Proceedings of the 15th Workshop on Building and Using Comparable Corpora (BUCC 2022) @LREC2022 , pages 33–42
Marseille, June 25, 2022

Multilingual Comparative Analysis of Deep-Learning Dependency Parsing Results Using Parallel Corpora

Diego Alves, Božo Bekavac, Marko Tadić
Faculty of Humanities and Social Sciences, University of Zagreb
Ulica Ivana Lučića 3, 10000, Zagreb
{dfvalio, bbekavac, mtadic}@ffzg.hr

Abstract

This article presents a comparative analysis of dependency parsing results for a set of 16 languages, coming from a large variety of linguistic families and genera, whose parallel corpora were used to train a deep-learning tool. Results are analyzed in comparison to an innovative way of classifying languages concerning the head directionality parameter used to perform a quantitative syntactic typological classification of languages. It has been shown that, despite using parallel corpora, there is a large discrepancy in terms of LAS results. The obtained results show that this heterogeneity is mainly due to differences in the syntactic structure of the selected languages, where Indo-European ones, especially Romance languages, have the best scores. It has been observed that the differences in the size of the representation of each language in the language model used by the deep-learning tool also play a major role in the dependency parsing efficacy. Other factors, such as the number of dependency parsing labels may also have an influence on results with more complex labeling systems such as the Polish language.

Keywords: dependency parsing, typology, multilingualism

1. Introduction

Dependency parsing is an important part of Natural Language Processing (NLP) chains which consist of annotating raw texts from tokenization to dependency relations. This specific task concerns the process to analyze the grammatical structure in a sentence and identify syntactic heads as well as the type of the relationship between them (syntactical analysis) (Jurafsky and Martin, 2009).

Since the 1980s, the NLP field has increasingly relied on statistics, probability, and machine learning methods which require a large amount of linguistic data. Unlike other annotation tasks such as POS tagging, dependency parsing annotation is much more complex and expensive. Furthermore, from 2015 onward, the usage of deep learning techniques has been dominant in this field which has provided a great improvement in overall results even for under-resourced languages (Otter et al., 2018).

The focus of the majority of studies regarding dependency parsing is on new methods to improve overall results using existing data. Methods and algorithms are compared in terms of results, however, usually, there is no comparison or analysis of the obtained results considering the syntactic complexity of languages. This is due to the fact that, in general, systems are trained using different data-sets (in terms of size and content) for different languages. The lack of data for under-resourced languages is the usual explanation for worse results with respect to dependency parsing metrics. It is undeniable that the amount of training data plays a crucial role in the performance of deep learning models, however, it is not clear how models deal with different structures of languages when the same type and amount of linguistic data is provided for different languages.

Therefore, our aim in this paper is to propose a multilingual analysis of dependency parsing results considering the syntactic structure of languages (using head directionality parameter). By using parallel annotated corpora, our idea is to scrutinize parsing results obtained with a deep learning model to check how different language structures influence the performance of the chosen tool. Also, our aim is to correlate it with the syntactical characterization of languages concerning the specific syntactic feature of head and dependent position. As presented by Jurafsky and Martin (Jurafsky and Martin, 2021) this is one of the features that plays a role in the performance of graph-based parsers. The idea is to check the degree of influence in dependency parsing of this specific language characteristic.

The paper is composed as follows: Section 2 presents an overview of the related work to this topic. Section 3 describes the campaign design: language and data-sets selection, dependency parsing annotation, and syntactic typological characterization; 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

The Universal Dependencies (UD) framework (Nivre et al., 2016) proposes a robust set of rules for annotating parts of speech, morphological features, and syntactic dependencies across different human languages allowing multi-lingual data to be annotated with the same set of tags. If the framework can be used to annotate, in a homogeneous way, different languages, there is a lack of annotated parallel corpora that can be used for more precise multilingual comparison studies.

As mentioned in the previous section, many studies

concerning dependency parsing metrics present multilingual perspectives but results cannot be compared in terms of language structure as training sets come from different sources and present different sizes and genres. An example of it is the article presenting UDify tool (Kondratyuk and Straka, 2019) which is a software conceived for PoS-MSD and dependency parsing tagging integrating Multilingual BERT (mBERT) language model (104 languages) (Pires et al., 2019). It is also the case for mainstream NLP tools such as Stanford Core NLP (Manning et al., 2014), UDPipe (Straka and Straková, 2017), sPacy (Honnibal and Montani, 2017) and NLPCube (Boroș et al., 2018).

In the article "Evaluating Language Tools for Fifteen EU-official Under-resourced Languages" (Alves et al., 2020), the authors have compared tools to check the reproducibility of presented results in the official respective articles. The authors, however, used the same heterogeneous corpora as the developers of the tools to train the models, the focus was on the analysis of the discrepancy between obtained results and claimed ones by the tool creators.

Parallel corpora are most often used in machine translation (MT) tasks. Therefore, many studies considering the quality, availability, and performance of tools using this type of data-set do not consider dependency parsing. It is the case of the studies presented by Heiki-Jaan Kaalep and Kaarel Veskis (Kaalep and Veskis, 2007) and Wolfgang Teubert (Teubert, 1996). When parallel corpora are considered for parsing, the analysis is, most generally, focused on the improvement of overall results, not on language comparison, as in (Kuhn, 2005).

