Investigation of Transfer Languages for Parsing Latin: Italic Branch vs. Hellenic Branch

Antonia Karamolegkou and Sara Stymne
Department of Linguistics and Philology
Uppsala University
Sweden
antoniakrm16@gmail.com ,sara.stymne@lingfil.uu.se

Abstract

Choosing a transfer language is a crucial step in cross-lingual transfer learning. In much previous research on dependency parsing, related languages have successfully been used. However, when parsing Latin, it has been suggested that languages such as ancient Greek could be helpful. In this work we parse Latin in a low-resource scenario, with the main goal to investigate if Greek languages are more helpful for parsing Latin than related Italic languages, and show that this is indeed the case. We further investigate the influence of other factors including training set size and content as well as linguistic distances. We find that one explanatory factor seems to be the syntactic similarity between Latin and Ancient Greek. The influence of genres or shared annotation projects seems to have a smaller impact.

1 Introduction

There have been multiple projects exploiting the benefits of multilingual dependency parsing[1] (Ammar et al., 2016; Ponti et al., 2018) and especially the use of transfer learning in low-resource scenarios (Guo et al., 2015; Ponti et al., 2018). Transfer learning in the context of parsing low-resource languages uses knowledge from a transfer language in order to parse the low-resource target language (Pan and Yang, 2010). Determining the optimal transfer language for any target language is a crucial step usually leading to the selection of a language that belongs to the same language family as the target language (Dong et al., 2015; Guo et al., 2016; Dehouck and Denis, 2019). However, language proximity is not always the best criterion, since there are other properties that could lead to better results such as the content of the syntactic, geographical, or phonological distances, which is confirmed by studies both in Machine Translation (Bjerva et al., 2019) and Syntactic Parsing (Lin et al., 2019). Smith et al. (2018) noted that for Latin, it was useful to group it with other ancient languages such as ancient Greek and Gothic, but they did not provide a comparison with other potential transfer languages.

We perform an investigation of parsing Latin in a low-resource setting, with the goal of investigating if Greek languages are better as transfer languages than Italic languages. We also explore the role of factors such as treebank size, treebank content and linguistic distance measures. We find that ancient Greek, and also modern Greek, are indeed a better choice as transfer languages for Latin than the related Italic languages Italian and French. We further show that while using ancient Greek data from the same annotation project is preferable, it is not the sole cause of the strong results, since good results are had also across different annotation projects. These results also hold for different training data sizes. Finally we note that ancient Greek is syntactically more similar to Latin than Italian, which can be an explanatory factor.

2 Related Work

Multilingual parsing has been an active topic of research over the last decade, but there is a limited number of studies that focus on transfer language selection. There are works that include language selection techniques for dependency parsing such as using a typological database to choose transfer languages based on their typological weight similarities to the target language (Søgaard and Wulff, 2012). Similarly, Agić (2017) use a part-of-speech sequence similarity method between the source and target language. A more detailed investigation on transfer language selection is performed by Lin et al. (2019). They attempt to build

[1] often mentioned as cross-lingual dependency Parsing

models that rank languages based on linguistic distance measures in order to predict the optimal transfer languages. Another option is to choose the most suitable single-source parser among a set of parsers, either at the level of language (Rosa and Žabokrtský, 2015) or for individual sentences (Litschko et al., 2020), often based on part-of-speech patterns.

3 Experimental Setup

Our main aim is to investigate the impact of different transfer languages on low-resource Latin parsing. In addition we explore the impact of training data size and content, as well as the connection to a number of distance measures between languages.

3.1 Parser

To train and evaluate the parsing models we use UUparser[2] (de Lhoneux et al., 2017). It is a transition-based parser using a two-layer BiLSTM to extract features, and a multi-layer perceptron to predict transitions. Words are represented by a word embedding, a character embedding and a treebank embedding. Treebank embeddings represent the source treebank of each token, and has been shown to be effective both in a multilinugal (Smith et al., 2018) and monolinugal (Stymne et al., 2018) settings. An arc-hybrid transition system with a swap transition and a static-dynamic oracle (de Lhoneux et al., 2017) is used. It can handle non-projectivity, which is quite common in Latin.

We keep the default hyperparameter settings of the parser from Smith et al. (2018). All embeddings are initialized randomly at training time. For evaluation, we use Labeled Attachment Score (LAS). All models are trained for 30 epochs. The best epoch is selected according to the best average development set LAS score.

