Typological Approach to Improve Dependency Parsing for Croatian Language

Diego Alves
FFZG, University of Zagreb
Zagreb, Croatia
dfvalio@ffzg.hr

Boke Bekavac
FFZG, University of Zagreb
Zagreb, Croatia
bbekavac@ffzg.hr

Marko Tadić
FFZG, University of Zagreb
Zagreb, Croatia
marko.tadic@ffzg.hr

Abstract

This article presents the results of the experiments concerning different typological approaches considering syntactic structures with the aim to identify similar languages which can be combined with Croatian to improve UAS and LAS metrics when using a deep learning tool. From the eight selected languages coming from different linguistic families and genera, we showed that Slovene and Irish are the best candidates which improved significantly dependency parsing results. Slovak is the only language presenting negative synergy when combined with Croatian. Both typological approaches presented in this study, using quantitative data concerning rules from context-free grammar extracted from corpora using Marsagram tool and using syntactic features from lang2vec language vectors, did not allow us to explain the observed synergy when the different languages were combined. The traditional genealogical classification does not explain either the improvement provided by Irish or the negative impact of the Slovak language in both considered metrics.

1 Introduction

Since the 1980s, NLP field has increasingly relied on statistics, probability, and machine learning methods which require a large amount of linguistic data. Furthermore, from 2015 onward, the usage of deep learning techniques has been dominant in this field (Otter et al., 2018). These approaches require a large amount of annotated data which can be problematic for some languages considered as low-resourced.

Linguistic manual annotation of texts can be very costly (Fort et al., 2014), therefore, other solutions for improving PoS-MSD (Part-of-Speech and Morphosyntactic descriptors) and Dependency Parsing tagging scores have been proposed. One way to overcome this issue is to combine data from similar languages according to established typological classifications (Smith et al., 2018)(Alzetta et al., 2020). Although some improvement can be observed, most of these studies, however, do not present a deep analysis of typological features which may play a significant role when corpora are combined. Furthermore, none has considered statistics concerning possible (or impossible) syntactic constructions inside the available training datasets as a possible typological classification.

Therefore, our aim in this paper is to propose an innovative way of considering typological aspects when combining datasets for dependency parsing improvement. The study is focused on the Croatian language and its association with several European languages from different linguistic families. Our hypothesis is that by comparing syntactic rules automatically extracted from Universal Dependencies datasets by inferring context-free grammars (together with its statistics), we are able to classify languages according to these syntactic criteria. Combining languages closer in terms of syntactic structure to train deep learning parsing models should improve final LAS and UAS metrics.

The paper is composed as follows: Section 2 presents related work to this topic. Section 3 describes the campaign design: datasets selection, typological classification strategies, and extrinsic evaluation using trained models; Section 4 present the obtained results which are discussed in Section 5. In Section 6 we provide conclusions and possible future directions for research.

2 Related Work

Combining data from multiple languages has the ultimate aim of creating Universal Morphological and Dependency Parsing systems by considering the relationship between different languages morphology and syntactic structure (Otter et al., 2018). The Universal Dependencies (UD) framework (Nivre et al., 2020) proposes a robust set of rules for annotating parts of speech, morphological features, and syntactic dependencies across different human languages, and is inserted in this strategy as it allows multi-lingual data to be annotated with the same set of tags.

Udify tool (Kondratyuk and Straka, 2019) proposes an architecture aimed for PoS-MSD and dependency parsing tagging integrating Multilingual BERT language model[1] (104 languages) (Pires et al., 2019). It can be fine-tuned using specific corpora (mono or multilingual) to enhance overall results. The authors showed that by using a corpus composed by the association of all Universal Dependencies training sets, there is a considerable improvement in the results of parsing for low-resourced languages. Nevertheless, the authors did not conduct an experiment based on typological features to test the potential of the model when only similar languages are combined.

An interesting example of the usage of typological features to improve results of NLP methods was presented by (Üstün et al., 2020). They proposed UDapter, a tool that uses a mix of automatically curated and predicted typological features obtained via URIEL language typology database (Littell et al., 2017). These features were used as direct input to a neural parser as language-typology vectors and results showed that they were crucial for the improvement of the dependency parsing accuracy for low-resourced languages. A similar study, using different deep learning architecture had been performed by (Ammar et al., 2016), however, in both cases, there is no detailed analysis on which features were the most relevant.

