $\approx 6,000$ manually annotated tokens, depending on the objectives of the syntactic analysis (LAS or UAS).

It is worth noting that the cross-lingual results shown in the previous learning curves corresponded to single models. In several cases they could be replaced by the combined (*All*) parsers, increasing the threshold of needed training data up to $\approx 9,000$ (LAS) and $> 16,000$ (UAS) tokens (e.g., in European Portuguese). These values heavily depend on the language, since the combined models do not always perform better than the best single ones (see Tables 3 and 4).

This fact leads us to carry out a previous analysis of the source and the target treebanks in order to build a focused combination that could surpass both single models and the *all versus one* concatenation. These experiments, which are partially based on the results that have been presented, are introduced in the next section, aimed at reducing the effort of creating a UD treebank for a new language.

4 Galician UD treebank

As has been shown in the previous experiments, Romance languages can be reasonably well parsed using cross-lingual resources built with harmonized annotation, such as universal dependencies. This fact allows researchers to obtain labeled data for a new language without manual annotation. Even if this process does not always provide high-quality resources for training and testing statistical parsers (and also the annotation may be biased (Berzak, Huang, Barbu, Korhonen and Katz 2016)), it can be seen as a good starting point for creating a treebank for a new language.

In this regard, this section presents a case study of the development of a new UD treebank for Galician. The treebank was built by means of cross-lingual parsing and manual correction, applying an iterative bootstrapping strategy.

First, we briefly present the main properties of the Galician language as well as of the corpus we used as source, followed by the core guidelines for annotating Galician with UD. After that, we use the Galician data to carry out similar experiments to those performed in Section 3, thus extending the cross-lingual parsing evaluations, verifying whether the new treebank has a similar behaviour to that of the other Romance corpora, and testing the proposed approach in a real scenario.

4.1 Galician language

Galician is an Indo-European linguistic variety and a part of the Western Ibero-Romance group evolved from Vulgar Latin. It derives from the medieval Galician-Portuguese language (Teyssier 1982).

Modern Galician is spoken in the Spanish Autonomous Region of Galicia by about 2.5 million people (Xunta de Galicia 2004), and it is the official language together with Spanish, which has a strong influence on different Galician characteristics such as its syntax, morphology or phonology (Freixeiro Mato 2000; Figueroa 1997).

With an eye on cross-lingual parsing, it is important to note that the current spelling of Galician is based on the Spanish one, and also that some linguists still consider Galician as a variety of (Galician-)Portuguese language (Cintra and Cunha 1984), due to its common origin and present similarity. In this regard, some studies have used Portuguese resources for NLP in Galician, taking advantage of the high relatedness between these linguistic varieties (Malvar, Pichel, Senra, Gamallo and Garcia 2010).

So, theoretically, both Portuguese and Spanish treebanks may be the best sources for cross-lingual parsing in Galician, as also the numbers in Tables 1 and 2 reinforce.

4.2 Source corpus

In order to build a UD treebank for Galician, we first selected a corpus with some manually corrected linguistic information as the starting point. This choice allowed us to reduce the effort of manually annotating all the information of the corpus.

The selected resource was XIADA, a Galician corpus with annotation of lemmas as well as POS-tags with rich morphological information reviewed by experts (Rojo,

Martínez, Noya and Barcala 2015). The current version of this corpus has 741,833 tokens, and it is divided in four sets, each one belonging to a different typology: (i) generic press, (ii) economic press, (iii) short stories, and (iv) some free unrelated sentences. The first subcorpus (of generic press, called *xeral*) is the first one that we have begun enriching with syntactic information.

We have programmed a script to convert XIADA to the CoNLL-U format, extracting both the UD POS-tags and the morphological features from the original POS tagset.

4.3 Annotation Guidelines

Before starting the addition of syntactic information to XIADA, we defined the annotation guidelines for labeling the Galician corpus using the UD 1.4 version. These guidelines are based on three main foundations (Garcia 2016):

1. Use of the UD recommendations whenever possible.
2. Use the smallest possible number of language-specific relations.
3. For labeling structures with more than one possibility of analysis, make the corpus coherent with the European Portuguese and Spanish ones (in this order).

Taking the above into account, the main properties of the UD labeling for the *Galician-TreeGal* treebank are the following:

- Tokenization: The current version of the corpus maintains the original tokenization of XIADA, which joins compound proper nouns and some multiword expressions into single tokens (e.g, "John_Lennon", or "a_as_veces", *sometimes*). As UD recommends to split these cases, these disagreements are being corrected for adapting our treebank to the UD 2.0 version.
- Pseudo-copulative verbs: Verbs belonging to this class are tagged as *cop* (copulative) when they function as *copulae* (e.g., "Miguel_Barros permanecerá relegado", *Miguel Barros will remain relegated*).
- Modal, temporal and aspectual verbs: These verbs are considered *aux* (auxiliary) of the main verb they depend on (e.g., "debe conducirnos", *should drive us*, or "deixa de ser", *stop being*). Similarly, auxiliary verbs in verbal periphrasis are also tagged as *aux* (e.g., "vai gañar", *will win*).
- Years are marked as *nmod* (nominal modifier). In further versions, they could be labeled as *nmod:tmod*, a subtype relation used in other UD treebanks for identifying temporal expressions.
- Objects: UD 1.4 recommends labeling as *dobj* (direct object) dative objects when they occur in a sentence with no explicit direct object ("a tarefa$_{\text{nsubj}}$ corresponde$_{\text{root}}$ lle$_{\text{expl}}$ a o goberno$_{\text{dobj}}$", *the task falls to the government*). Nevertheless, we preferred to mark them as *iobj* (indirect object) because it facilitates both the linguistic analysis (e.g., transitivity) and the information extraction from the treebank (and from other corpora with automatic