3.2 Language and Treebank selection

Latin is used as the target/low-resource language and we choose two transfer languages from each language family. The languages from the Italic branch, Italian and French, belong to a branch with languages historically evolved from Latin and are relatively closely related to the target language. Ancient Greek and its descendant language, modern Greek, on the other hand, belong to the Hellenic branch of the Indo-European Languages, and these languages are not as closely related as languages from the Italic branch (Nordhoff and Hammarström, 2011; Dehouck and Denis, 2019).

We use corpora from the Universal Dependencies (UD) project (Nivre et al., 2020) version 2.5 (Zeman et al., 2019). The data is sampled by choosing the first n sentences from each treebank. In two cases the Latin and ancient Greek datasets come from the same annotation projects. The Perseus treebanks have parallel texts from the Bible and classical writers (Bamman and Crane, 2011), while the PROIEL treebanks have similar texts from the new testament, but they also include texts from different authors (Haug and Jøhndal, 2008). Both the text overlap and supposedly similar annotation styles between these treebanks have been hypothesized as one possible cause of the fact that combining Latin and ancient Greek is useful (Smith et al., 2018).

We also want to investigate the effect of the size of training data, both for the target and transfer treebanks. For the target treebank, where we focus on a low-resource scenario, we use 250 and 500 sentences, respectively, while we use 2.5K and 10K sentences for the transfer languages. In the latter scenario we focus on Italian and ancient Greek, due to the small size of the modern Greek treebank and the poor performance with French as a target language. Table 1 contains information about the treebanks. All development and test sets include 250 sentences.

3.3 Linguistic Distances

Linguistic distance defines how distant a set of languages is based on genealogical, geographical, or typological features created with linguistic analysis (Lin et al., 2019). Littell et al. (2017) provide various vector information on linguistic features in URIEL Typological database which can be used to calculate how distant are the languages.[3] In this work using the URIEL database we use the following linguistic distances:[4]

- **Geographic distance** (d_{geo}): The spherical distance among languages on Earth's surface, divided by the diametrically opposite Earth's distance. The language points are abstractions, and not precise facts, derived from

[2]https://github.com/UppsalaNLP/uuparser

[3]https://github.com/antonisa/lang2vec

[4]Inventory distance was not used in this study, since it is similar to phonological distance, but the phonological feature vectors are derived from PHOIBLE database

Language	Treebank	Size	Genre	Exp1	Exp2	Exp3
Latin	la_Perseus	2,273	Bible, Classical texts	250	500	500
	la_proiel	18,411	New Testament, Classical texts	250	500	500
	la_ittb	26,977	Classical texts	250	500	500
Italian	it_isdt	14,167	News, legal, wiki	2,500	2,500	10,000
	it_vit	10,087	News, Politics, Literary	2,500	2,500	10,000
Ancient Greek	grc_Perseus	13,919	Bible, Classical texts	2,500	2,500	10,000
	grc_proiel	17,080	New testament, Classical texts	2,500	2,500	10,000
Modern Greek	el_gdt	2,521	News, Politics, Health	2,500	2,500	–
French	fr_ftb	18,535	News, Politics	2,500	2,500	–

Table 1: Treebank information and the number of sentences used in each experiment.

	d_{geo}	d_{gen}	d_{fea}	d_{pho}	d_{syn}
Italian	0.0	0.5	0.7	0.2	0.52
French	0.1	0.68	0.8	0.54	0.71
ancient Greek	0.0	0.8	0.3	0.2	0.35
modern Greek	0.1	1	0.8	0.59	0.64

Table 2: Distances between Latin and the other languages according to the URIEL typological database

existing databases with declarations on language location (Littell et al., 2017).

- **Genetic distance** (d_{gen}): The genealogical distance among languages, according to the hypothesized world language family tree in the Glottolog catalogue (Nordhoff and Hammarström, 2011).
- **Phonological distance** (d_{pho}): The cosine distance among the phonological vectors extracted from the World Atlas of Language Structure (WALS) and Ethnologue databases (Dryer and Haspelmath, 2013; Lewis, 2009).
- **Syntactic distance** (d_{syn}): The cosine distance among vectors mostly extracted from the syntactic structures of the languages according to WALS (Dryer and Haspelmath, 2013).
- **Featural distance** (d_{fea}): The cosine distance between feature vectors from a combination of the linguistic features described above (geographic, genetic, syntactic, phonological, inventory) extracted from the URIEL database.

All the leveraged information from the URIEL database can be found in Table 2, where the values range from 0.0 to 1.0.; numbers close to 0.0 represent proximity and vice versa. The language codes are based on the ISO-639-3 codes.[5] In order to examine whether these linguistic distances are related to the parsing results, the Pearson Correlation Coefficient will be used.