The above-mentioned language typology database offers the lang2vec tool (Littell et al., 2017) which provides uniform, consistent and standardized information about languages drawn from typological, geographical and phylogenetic databases. Its sources include WALS (Dryer and Haspelmath, 2013), PHOIBLE (Moran and McCloy, 2019), Ethnologue (Lewis, 2009), and Glottolog (Hammarström et al., 2020). While (Üstün et al., 2020) used lang2vec in an automatized way to cluster languages, (Naseem et al., 2012) selected specific typological features to fine-tune effective automatic annotation of data from languages with no available training sets.

The strategy proposed by (de Lhoneux et al., 2018) concerns sharing 27 parameters using Uppsala parser using pairs of languages from the same linguistic family, showing that general typological classifications can already contribute to enhancing final results on low-resourced languages. They also observed that by combining features even from unrelated languages overall scores can be improved in some specific cases. Nevertheless, as it is the case for most of the similar studies, no specific linguistic analysis was presented in order to explain why languages coming from different families can improve overall results.

An interesting and detailed experiment was conducted by (Lynn et al., 2014) concerning the Irish language. The authors performed a series of cross-lingual direct transfer parsing for the Irish language and the best results were achieved when using Indonesian, a language from the Austronesian language family. They also propose some analysis considering similarities between the treebanks of both languages in terms of dependency parsing labels, however, detailed statistical analysis of corpora and complete comparison of specific typological features were not carried out.

Concerning syntax more specifically, (Alzetta et al., 2020) presented a study whose main objective was to identify cross-lingual quantitative trends in the distribution of dependency relations in annotated corpora from distinct languages by using an algorithm (LISCA - LInguiStically– driven Selection of Correct Arcs) (Dell'Orletta et al., 2013) capable of detecting patterns of syntactic constructions in large datasets. Only four Indo-European languages were scrutinised but some interesting insights concerning languages peculiarities were observed.

Another approach to extract and to compare syntactic information from treebanks is proposed by (Blache et al., 2016) by inferring context-free grammars (together with its statistics) from syntactic struc-

[1] https://github.com/google-research/bert/blob/master/multilingual.md

Language	Linguistic Family	Genus	UD Dataset	Corpus size (K tokens)
Bulgarian	Indo-European	Slavic	BTB	156
Croatian	Indo-European	Slavic	SET	199
Greek	Indo-European	Greek	GDT	63
Hungarian	Uralic	Ugric	Szeged	42
Irish	Indo-European	Celtic	IDT	115
Latvian	Indo-European	Baltic	LVTB	220
Maltese	Afro-Asian	Semitic	MUDT	44
Slovak	Indo-European	Slavic	SNK	106
Slovene	Indo-European	Slavic	SSJ	140

Table 1: Selected Languages, corresponding Linguistic Families and Genus, and corresponding UD datasets information (v.2.7).

tures inside annotated corpora. The analysis comparing 10 different languages showed the potential of the proposed tool (MarsaGram), however, like (Alzetta et al., 2020), the authors do not explore how this information can be used to improve existing NLP tools, which is the main objective of this paper.

3 Campaign Design

In this section, we describe the corpora that have been selected, the typological classification methods that were considered, and the experimental design used to evaluate the effects on dependency parsing metrics of the combination of different training datasets.

3.1 Languages and Datasets selection

As mentioned before, the focus of this study is the Croatian language. The main idea is to combine its training dataset with other European languages to improve UAS and LAS scores. From all 24 European Union official languages, we have chosen the following ones for our experiments: Bulgarian, Greek, Hungarian, Irish, Latvian, Maltese, Slovak, and Slovene. We have decided to work with European languages as this ensemble already provides languages from diverse linguistic families and allows us to test our hypothesis.

All the selected languages have Universal Dependencies datasets (version 2.7) and were chosen as they have only one UD corpus. Slovene is the exception, it has two different UD datasets but one is composed of spoken language, therefore, the other available corpus (written language) was used. The choice of including Slovene is also due to its genealogical proximity to Croatian.

Table 1 presents the languages involved in the experiment, with the respective linguistic family and genus (from Worl Atlas of Language Structure Online[2]) and the size of their UD corpora (Version 2.7).