parsing), as in the XIADA corpus, both direct and indirect objects can be introduced by the same preposition (usually *a*): "apuntando$_{\text{root}}$ a o pobo$_{\text{dobj}}$"[17] (*pointing to the people*), and "correspóndelle a o goberno$_{\text{iobj}}$" (labeled as *dobj* above). Note, however, that these cases can be automatically converted to *dobj* in case it could be needed. Apart from that, we followed the UD recommendation of annotating *reflexive*, *reciprocal* and *expletive* pronouns as *expl* (expletive).

These guidelines (as well as the UD 1.4 recommendations) were used for labeling the first 1,000 sentences of XIADA.

4.4 Annotation process

Instead of starting the labeling process of XIADA from scratch, we applied a cross-lingual parser trained with a combination of the European Portuguese and Spanish treebanks, due to their similarities with the target language. Some labels of this combined corpus were automatically adapted to the Galician guidelines (we simplified the subtype relations by their core dependency, and automatically replaced the annotation of the reflexive pronouns with *expl*) in order to avoid the use of unwanted dependencies.

This combination was used as training data for building a cross-lingual parser, then applied to the first $\approx$ 1,000 tokens of the Galician corpus (from the *xeral* subcorpus). These sentences were manually corrected by one of the authors, and then added to the training corpus for automatically labeling the next $\approx$ 1,000 tokens. This bootstrapping process was repeated until achieving 1,000 sentences ($>$ 24,000 tokens).

4.5 Experiments

Once a gold standard treebank for Galician had been obtained, we performed various tests with the following objectives:

- To know what is the best single model for cross-lingual parsing in this language.
- To verify to what extent we can adapt and combine source treebanks in order to increase parsing performance.
- To analyze the impact of the amount of training data in Galician.
- To check how a small treebank behaves in cross-lingual parsing on the same language family.

Note that these evaluations complement those performed in Section 3, providing new information about the cross-lingual parsing in Romance languages.

As in the previous experiments, the parsers were built with the *train* sets of UD

[17] The use of the preposition in some of these cases is often analyzed as an influence of Spanish, so distinguishing between *iobj* and *dobj* is also useful for identifying this phenomenon.

Table 6. Cross-lingual results of different parsers of Romance languages on the Galician gold standard, both using predicted and gold POS-tags. In each metric row (*pred* and *gold* LAS and UAS), top numbers correspond to full (lexicalized) models, and the bottom ones to delexicalized models. Bold numbers mark the best source languages, while underlined values denote those models with better results in the delexicalized scenario.

Metric	POS	CA	ES	FR	IT	BP	EP	RO	*All*
LAS	*pred*	60.61	64.85	60.22	64.97	63.96	**65.27**	59.66	**68.72**
		61.74	62.73	61.15	65.05	61.39	63.42	60.31	64.50
	gold	67.31	71.07	66.56	71.39	70.18	71.40	64.97	**74.92**
		68.12	69.16	67.11	**71.46**	67.47	69.55	65.91	71.20
UAS	*pred*	70.71	75.17	70.33	72.20	73.88	**75.56**	69.48	**77.54**
		71.92	72.74	70.30	73.43	70.98	72.86	70.83	73.65
	gold	75.20	79.13	74.66	76.70	78.12	**79.35**	73.00	**81.22**
		76.09	77.20	74.08	77.94	75.01	76.75	74.85	78.12

1.3, using both MaltParser and UDPipe (whose results are not reported here due to their similarity). In the case of Galician, we evaluated the parsing performance using both predicted and gold POS-tags (since the selected corpus already had manually reviewed POS annotation). We used the Galician POS-tagger provided by LinguaKit (Garcia and Gamallo 2015), achieving a precision of $\approx 93\%$.[18] For testing, we used all the manually reviewed data as gold standard for Galician, except in a learning curve analysis (Figure 3).

Table 6 shows the cross-lingual results on the Galician data of direct transfer from the other Romance languages. On average, the best results are achieved using the European Portuguese treebank as source, although Italian (especially in LAS) and Spanish also reach high values. The *All* combination produces better results (both LAS and UAS) than the best single model. In some way the best single results (EP and ES) are in accordance with the linguistic relatedness these languages have with Galician, and also with the lexical distance between them. But these factors cannot be generalized, since e.g., Italian (whose lexical similarity is lower than most of the other languages) also achieves good cross-lingual results in our gold standard (better than Catalan, with higher lexical coverage).