[5]https://iso639-3.sil.org/code_*tables*/639/*data*

4 Results

Table 3 shows results from training a monolingual model for each Latin treebank with a small amount of data. As expected, the scores are quite low, given the limited training data size, but there is a large improvement from doubling the data from 250–500 sentences of up to 8.4 LAS points. There is a large difference in performance between the treebanks, where the Persues treebank seems to have the most challenging test set.

Table 4 shows results with a cross-lingual model with 2.5K transfer language sentences and Table 5 shows the results with 10K transfer language sentences. In all cases, one of the ancient Greek treebanks give the best results, with improvements of up to 16.9 LAS points compared to the monolingual baseline for Latin PROIEL. In all but one case, modern Greek also surpasses the results of all Italic treebanks, and also beats all monolingual baselines. Italian helps for the PROIEL and ITTB Latin treebanks, but in most cases hurts slightly for the Persues treebank. French, on the other hand leads to very poor results in all cases, mostly giving worse results than the monolingual baseline.

Concerning the impact of training data size, we can usually see a large improvement, when doubling the target data, just as in the monolingual case. Overall the improvements are larger for the poor models than for the stronger ones. Increasing the size of the transfer language from 2.5K to 10K further improves the results in most cases when ancient Greek is used as transfer language. The improvements are typically smaller than when increasing the size of the target language, though. When using Italian as the transfer language, however, the results do not show much change compared to using less Italian data, sometimes even leading to worse results. It thus seems that using more data from the transfer language is only useful for transfer languages that are a good fit to the

Training sentences:	250	500
la_Perseus	17.9	26.1
la_proiel	39.9	43.1
la_ittb	33.1	41.6

Table 3: LAS scores for monolingual training with 250 and 500 sentences.

	la_Perseus		la_proiel		la_ittb	
Target sent.	250	500	250	500	250	500
it_isdt	19.9	25.9	46.5	55.6	38.1	46.4
it_vit	17.8	24.5	44.2	54.7	36.9	44.3
grc_Perseus	**30.1**	**32.4**	**50.4**	58.1	**39.9**	45.4
grc_proiel	**27.6**	**31.9**	**50.9**	**60**	**40.4**	**47.6**
el_gdt	23.6	27.2	48.5	**58.4**	36.6	**46.6**
fr_ftb	12.8	22.8	39.8	50.3	13.7	40.2

Table 4: LAS scores for multilingual experiments with 2.5K sentences from the transfer language, and 250 or 500 sentences from the target language.

target language.

For Latin PROIEL and Perseus, where there are ancient Greek treebanks from the corresponding annotation projects, it is always preferable to use the matching treebank. However, the gaps are typically not large, ranging from 0.4 to 3.9 LAS points, with the scores for the non-matching treebank in most cases beating the scores for treebanks from all other languages. Also for the Latin ITTB treebank, the scores for both non-matching ancient Greek treebanks are among the highest scores, with the PROIEL treebank being the best match. This indicates that the impact of annotation project, with content and annotation styles matching, adds to the performance, but is not the main explanatory factor for the usefulness of ancient Greek. It is also worth noting that the treebanks for Italian VIT, modern Greek and French have similar content, but very different parsing results, indicating that language choice is more important than the genres of the treebanks.

Table 6 shows Pearson correlations between the distance measures and the parsing scores for the Latin PROIEL treebank using 500 sentences and 2.5K transfer language sentences. There is a strong negative correlation of -.76 between the syntactic distance of the languages and the parsing results, even though it is not significant. This finding seems reasonable since syntactic features of a language are intuitively important for parsing. Ancient Greek and Latin actually have a closer syntactic distance than Italian and Latin, see Table 2. The same applies to the featural distance, which is

	la_Perseus	la_proiel	la_ittb
it_isdt	24.3	55.3	44.7
it_vit	24	55.4	42.7
grc_Perseus	**36.9**	**60.7**	**46.6**
grc_proiel	**33**	**62.3**	**47.3**

Table 5: LAS scores from multilingual experiments with 10K sentences from the transfer language and 500 from the target language

	R	Strength	P-value
d_{geo}	-0.47	weak	0.34
d_{gen}	0.57	moderate	0.23
d_{fea}	-0.91	strong	0.011
d_{pho}	-0.44	weak	0.382
d_{syn}	-0.76	strong	0.073

Table 6: Pearson correlation and p-value between parsing scores and linguistic distance measures for the Latin PROIEL treebank.

a combination of various features (including syntactic, phonological, inventory, geographic, and genealogical), and has a strong significant negative correlation of -.91. While this finding is quite intuitive, it is contrary to the finding of Lin et al. (2019) who found that geographic and genetic distances were more important than syntactic or featural distance, however, for 0-shot parsing with a higher number of languages. It is, however, in accordance with (Bjerva et al., 2019) who indicated that structural similarity is a better predictor of language representation similarities compared to genetic similarity. The strong performance for Hellenic languages is especially interesting since they do not share script with Latin, which means that the character embeddings in UUparser are less useful than for Italian.