3.2 Typological Analysis

In this study, we propose to compare the chosen languages using two different typological approaches. One considers the statistical analysis of context-free grammar rules extracted from dependency parsing trees using the software Marsagram, while the other strategy uses information from lang2vec tool language vectors.

3.2.1 Statistical comparison of Dependency Parsing Trees

Marsagram is a tool for exploring treebanks, it extracts context-free grammars (CFG) from annotated datasets that allow statistical comparison between languages as proposed by (Blache et al., 2016). We have used the latest release of this software[3] developed by ORTOLANG. This software has been chosen as it allows easy extraction and analysis of surface word order patterns which have never been used before as a way to interpret results of language combination for training deep learning models.

[2]https://wals.info/languoid/genealogy

[3]Available at: https://www.ortolang.fr/market/tools/ortolang-000917

Approach	Number of Rules
All rules, all properties	714 399
All rules, only linear properties	96 789
Common rules, all properties	1 912
Common rules, only linear properties	247

Table 2: Different approaches for the statistical typological approach and the respective number of the considered syntactic rules.

For this analysis, we have combined train, development, and test sets, and extracted quantitative information about its syntactic properties for each language. Distance matrices were, then, generated using R.

This software identifies four types of properties: precede, require, exclude, and unicity. The extracted syntactic rules contain information concerning part-of-speech and dependency parsing label as well as the associated property type. For example: *NOUN-conj_precede_CCONJ-cc_DET-det* which means that a *CCONJ* which has the dependency relation *cc* precedes a *DET* with *det* as dependency label in the context of a node having *NOUN* as head. Marsagram also indicates the frequency of each rule inside the corpus.

In the previous work (Blache et al., 2016), the authors proposed two different analyses: considering all possible properties or taking into account only the linear property (precede). They have shown that the linear approach was better for classifying languages typologically as results were closer to classic genealogical lists. Nevertheless, in our study, we still consider both scenarios in order to analyse which one is better when the aim is to combine languages for improving dependency parsing metrics.

For each language, Marsagram generates a specific set of rules and the percentage corresponding to its frequency inside the corpus. Some rules are common to all languages and some of them appear only in one or a few corpora. Therefore, the typological classification can be done by considering all possible identified rules (frequency equal to zero for languages in which the rule does not appear) or, considering only the rules present in all corpora (common rules).

Thus, we have 4 different possible comparisons which are presented in Table 2 together with the number of syntactic rules and considered properties used in each one.

3.2.2 Comparison using Language Vectors

Lang2vec is a library[4] that allows simple queries of the URIEL database which are presented as language vectors (Littell et al., 2017). For this study, we have considered syntactic information (*syntax_average* option). For example: *S_NEGATIVE_SUFFIX* which gives a value of 1 if the language has a negative suffix and 0 if it does not have, and *S_SUBJECT_AFTER_VERB*, 1 for languages in which the subject appears after the verb and 0 otherwise.

One disadvantage of this tool is that for some languages, not all information is available. If all official European Union are considered, the number of existing syntactic properties in lang2vec is 103. However, Croatian has values for only 12 of them. As our focus is this language, we have considered the syntactic features for which Croatian has associated values [5]. The distance between languages was calculated using cosine similarity. Among the other selected languages, only Maltese and Slovak do not have values for all these features and, therefore, were discarded for this specific analysis.

3.3 Training Models

We have selected Udify tool to train dependency parsing modules using the combined corpora as it allows fine-tuning of Multilingual BERT language model and for which the authors showed that multilingual

[4] https://pypi.org/project/lang2vec/

[5] Selected syntactic features: *S_SVO, S_SOV, S_VSO, S_VOS, S_OVS, S_OSV, S_SUBJECT_BEFORE_VERB, S_SUBJECT_AFTER_VERB, S_OBJECT_AFTER_VERB, S_OBJECT_BEFORE_VERB, S_SUBJECT_BEFORE_OBJECT, and S_SUBJECT_AFTER_OBJECT*.