The following evaluation analyzes the impact of two adaptations of the source treebanks aimed at approximating them to the target language. First, we applied a tool for converting the Portuguese orthography to the Galician one —which, in turn, is based on the Spanish spelling—, obtaining a new EP treebank with

[18] The output of LinguaKit —which does not use UD— was automatically converted to the universal POS-tags, so this is likely to be the reason for these low results.

Galician-like orthography (Malvar *et al.* 2010):[19] the lexicon of this new treebank covers 82.6% of the Galician one, up from 63.4% of the original version. The tool was also applied to the European Portuguese dictionary, increasing the similarity with respect to the Galician one from 29.3% to 57.5%. This new model, built by means of a lexico-orthographic adaptation, is called AP (Adapted Portuguese).

The other adaptation concerns the annotation of the reflexive pronouns in both European Portuguese and Spanish treebanks, which differs from the UD guidelines applied to Galician. In this respect, we replaced the dependency label of these pronouns (*dobj* or *iobj*) with *expl*. These new models are referred as EX variants.

The first columns of Table 7 show that the lexico-orthographic adaption of the EP treebank allows the cross-lingual parsing to increase its performance between 1.17% (UAS) and almost 2% (LAS), depending on the POS.[20] These results suggest again that lexical distance has an impact on cross-lingual parsing, even if it is not a decisive factor for transferring a parser from one language to another.

Adapting the labeling of the reflexive pronouns also boosted the European Portuguese and Spanish models, namely the LAS results, fact which is understandable because it was a simple label change. In this regard, the Spanish and Portuguese EX variants increased their LAS scores by about 0.9% and 0.1%, respectively.

Concerning the combined models, it is worth noting that the *All* parser (Table 6) still performs better than these ES and EP adaptations. However, the combination of the EX variants of Spanish and Adapted Portuguese outperforms the *All* model results by $\approx 1.1\%$.

Finally, the last evaluated combination adds the Italian treebank —which uses similar guidelines for labeling the *expl* pronouns as the Galician one— to the AP_ES_{ES} data. This new model (named $\text{AP\_ES\_IT}_{\text{EX}}$) increases both LAS ($\approx 0.4\%$) and UAS ($\approx 0.1\%$) in Galician, achieving the best results: 76.54% LAS and 82.43% UAS using the gold POS-tags provided by the original corpus.

In order to estimate the impact of the amount of training data in Galician, we created learning curves using a random split of the corrected treebank (with $\approx 5,000$ lines: 4,922 tokens and 191 sentences) as gold standard, and the remaining ($\approx 20,000$ lines: 19,297 tokens and 809 sentences) for training, adding $\approx 1,000$ tokens in each iteration.

Figures 3 and 4 show the LAS and UAS learning curves of Galician (using predicted and gold POS-tags, respectively), together with the values of the best single model and of the best combination (Table 7).[21]

With $\approx 20,000$ tokens, the Galician data achieves a parsing performance of 70.68%/77.22% (*pred*), and 78.12%/82.1% (*gold*), higher values than some of those obtained in other languages (similar, however, to Brazilian Portuguese, and lower

[19] `http://gramatica.usc.es/~gamallo/port2gal.htm`

[20] Interestingly, this approximation of the Portuguese spelling to the Galician/Spanish one also has a positive impact on Spanish parsing (using EP as source) of $\approx 0.7\%$ in both LAS and UAS.

[21] The results of the EP and of the combined models in these figures were updated using the same gold standard as the Galician learning curves.

Table 7. Results of several combined and adapted cross-lingual models on the Galician gold standard, both using predicted and gold POS-tags. AP is Adapted Portuguese, and EX means labeling adaptation of the *expl* relation.

Metric	POS	AP	ES_{EX}	AP_{EX}	AP_ES_{EX}	$AP\_ES\_IT_{EX}$
LAS	*pred*	66.92	65.78	67.03	69.76	**70.16**
	gold	73.28	71.98	73.44	76.08	**76.54**
UAS	*pred*	76.73	75.18	76.73	78.56	**78.63**
	gold	80.52	79.06	80.56	82.26	**82.43**

than Italian, which reaches almost 79%/84% with the same amount of training data and gold POS-tags).

Apart from that, Figures 3 and 4 also show that to beat the best single cross-lingual models we would need the following amount of Galician data: $\approx 3,000$ (LAS) and $\approx 6,000$ (UAS) tokens if we have predicted POS-tags, and $\approx 4,000$ (LAS) and $\approx 10,000$ (UAS) if we use gold POS-tags. These values follow the same tendency as mentioned for other Romance languages in Section 3.

However, the results obtained by the focused adaptation and combination of the source treebanks allows the parser to noticeably increase its performance, obtaining LAS results equivalent to the Galician parser at $> 16,000$ tokens, and better UAS values than using the full set of $\approx 20,000$. This combination achieves 70.63%/78.71% (*pred*) and 76.99%/82.44% (*gold*), but these values reach 71.45%/78.57% (*pred*) 78.35%/83.08% (*gold*) if we add the $\approx 20,000$ training tokens of Galician to the combined model.