5 Conclusion

We have shown that using Hellenic languages is preferable to using Italic languages when training a multilingual parsing model for Latin in a low resource scenario. While we see the best results when we use ancient Greek treebanks from the same annotation project as the Latin treebanks, we also see very competitive results when training across annotation projects, mostly surpassing all other languages explored. We also see that it is more useful to increase the training data size of the target language than the transfer language, and that increasing the size of the target language is only useful when it is a good match. Finally we show that there are strong correlations between the parsing result and the featural and syntactic dis-

tance of the target and transfer language, which could explain the usefulness of ancient Greek, the most syntactically similar language to Latin in our sample.

In this study we only explored a low-resource setting, using a limited amount of Latin data. It would be interesting to see if the findings hold also when we use all available data, as indicated by the results of Smith et al. (2018). We would also like to add pre-trained word embeddings, either cross-lingual static embeddings, or multilingual contextual embeddings, to see what the impact is, compared to our current experiments where we do not use any pre-trained embeddings. Another direction would be to investigate if ancient Greek is a good transfer language for Latin also for other tasks, which might be less sensitive to syntactic distance.

References

Željko Agić. 2017. Cross-lingual parser selection for low-resource languages. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 1–10, Gothenburg, Sweden. Association for Computational Linguistics.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

David Bamman and G. Crane. 2011. The Ancient Greek and Latin dependency treebanks. In *Language Technology for Cultural Heritage*.

Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019. What do language representations really represent? *Computational Linguistics*, 45(2):381–389.

Mathieu Dehouck and Pascal Denis. 2019. Phylogenic multi-lingual dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 192–203, Minneapolis, Minnesota. Association for Computational Linguistics.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. 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)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Matthew S Dryer and Martin Haspelmath. 2013. *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 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)*, pages 1234–1244, Beijing, China. Association for Computational Linguistics.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2016. A representation learning framework for multi-source transfer parsing. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2734–2740. AAAI Press.

Dag T. T. Haug and Marius L. Jøhndal. 2008. Creating a parallel treebank of the old indo-european bible translations. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH)*, pages 27–34.

M. Paul Lewis, editor. 2009. *Ethnologue: Languages of the World*, sixteenth edition. SIL International, Dallas, Texas, USA.

Miryam de Lhoneux, Sara Stymne, and Joakim Nivre. 2017. Arc-hybrid non-projective dependency parsing with a static-dynamic oracle. In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 99–104, Pisa, Italy. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Robert Litschko, Ivan Vulić, Željko Agić, and Goran Glavaš. 2020. Towards instance-level parser selection for cross-lingual transfer of dependency parsers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3886–3898, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Sebastian Nordhoff and Harald Hammarström. 2011. Glottolog/langdoc: Defining dialects, languages, and language families as collections of resources. In *Proceedings of the First International Workshop on Linked Science 2011*, volume 783 of *CEUR Workshop Proceedings*.

S. J. Pan and Q. Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Edoardo Maria Ponti, Roi Reichart, Anna Korhonen, and Ivan Vulić. 2018. Isomorphic transfer of syntactic structures in cross-lingual NLP. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1531–1542, Melbourne, Australia. Association for Computational Linguistics.

Rudolf Rosa and Zdeněk Žabokrtský. 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)*, pages 243–249, Beijing, China. Association for Computational Linguistics.

Aaron Smith, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan Shao, and Sara Stymne. 2018. 82 treebanks, 34 models: Universal dependency parsing with multi-treebank models. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Anders Søgaard and Julie Wulff. 2012. An empirical study of non-lexical extensions to delexicalized transfer. In *Proceedings of COLING 2012: Posters*, pages 1181–1190, Mumbai, India. The COLING 2012 Organizing Committee.

Sara Stymne, Miryam de Lhoneux, Aaron Smith, and Joakim Nivre. 2018. Parser training with heterogeneous treebanks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 619–625, Melbourne, Australia. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, et al. 2019. Universal dependencies 2.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.