Combination	Number of added sentences	Ratio
Smaller	450	94% Croatian, 6% other language
Medium	909	88% Croatian, 12% other language
Larger	1 662	81% Croatian, 19% other language

Table 3: Information concerning the different combinations of the Croatian training set and other languages.

corpus can potentially enhance overall results (specially for under-resourced languages) (Kondratyuk and Straka, 2019). Training parameters were defined as:

- Number of epochs: 80
- Warmup: 500
- Baseline training set: Croatian SET
- Development and test sets: Croatian SET

Our baseline is the result obtained by training Udify using the Croatian Universal Dependencies training set (SET) which contains 6 914 sentences. To obtain statistical significance, for each test using a specific dataset we have conducted 6 experiments varying the Random Seed value in the configuration file of Udify: standard value, 13370 (proposed by the developers), 10, 100, 1000, and 100000. For each test, we have calculated the standard deviation and the p-value when compared to the baseline.

As explained before, the objective is to combine the Croatian dataset with annotated data of the other selected languages. We have combined its training set with three different sizes of the other languages annotated data as detailed in table 3.

One problem is that each training set has a different size, thus, to have homogeneity in terms of size to allow results to be compared, we have decided to add the first 909 sentences of the second language training corpus to the Croatian one. This value corresponds to the size of the Hungarian training set (the smallest one among the chosen languages and, therefore, being totally used), this limitation concerning the Hungarian language is what determined the ratio of all language combinations.

The final size of the combined training sets is 7 823 sentences (88% Croatian and 12% from the other language).

4 Results

In this section, we present the typological classification of the languages obtained using the methods presented previously followed by the results of the combination of the different datasets.

4.1 Typological classification using statistics from syntactic trees

Tables 5 shows the distance between each language and Croatian concerning the different choices of rules and properties selection using Marsagram.

In the scenario considered in the second column of table 4 (considering all rules and properties), we observe that Slovene and Slovak are closest to Croatian (all Slavic languages), however, Bulgarian, which is also Slavic, comes after Greek, Maltese, Hungarian and Latvian which are from different genealogical families.

The third column of table 4 shows the results of the analysis of all rules but considering only the linear properties (*precede*). Again, Slovene and Slovak are the most similar to Croatian, followed by Greek. When only linear properties are considered, Latvian and Irish are classified as closer to Croatian compared to the previous scenario. Bulgarian, again beside being Slavic, occupies the second to last position.

When only common rules are considered (fourth column of table 4), Slovene is still the closest one to Croatian, however, in this case, Bulgarian is classified as much closer. Slovak loses the second position to Latvian. Maltese, Greek, and Hungarian are the most distant languages.

Language	d(All/All)	d(All/Linear)	d(Common/All)	d(Common/Linear)
Slovene	68.0	21.4	4.0	1.1
Slovak	69.4	24.0	4.9	1.2
Greek	70.0	24.5	5.9	1.6
Maltese	73.7	24.8	5.8	1.1
Hungarian	77.2	26.4	6.2	1.5
Latvian	78.5	24.5	4.3	1.0
Bulgarian	80.0	25.5	4.4	1.1
Irish	80.6	25.2	5.3	1.7

Table 4: Distance from Croatian using Marsagram results, first word correspond to the type of rules considered and the second word to the type of properties.

Language	Distance
Slovene	0.01
Bulgarian	0.03
Latvian	0.11
Greek	0.11
Hungarian	0.12
Irish	0.51

Table 5: Cosine distances calculated between Croatian language vector and other languages considering syntactic features from lang2vec.

Finally, when only common rules and linear properties are taken into account (fifth column of Table 4), we observe important changes in the classification. Slovene is no longer classified as the closest to Croatian. Maltese, and Bulgarian are the closest ones (second and third position) behind Latvian only.

Typological classification differs when different sets of rules and properties are considered. Slovene and Slovak are most of the time the closest languages to Croatian which was expected considering that they are all Slavic languages. These results show that it is difficult to determine which type of choice concerning rules and properties is the most adapted for syntactical language classification. Results may be biased by size, genre, and also the type of sentences composing the corpora (for example: length of sentences and syntactic complexity).

4.2 Typological classification using similarity between language vectors

By using cosine distance between the language vectors built with syntactic features from lang2vec, we obtain the classification present in Table 5.