The results of these evaluations show that a parser built by means of a combination of adapted treebanks is a good alternative to manually annotating a large corpus for a different language. Furthermore, a bootstrapping process with some manual revision allows researchers to evaluate the best source treebanks as well as to add these new data to improve the cross-lingual model.

Finally, we also carried out an evaluation of a parser for Galician, trained using the whole gold standard in a cross-lingual scenario: parsing all the other Romance languages. These results can be seen in Table 8, which complement those presented in Tables 3 and 4.

Even if the training data is small ($\approx 24,000$ tokens), and it has fewer dependency labels than the *test* treebanks, the results also follow the same tendency as the other Romance languages. Italian is the language on which the best results are obtained (both in LAS and UAS), followed by Portuguese and Spanish (with small differences between LAS and UAS results), and the Romanian numbers are worse than those for all the other varieties. Again, the results show that the Galician delexicalized models did not improve the parsing of any Romance language, and also that lexical

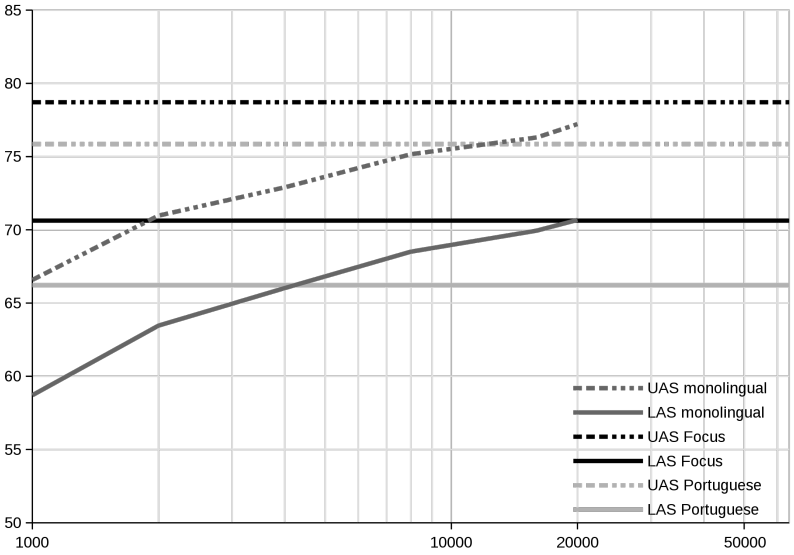

Fig. 3. LAS and UAS monolingual learning curves of Galician ($0 - 20,000$ tokens —on a $0 - 64,000$ scale, for comparison with Figures 1 and 2) *versus* LAS and UAS of the best cross-lingual single model (European Portuguese) and the focused combination (AP_ES_IT$_{\text{EX}}$). POS-tags are predicted.

Table 8. Cross-lingual parsing results on Romance languages using a model trained with the Galician gold standard (1,000 sentences). Each metric row (LAS and UAS) contains both lexicalized (top) and delexicalized results (bottom). Results were obtained using predicted POS-tags. Galician values were obtained with the $1,000$ sentences splitted in 800 (train) and 200 (test), also with predicted POS-tags ($\approx 93\%$).

Metric	CA	ES	FR	IT	BP	EP	RO	GL
LAS	58.45	60.79	57.05	66.10	61.69	62.69	45.70	*70.68*
	56.73	59.19	56.66	65.84	59.30	61.34	42.76	*63.12*
UAS	69.09	70.14	68.13	73.56	68.72	72.19	55.86	*77.22*
	66.56	68.42	66.97	72.99	66.47	71.62	53.29	*70.52*

distance does not seem to be a crucial factor for the cross-lingual analysis of varieties from the same linguistic family.

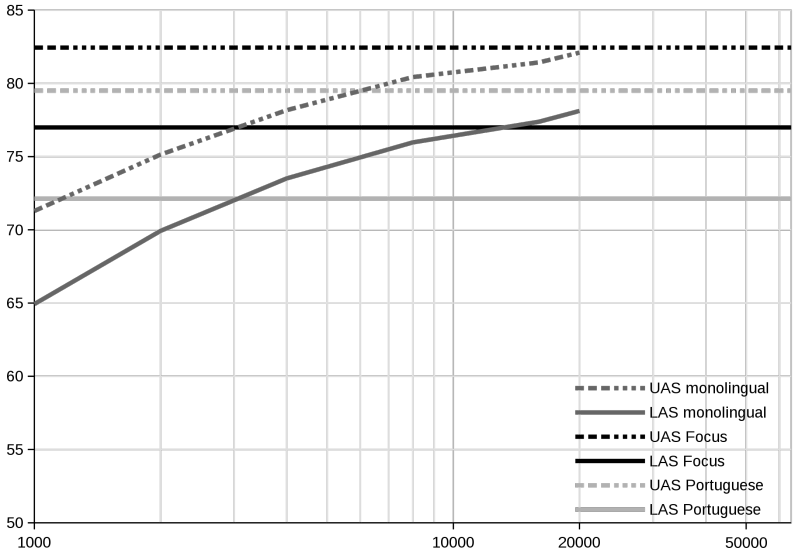

Fig. 4. LAS and UAS monolingual learning curves of Galician ($0 - 20,000$ tokens —on a $0 - 64,000$ scale, for comparison with Figures 1 and 2) *versus* LAS and UAS of the best cross-lingual single model (European Portuguese) and the focused combination (AP_ES_IT$_{\text{EX}}$). POS-tags are gold.