Both Slavic languages (Slovenian and Bulgarian) are the most similar to Croatian, therefore more coherent to the typical genealogical classification of languages. As mentioned before, Slovak, also Slavic, does not have values for the analysed features and was therefore excluded from this comparison. Latvian, Greek, and Hungarian have similar distances, but much higher than the ones concerning Slavic languages and Irish is the most distant one.

4.3 Dependency parsing results with combined corpora

In tables 6, 7, and 8 we present the UAS and LAS values obtained when Udify was trained using the Croatian training set alone (baseline) and with the combined datasets (Croatian associated with another language) with three different ratios, as well as the delta when compared to the baseline. Each result corresponds to the mean value calculated with the six different trials using different Random Seed initial values. Highlighted results concern the experiments for which p-value is inferior to 0.05. Development and test sets were purely Croatian.

When the smaller ratio is used to train Udify (94% Croatian, 6% other language), we observe that only Bulgarian, Greek and Irish contribute positively in increasing both UAS and LAS metrics. Association

Training Corpus	UAS	delta UAS	LAS	delta LAS
Croatian (baseline)	92.32	-	88.99	-
Croatian + Bulgarian (Smaller)	**92.38**	**0.06**	**89.05**	**0.06**
Croatian + Greek (Smaller)	**92.40**	**0.09**	**89.07**	**0.08**
Croatian + Hungarian (Smaller)	92.33	0.02	88.98	-0.01
Croatian + Irish (Smaller)	**92.42**	**0.11**	**89.09**	**0.10**
Croatian + Latvian (Smaller)	92.39	0.07	88.98	0.00
Croatian + Maltese (Smaller)	92.32	0.01	88.97	-0.01
Croatian + Slovak (Smaller)	92.24	-0.07	**88.89**	**-0,09**
Croatian + Slovene (Smaller)	92.36	0.05	89.02	0.04

Table 6: UAS and LAS metrics obtained by training Udify with different training datasets: Croatian alone and associated with other languages (94% Croatian, 6% other language).

Training Corpus	UAS	delta UAS	LAS	delta LAS
Croatian (baseline)	92.32	-	88.99	-
Croatian + Bulgarian (Medium)	92.35	0.03	89.02	0.03
Croatian + Greek (Medium)	92.35	0.03	89.98	-0.01
Croatian + Hungarian (Medium)	92.33	0.02	89.01	0.02
Croatian + Irish (Medium)	**92.43**	**0.12**	**89.07**	**0.08**
Croatian + Latvian (Medium)	92.26	-0.06	88.92	-0.06
Croatian + Maltese (Medium)	92.36	0.04	88.97	-0.01
Croatian + Slovak (Medium)	**92.21**	**-0.11**	**88.89**	**-0,09**
Croatian + Slovene (Medium)	**92.42**	**0.10**	**89.09**	**0.10**

Table 7: UAS and LAS metrics obtained by training Udify with different training datasets: Croatian alone and associated with other languages (88% Croatian, 12% other language).

of Croatian and Irish being the one providing the highest increase. Negative synergy is only observed for LAS metric when Croatian is combined with Slovak.

For the medium ratio (88% Croatian, 12% other language), combinations of Croatian with Irish and with Slovene provide a positive synergy. As for the smaller ratio, when Croatian is combined with Slovak, there is a negative synergy which is, this time, observed for both UAS and LAS.

Concerning the larger ratio (81% Croatian, 19% other language), again the combination of Croatian and Slovak decrease significantly both UAS and LAS metrics. The corpus composed by both Croatian and Irish no longer provides a positive synergy. The only significant increase is obtained for LAS metric when Croatian is combined with Slovene.

5 Discussion

By analysing the UAS and LAS results presented in the previous section, it is possible to observe that Bulgarian, Greek, Irish, and Slovene training corpora have the potential to improve UAS and LAS metrics when combined with the Croatian training dataset. However, results strongly depend on the ratio between Croatian sentences and the other combined language. Bulgarian and Greek languages provided a positive synergy only for the smaller ratio, while the combination with Irish was positive for both smaller and medium ratios. Slovene did not improve the metrics for the smaller ratio but had a positive impact for both medium and larger ones. What is clear for all three ratios is the strong negative impact of Slovak when this language is associated with Croatian.

In their article, (Kondratyuk and Straka, 2019) presented results for Croatian from a model which was trained combining 124 languages. The obtained UAS and LAS values are respectively 91.10 and 86.78. It is possible to see that all the models presented in this study are higher than these, even for our baseline and for the combination with Slovak. Thus, it seems that finding typological ways to combine languages wisely and on the smaller scale is more effective.

Training Corpus	UAS	delta UAS	LAS	delta LAS
Croatian (baseline)	92.32	-	88.99	-
Croatian + Bulgarian (Larger)	92.33	0.01	88.97	-0.02
Croatian + Greek (Larger)	92.34	0.03	89.99	-0.01
Croatian + Hungarian (Larger)	-	-	-	-
Croatian + Irish (Larger)	92.37	0.05	89.03	0.04
Croatian + Latvian (Larger)	92.33	0.01	88.97	-0.02
Croatian + Maltese (Larger)	-	-	-	-
Croatian + Slovak (Larger)	**92.20**	**-0.11**	**88.83**	**-0,16**
Croatian + Slovene (Larger)	92.36	0.04	**89.06**	**0.07**

Table 8: UAS and LAS metrics obtained by training Udify with different training datasets: Croatian alone and associated with other languages (81% Croatian, 19% other language). Hungarian and Maltese training corpora do not have enough annotated sentences to be combined with Croatian in this specific ratio.

Moreover, (de Lhoneux et al., 2018) included the Croatian language in their study and the LAS obtained was 77.9, also inferior to the values in our experiments. However, the combined languages were not the same.

In terms of typology, if we consider the traditional genealogical classification of languages, we can state that being part of the same linguistic family and genus do not guarantee a positive synergy when corpora are combined. Even though Bulgarian and, especially, Slovene can improve the final results when combined with Croatian, Slovak, which is also from the same genus, is the only language with a negative influence in all tested scenarios. Moreover, Irish, which is from a different genus is a good candidate for improving UAS and LAS results when combined with Croatian.

If we consider the classifications established using Marsagram, it is not possible to find any correlation between the classification lists considering the syntactic criteria with the observed results from Udify. Slovene is the closest language to Croatian when all rules are considered (with all properties considered and only linear ones too) and also when only common rules are compared. However, the calculated distances between Irish and Croatian do not explain the improvement obtained by associating both languages. Also, Slovak does not appear as being the most distant language when compared to Croatian, a result that would explain the negative synergy observed when its corpus is combined with the Croatian dataset.

One possible explanation for this lack of correlation may come from the fact that the distances were calculated using the results obtained by Marsagram which were composed of rules coming from the whole Universal Dependency datasets for each language. However, when Udify experiments were conducted, only a small part of the respective corpora have been used. Therefore, a more precise correlation may be possible if distances are calculated using only the sentences that have been added to the combined training corpus. Another aspect that may need further research concern the homogeneity of extracted rules using Marsagram from subcorpora of a dataset from a single language. It may be possible that the variation inside a corpus may be higher than when two different languages are compared. This case could be accommodated with the usage of controlled content, i.e. parallel corpora of languages investigated. However, this is not always available, particularly for under-resourced languages.

Furthermore, the selected corpora have different sizes and different contents. It may impact heavily the type of syntactic patterns that were extracted using Marsagram. The number of patterns obtained seems to be correlated with the size of the corpus. A comparison using parallel corpora could avoid this bias.

Moreover, positive synergies may not be caused by the whole ensemble of extracted rules but maybe by specific syntactic relations which are shared by the associated languages. Further qualitative analysis of similarities between Irish and Croatian Marsagram results should be conducted.

When analysing the typological classification using lang2Vec, Slovene and Bulgarian are the closest

to Croatian, which we can relate to the positive synergy observed in Udify results. However, Irish is the most distant one which is contradictory with the improvement obtained for both UAS and LAS in two different scenarios. Also, as Slovak does not have values for the selected syntactic features, it was impossible to check whether the combination with Croatian has any negative impact. Thus, even though this tool is a powerful instrument to compare languages, in the approach described here, it seems limited. The idea of combining corpora to improve parsing is most useful for under-resourced languages, and, unfortunately, some of these languages are also under-resourced in terms of language vector information in lang2vec. For example, from the 103 possible syntactic features, the Croatian language only has values for 12 of them.

Considering all the aspects presented above, we can affirm that none of the genealogical and typological approaches were able to explain precisely what was observed when different languages were combined to Croatian.