5 Discussion

The results of previous work on cross-lingual parsing had suggested that the good performance of this strategy on Romance languages might come from their structural similarities. Thus, in these syntactically related languages, lexical distance may play a crucial role when leveraging treebanks from one linguistic variety to another.

In this regard, the tests carried out with an adapted version of the European Portuguese treebank (automatically converted to a Spanish-like spelling, also used in Galician) seem to confirm the following hypothesis: if we approximate, in lexical terms, a source treebank to the target one, the transfer performance improves.

However, from the computed values of lexical similarity and treebank coverage (Section 3.1) together with the results of the experiments on cross-lingual parsing (Section 3.2), we can infer that lexical distance is not as important as other factors in the cross-lingual transfer of UD resources in Romance languages. Thus, the best source language for analyzing Spanish (both LAS and UAS) seems to be Italian, but this language only covers $\approx$ 2.1% of the Spanish dictionary (and 39.9% if we use the treebank-based comparison, which are the worst results on Spanish after Romanian). Moreover, European Portuguese is the variety with highest lexical coverage of Brazilian Portuguese (72% and 83%), but Spanish was the best source language for the Brazilian variety.

Concerning delexicalization, removing the tokens from the source treebanks does not improve, in most cases, cross-lingual parsing between Romance languages (in a similar way as it had been reported for three Slavic languages (Agić *et al.* 2014)). In the experiments performed, the only source language that clearly benefited from delexicalization was Romanian, whose results also followed different tendencies when compared to the other Romance languages. Furthermore, the improvements provided by lexicalized models were noticeably greater than those of the cases where the delexicalized parsers performed better. Interestingly, the impact of delexicalization is significantly correlated with the treebank lexical coverage between the source and the target varieties: the higher the coverage, the better it is to use a lexicalized parser, and vice-versa.

As pointed out in previous work (Tiedemann 2015b), the quality of the POS-tags is also critical for both cross-lingual and monolingual parsing. In our first set of experiments on Romance languages, the use of predicted POS-tags (with an average precision of $\approx 97\%$) involved drops of $\approx 1.6\%$ and $\approx 2.4\%$ in cross-lingual and monolingual parsing, respectively, if compared with parsing using gold POS-tags. In the analysis of Galician —where the predicted POS-tags only achieved $\approx 93\%$—, the differences were of ≈ 4.9 (cross-lingual) and $\approx 6.1\%$ (monolingual), again compared to the results on a corpus with gold POS-tags.

The size of the training corpus does not seem to be as crucial in cross-lingual parsing as in a monolingual scenario. Large training corpora (using combinations of treebanks with more than $300k$ tokens) achieve equivalent performance to using between $3,000$ and $7,000$ tokens of the target language (which is in accordance with the UAS results presented by Lacroix *et al.* (2016b)), so it can be said that more data from a different language does not continuously improve the syntactic analysis. Also, the cross-lingual experiments using the Galician gold standard as source (with less than $25,000$ tokens) demonstrate that even a small training corpus can perform relatively well when compared to other large datasets used for training.

Another factor that has an impact in the cross-lingual transfer of UD resources are possible divergences in the annotation of the treebanks (obviously, much lower than those that arise using other guidelines than UD). These divergences can derive (i) from linguistic phenomena and decisions of the annotators (e.g., the use of language-specific dependencies or decisions about how to label some phenomenon in a language), as well as (ii) from the properties of a treebank converted from another resource. An example of the latter can be seen in European and Brazilian Portuguese, which in their UD 1.3 version show divergences in the tokenization of contractions.

Taking the above into account and depending on the objectives of the cross-lingual parsing, some guidelines could be easily modified for improving treebank transfer (especially in terms of LAS). Moreover, the combination of different source treebanks, together with the mentioned adaptation of their annotation to the target language, noticeably improves parsing performance. This fact can be used by researchers to syntactically analyze a new language with little annotation effort.

6 Conclusions

In the present paper we have performed an analysis of cross-lingual parsing between Romance languages using Universal Dependencies. The results of this analysis have served us to start the creation of a new UD treebank for Galician, by means of a previous cross-lingual parsing with reasonably high accuracy.

The experiments carried out in this work were designed to know (i) the impact that both the lexical distance between two languages and their structural similarities have in cross-lingual parsing, (ii) whether it is beneficial to perform delexicalization in a cross-lingual scenario, and (iii) to what extent we can leverage cross-lingual resources for UD parsing in Romance languages.

In this regard, the results of several evaluations suggest that —even if it is important— the lexical distance between two languages is not a key factor for cross-lingual parsing, so other properties such as syntactic differences or divergences in the annotation guidelines —or even the textual typology of the training corpus— play a crucial role in the performance of the transfer of syntactic models. Apart from that, the experiments on Romance languages have also shown that the delexicalization process is not useful in most scenarios of cross-lingual parsing, and that its impact is significantly correlated with the lexical coverage between the source and target treebanks.

After comparing the performance of transferred parsers with the learning curves of monolingual ones, we can state that a direct cross-lingual transfer using just one source treebank behaves similar to $\approx 3,000$ (in LAS results) and $\approx 7,000$ (UAS values) manually revised tokens of the target language. However, a focused combination and adaptation of different treebanks to the target language can boost these results to more than double.

Thus, the case study on the development of a new UD treebank for Galician has shown that combining and adapting resources from related languages (in our case from European Portuguese, Castilian Spanish and Italian) leads us to parse a Galician corpus with results of 76.99% (LAS) and 82.44% (UAS) without data of the target language (which would be equivalent to the use of a Galician training corpus of $\approx 16,000$ tokens —for similar LAS results— or $> 20,000$ —for UAS). Moreover, the addition of $20,000$ tokens of Galician to these source treebanks increases the numbers up to 78.35% (LAS) and 83.08% (UAS), which are competitive results when compared to those obtained in other languages with large training corpora.

In sum, the experiments carried out in this paper point out that, depending on the objectives and on the available resources, it is possible to start the creation of a UD treebank (or a parser) for a new language by leveraging resources from related languages, reducing notoriously the manual effort with respect to building a treebank from scratch.

However, there are still some open questions which need further research. A more detailed analysis between varieties of the same language (such as EP and BP), and also between different treebanks of the same linguistic variety, will shed some light on the effect of syntactic differences in cross-lingual parsing. Also, extending the

experiments performed in this paper to other linguistic families could also bring interesting information concerning cross-lingual transfer.

Specifically for Galician, it would be important to enlarge and improve the manual annotation of the corpus, in order to have a better view of its learning curve compared to the cross-lingual focused combinations.

In this respect, it is worth noting that this article contributes to the UD project by releasing a new treebank for Galician (*Galician-TreeGal*), with a manually corrected gold standard of 1,000 sentences (24,219 tokens).

References

Agić, Ž., Hovy, D., and Søgaard, A. 2015. If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2015). Short Papers*. Beijing: Association for Computational Linguistics, pp 268–72.

Agić, Ž., Johannsen, A., Plank, B., Martínez Alonso, H., Schluter, N., and Søgaard, A. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–12.

Agić, Ž., Tiedemann, J., Merkler, D., Krek, S., Dobrovoljc, K., and Moze, S. 2014. Cross-lingual Dependency Parsing of Related Languages with Rich Morphosyntactic Tagsets. In *Proceedings of the EMNLP'2014 Workshop on Language Technology for Closely Related Languages and Language Variants*. Doha: Association for Computational Linguistics, pp. 13–24.

Ammar, W., Mulcaire, G., Ballesteros, M., Dyer, C., and Smith, N. A. 2016. Many Languages, One Parser. *Transactions of the Association for Computational Linguistics*, 4:431–44.

Aufrant, L., Wisniewski, G., and Yvon, F. 2016. Zero-resource Dependency Parsing: Boosting Delexicalized Cross-lingual Transfer with Linguistic Knowledge. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka: Association for Computational Linguistics, pp 119–30.

Bejček, E., Panevová, J., Popelka, J., Straňák, P., Ševčíková, M., Štěpánek, J., and Žabokrtskỳ, Z. 2012. Prague Dependency Treebank 2.5–a revisited version of PDT 2.0. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*. Bombay: Association for Computational Linguistics, pp. 231–46.

Berzak, Y., Huang, Y., Barbu, A., Korhonen, A., and Katz, B. 2016. Anchoring and Agreement in Syntactic Annotations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*. Austin: Association for Computational Linguistics, pp. 2215–24.

Cintra, L. F. L. and Cunha, C. 1984. *Nova gramática do português contemporâneo*. Lisbon: Livraria Sá da Costa.

Cohen, S. B., Das, D., and Smith, N. A. 2011. Unsupervised Structure Prediction with Non-Parallel Multilingual Guidance. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*. Edinburgh: Association for Computational Linguistics, pp. 50–61.

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the 9th edition of the International Language Resources and Evaluation Conference (LREC 2014)*, volume 14. Reykjavik: European Language Resources and Evaluation, pp. 4585–92.

de Marneffe, M.-C., MacCartney, B., Manning, C. D. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th edition of the International Language Resources and Evaluation Conference (LREC 2006)*, volume 6. Portorož: European Language Resources and Evaluation, pp. 449–54.

de Marneffe, M.-C. and Manning, C. D. 2008. The Stanford typed dependencies representation. In *Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation at COLING 2008*. Manchester: Association for Computational Linguistics, pp. 1–8.

Duong, L., Cohn, T., Bird, S., and Cook, P. 2015. A Neural Network Model for Low-Resource Universal Dependency Parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*. Lisbon: Association for Computational Linguistics, pp. 339–48.

Durrett, G., Pauls, A., and Klein, D. 2012. Syntactic transfer using a bilingual lexicon. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)*. Jeju Island: Association for Computational Linguistics, pp. 1–11.