6 Conclusions and Perspectives

In this article, we presented different approaches to identify languages that can be combined with Croatian to improve dependency parsing evaluation metrics (UAS and LAS) when using Udify deep learning tool.

The possible typological classifications were compared to the results obtained when combining the Croatian training dataset to other European languages from different linguistic families to train Udify models. Three different association ratios were used.

We showed that the association of Croatian with Irish and Slovene languages showed the best positive synergy, increasing UAS and LAS for at least two different combination ratios. Moreover, from all selected languages, the only one which decreased significantly in both metrics is Slovak.

These results show that the classical genealogical classification of languages is not enough to explain the observed phenomena. Slovak and Slovene are from the same linguistic family and genus as Croatian but with totally different impacts on the final results. Also, the Irish language does not belong to the same genus as Croatian, nevertheless, it helped improve UAS and LAS significantly.

The two typological approaches proposed in this paper, using rules from a context-free grammar with Marsagram and comparing lang2vec syntactic features of language vectors, also did not allow us to predict the results obtained when languages were combined. Slovene is identified as the closest language to Croatian in three out of four different analysed Marsagram scenarios. However, the classification of Irish and Slovak does not correspond to the influence these languages have when combined with Croatian. Moreover, the lang2vec classification shows Irish as being the least similar to Croatian, and, unfortunately, Slovak was not analysed due to the lack of syntactic information of this language in this tool.

The study presented in this article was conducted only for Croatian, therefore, we intend to test this approach with other under-resourced languages, also enlarging the selection of languages to be combined to understand better the existing synergies and, also, possible exceptions as the one that has been identified in this article concerning the association between Croatian and Irish.

For future research we will check the quality of Slovak data because it consistently differ from other Slavic languages although genealogically and culturally Slovak is closely connected to Croatian.

Furthermore, our aim is to conduct a more detailed analysis concerning Marsagram results, first, checking the homogeneity of rules extracted from different subcorpora of the same language, and, secondly, using only the sentences that were appended to the combined training corpora to calculate the distances.

Acknowledgements

The work presented in this paper has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 812997 and under the name CLEOPATRA (Cross-lingual Event-centric Open Analytics Research Academy).

References

Chiara Alzetta, Felice Dell'Orletta, Simonetta Montemagni, Petya Osenova, Kiril Simov, and Giulia Venturi. 2020. Quantitative linguistic investigations across universal dependencies treebanks. In Johanna Monti, Felice Dell'Orletta, and Fabio Tamburini, editors, *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021*, volume 2769 of *CEUR Workshop Proceedings*. CEUR-WS.org.

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.

Philippe Blache, Stéphane Rauzy, and Grégoire Montcheuil. 2016. MarsaGram: an excursion in the forests of parsing trees. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2336–2342, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Miryam de Lhoneux, Johannes Bjerva, Isabelle Augenstein, and Anders Søgaard. 2018. Parameter sharing between dependency parsers for related languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4992–4997, Brussels, Belgium, October-November. Association for Computational Linguistics.

Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2013. Linguistically–driven selection of correct arcs for dependency parsing. *Computación y Sistemas*, 17.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Karen Fort, Bruno Guillaume, and Hadrien Chastant. 2014. Creating zombilingo, a game with a purpose for dependency syntax annotation. In *Gamification for Information Retrieval (GamifIR'14) Workshop*.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2020. *Glottolog 4.3*. Jena.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally.

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

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, April. Association for Computational Linguistics.

Teresa Lynn, Jennifer Foster, Mark Dras, and Lamia Tounsi. 2014. Cross-lingual transfer parsing for low-resourced languages: An Irish case study. In *Proceedings of the First Celtic Language Technology Workshop*, pages 41–49, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Steven Moran and Daniel McCloy, editors. 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–637, Jeju Island, Korea, July. 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, May. European Language Resources Association.

Daniel W. Otter, Julian R. Medina, and Jugal K. Kalita. 2018. A survey of the usages of deep learning in natural language processing. *CoRR*, abs/1807.10854.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July. 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, October. Association for Computational Linguistics.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. UDapter: Language adaptation for truly Universal Dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2302–2315, Online, November. Association for Computational Linguistics.