Erjavec, T. 2012. MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Language Resources and Evaluation*, 46(1):131–42.

Figueroa, T. V. 1997. Estruturas fonéticas de tres dialectos de Vigo. *Verba*, (24):313–32.

Freixeiro Mato, X. R. 2000. *Gramática da lingua galega II. Morfosintaxe*. Vigo: A Nosa Terra.

Ganchev, K., Gillenwater, J., and Taskar, B. 2009. Dependency grammar induction via bitext projection constraints. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009)*, volume 1. Singapore: Association for Computational Linguistics, pp. 369–77.

Garcia, M. 2016. Universal Dependencies Guidelines for the Galician-TreeGal treebank. Technical Report, LyS Group, Universidade da Coruña.

Garcia, M., and Gamallo, P. 2015. Yet Another Suite of Multilingual NLP Tools. In *Languages, Applications and Technologies. Communications in Computer and Information Science*, 563. Switzerland: Springer: 65–75.

Gimpel, K. and Smith, N. A. 2014. Phrase Dependency Machine Translation with Quasi-Synchronous Tree-to-Tree Features. *Computational Linguistics*, 40(2):349–401.

Guo, J., Che, W., Yarowsky, D., Wang, H., and Liu, T. 2015. Cross-lingual Dependency Parsing Based on Distributed Representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing: Association for Computational Linguistics, pp. 1234–44.

Guo, J., Che, W., Yarowsky, D., Wang, H., and Liu, T. 2016. A Representation Learning Framework for Multi-Source Transfer Parsing. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI 2016)*. Phoenix: Association for the Advancement of Artificial Intelligence, pp. 2734–40.

Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., and Kolak, O. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(03):311–25.

Kendall, M. G., 1938. A new measure of rank correlation. *Biometrika* 30(1/2):81–93.

Klein, D. and Manning, C. D. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL 2004)*. Barcelona: Association for Computational Linguistics, pp. 479–86.

Lacroix, O., Aufrant, L., Wisniewski, G., and Yvon, F. 2016a. Frustratingly Easy Cross-Lingual Transfer for Transition-Based Dependency Parsing. In *Proceedings of the 15th*

Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2016). San Diego: Association for Computational Linguistics, pp. 1058–63.

Lacroix, O., Wisniewski, G., and Yvon, F. 2016b. Cross-lingual Dependency Transfer: What Matters? Assessing the Impact of Pre- and Post-processing. In *Proceedings of the Workshop on Multilingual and Cross-lingual Methods in NLP at the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2016)*. San Diego: Association for Computational Linguistics, pp. 20–9.

Malvar, P., Pichel, J. R., Senra, Ó., Gamallo, P., and Garcia, A. 2010. Vencendo a escassez de recursos computacionais. Carvalho: Tradutor Automático Estatístico Inglês-Galego a partir do corpus paralelo Europarl Inglês-Português. *Linguamática*, 2(2):31–8.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K. B., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., and Lee, J. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*. Sofia: Association for Computational Linguistics, pp. 92–7.

McDonald, R., Petrov, S., and Hall, K. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*. Edinburgh: Association for Computational Linguistics, pp. 62–72.

McEnery, T. and Hardie, A. 2011. *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Muniz, M. C., Nunes, M. D. G. V., and Laporte, E. 2005. UNITEX-PB, a set of flexible language resources for Brazilian Portuguese. In *Workshop on Technology on Information and Human Language (TIL)*. São Leopoldo: Sociedade Brasileira de Computação, pp. 2059–68.

Naseem, T., Barzilay, R., and Globerson, A. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1 (ACL 2012)*. Jeju Island: Association for Computational Linguistics, pp. 629–37.

Nguyen, T.-V. T., Moschitti, A., and Riccardi, G. 2009. Convolution kernels on constituent, dependency and sequential structures for relation extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, volume 3. Singapore: Association for Computational Linguistics, pp. 1378–87.

Nivre, J. 2004. Incrementality in deterministic dependency parsing. In *Proceedings of the Workshop on Incremental Parsing: Bringing Engineering and Cognition Together*. Barcelona: Association for Computational Linguistics, pp. 50–7.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the 10th edition of the International Language Resources and Evaluation Conference (LREC 2016)*. Portorož: European Language Resources and Evaluation, pp. 1659–66.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., and Marsi, E. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(02):95–135.

Padró, L. and Stanilovsky, E. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the 8th edition of the International Language Resources and Evaluation Conference (LREC 2012)*. Istambul: European Language Resources and Evaluation, pp. 2473–9.

Petrov, S., Das, D., and McDonald, R. 2012. A Universal Part-of-Speech Tagset. In *Proceedings of the 8th edition of the International Language Resources and Evaluation Conference (LREC 2012)*. Istambul: European Language Resources and Evaluation, pp. 2089–96.

Rasooli, M. S. and Collins, M. 2015. Density-driven cross-lingual transfer of dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*. Lisbon: Association for Computational Linguistics, pp. 328–38.

Rojo, G., Martínez, M. L., Noya, E. D., and Barcala, F. M. 2015. Corpus de adestramento do Etiquetador/Lematizador do Galego Actual (XIADA), versión 2.6. `http://corpus.cirp.es/xiada/corpus_xiada_2_6.tar.gz`. Centro Ramón Piñeiro para a Investigación en Humanidades.

Rosa, R., Masek, J., Marecek, D., Popel, M., Zeman, D., and Zabokrtskỳ, Z. 2014. HamleDT 2.0: Thirty Dependency Treebanks Stanfordized. In *Proceedings of the 9th edition of the International Language Resources and Evaluation Conference (LREC 2014)*. Reykjavik: European Language Resources and Evaluation, pp. 2334–41.

Rosa, R., Žabokrtskỳ, Z. 2015. KLcpos3 - a Language Similarity Measure for Delexicalized Parser Transfer. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Beijing: Association for Computational Linguistics, pp. 243–9.

Rosa, R., Žabokrtskỳ, Z. 2015b. MSTParser Model interpolation for multi-source delexicalized transfer. In *Proceedings of the 14th International Conference on Parsing Technologies*. Bilbao: Association for Computational Linguistics, pp. 71–5.

Smith, D. A. and Eisner, J. 2009. Parser adaptation and projection with quasi-synchronous grammar features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, volume 2. Singapore: Association for Computational Linguistics, pp. 822–31.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. 2013. Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*. Seattle: Association for Computational Linguistics, pp. 1631–42.

Søgaard, A. 2011. Data point selection for cross-language adaptation of dependency parsers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers (ACL HLT 2011)*, volume 22. Portland: Association for Computational Linguistics, pp. 682–6.

Søgaard, A., Agić, Ž., Martínez Alonso, H., Plank, B., Bohnet, B., and Johannsen, A. 2015. Inverted indexing for cross-lingual NLP. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing: Association for Computational Linguistics, pp. 1713–22.

Straka, M., J. Hajič, and J. Straková, J. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož: European Language Resources Association, pp. 4290–7.

Täckström, O., McDonald, R., and Uszkoreit, J. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2012)*. Montreal: Association for Computational Linguistics, pp. 477–87.

Täckström, O., McDonald, R., and Nivre, J. 2013. Target Language Adaptation of Discriminative Transfer Parsers. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2013)*. Atlanta: Association for Computational Linguistics, pp. 1061–71.

Teyssier, P. 1982. *História da língua portuguesa.* Lisbon: Livraria Sá da Costa.

Tiedemann, J. 2014. Rediscovering annotation projection for cross-lingual parser induction. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*. Dublin: Association for Computational Linguistics, pp. 1854–64.

Tiedemann, J. 2015. Improving the cross-lingual projection of syntactic dependencies. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, No. 109. Vilnius: Linköping University Electronic Press, pp. 191–9.

Tiedemann, J. 2015b. Cross-Lingual Dependency Parsing with Universal Dependencies and Predicted PoS Labels. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*. Uppsala: Association for Computational Linguistics, pp. 340–9.

Tiedemann, J., and Agić, Ž. 2016. Synthetic Treebanking for Cross-Lingual Dependency Parsing. *Journal of Artificial Intelligence Research (JAIR)*, 55:209–48.

Tiedemann, J., Agić, Ž, and Nivre, J. 2014. Treebank Translation for Cross-Lingual Parser Induction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning (CoNLL 2014)*. Baltimore: Association for Computational Linguistics, pp. 130–40.

Tsarfaty, R. 2013. A Unified Morpho-Syntactic Scheme of Stanford Dependencies. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*. Sofia: Association for Computational Linguistics, pp. 578–84.

Vilares, D., Alonso, M. A., and Gómez-Rodríguez, C. 2016. One model, two languages: training bilingual parsers with harmonized treebanks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*. Berlin: Association for Computational Linguistics, pp. 425–31.

Xunta de Galicia (AA.VV). 2004. *Plan xeral de normalización da lingua galega.* Xunta de Galicia, Consellería de Educación e Ordenación Universitaria, Dirección Xeral de Política Lingüística.

Yarowsky, D., Ngai, G., and Wicentowski, R. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research (HLT 2001)*. San Diego: Association for Computational Linguistics, pp 1–8.

Zeman, D. 2008. Reusable Tagset Conversion Using Tagset Drivers. In *Proceedings of the 6th edition of the International Language Resources and Evaluation Conference (LREC 2008)*. Marrakech: European Language Resources and Evaluation, pp. 213–18.

Zeman, D., Dušek, O., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtskỳ, Z., and Hajič, J. 2014. HamleDT: Harmonized multi-language dependency treebank. *Language Resources and Evaluation*, 48(4):601–37.

Zeman, D. and Resnik, P. 2008. Cross-Language Parser Adaptation between Related Languages. In *Proceedings of the Workshop on NLP for Less Privileged Language at the 3rd International Joint Conference on Natural Language Processing (IJCNLP 2008)*. Hyderabad: Asian Federation of Natural Language Processing, pp. 35–42.

Zhang, Y., and Barzilay, R. 2015. Hierarchical Low-Rank Tensors for Multilingual Transfer Parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*. Lisbon: Association for Computational Linguistics, pp. 1857–67.