Liu and Xu proposed a quantitative syntactic typological analysis of Romance languages using information from corpora annotated for dependency syntactic relations (Liu and Xu, 2012). They have analyzed the overall distribution of dependency directions which enabled them to correlate with the degree of inflectional variation of a language and to classify them diachronically (compared to Latin) and synchronically. Moreover, in a different article (Alzetta et al., 2020), the authors presented a study whose main objective was to identify cross-linguistic quantitative trends in the distribution of syntactic relations in annotated corpora from distinct languages (4 Indo-European ones) 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. However, results were not correlated to scores of dependency parsing tools and corpora used were not parallel, thus the content part of texts was not a controlled variable.

Typological information has been used in different ways in many studies intending to improve dependency parsing results. It has been proved that typological comparison of languages is a powerful way of increase overall metrics concerning dependency parsing automatic annotation, especially regarding unannotated languages (which do not have any corpora annotated in terms of syntactic relations) and low-resource ones.

One example is the method proposed by Agić (Agić, 2017) where he combines three language comparison techniques to determine the best single source for an unannotated language: part-of-speech trigrams, a language identification software (lang.py tool, developed by (Litschko et al., 2020)), and WALS features. It considers the whole corpus of the unannotated language to determine the best (most similar in terms of the described comparison features) training corpus. Later, it has been showed by (Litschko et al., 2020) that better results are obtained when typologically analysing each sentence of the unannotated language in comparison with the available annotated corpora, defining, for each instance the best model (and not the same one for the whole text). In both studies, only qualitative typological features and surface level word order (part-of-speech trigrams) are considered.

While the studies mentioned in the previous paragraph focus on part-of-speech trigrams to compare languages, (Wang and Eisner, 2018) proposed a method to compare word order (in terms of part-of-speech possible combinations) by using a deep-learning algorithm (multilayer perceptron architecture) that classify languages in an unsupervised way with the information extracted from delexicalized corpora. This model is, then, used to the identification of the best language to serve as the source of the best training corpus for the target one. Their major aim was to prove that part-of-speech (POS) sequences carry useful information about the syntax of a language.

A different approach, using only typological information from URIEL database (lang2vec tool, (Littell et al., 2017)), was presented by (Glavaš and Vulić, 2021). Their method consists of comparing the vector composed by the values of the linguistic features of the target language with vectors from other well-resourced languages. The idea is not to select the best corpus but to combine the most similar ones from different languages as long as the similarity respects a determined threshold.

These studies have in common the objective of choosing the best combination of languages to improve dependency parsing results, there is no specific analysis concerning the influence of the chosen features used to compare languages on the final results.

In a different perspective, (de Lhoneux et al., 2018) compared how typological features are related to the dependency parsing results when twenty-seven diverse deep-learning parameters are used for cross-lingual parameters sharing. They were divided in three sets: character based one-layer (bidirectional LSTM), word based two-layer (bidirectional LSTM), and multi-layered perceptron (MLP) with a single layer. The authors have shown that the linguistic intuition that character- and word-level LSTMs are highly sensitive

to phonological and morphosyntactic differences (such as word order), whereas the MLP learns to predict less idiosyncratic, hierarchical relations from relatively abstract representations of parser configurations. Languages were compared in terms of their genealogical family and subject, verb and object order (qualitative classification).

3. Experimental Design

In this section, we describe the corpora that have been used in this study, the dependency parsing task evaluation using UDify software, and the typological classification method that has been employed.

3.1. Languages and Data-set Selection

The data-sets used for all experiments are part of the Parallel Universal Dependencies (PUD) tree-banks created for the CoNLL 2017 shared task on Multilingual Parsing from Raw Text to Universal Dependencies. They are composed of 1000 sentences for each language, always in the same order, coming from the news domain and Wikipedia (Zeman et al., 2017).

The first 750 sentences are originally in English and the rest are in German, French, Italian or Spanish. Sentences were mostly translated by professional translators via the English text. The data has been annotated morphologically and syntactically by Google according to Google universal annotation guidelines, afterwards, labels were converted to Universal Dependencies v2 guidelines[1].

The corpora were composed to serve as test sets to the mentioned shared task. Due to their relatively small size, the creators have suggested that a ten-fold cross-validation should be employed should these sets be used as training ones (as it is the case in this article). As our aim was to focus on one specific syntactic feature, the idea was to use only parallel corpora so that there would be no bias concerning the size or the domain of corpora. No data augmentation technique was used as there is no other parallel annotated corpora covering all PUD languages.

The list of PUD languages, their ISO-639-3 code and their genealogical information according to WALS[2] (Dryer and Haspelmath, 2013) is presented in the table 1. Although WALS database provides valuable information of word order patterns, there is no information regarding the relative position of the head and dependent in a broader way. Their focus is on word order position of subject, object and verb (and other type of syntactic functions), not exactly specifying the behavior of the ensemble of heads and dependents.

All corpora have been tagged in terms of core part-of-speech categories (UPOS) and dependency relation (deprel) using Universal Dependencies labels. The number of UPOS and deprel labels varies depending on the language, their distribution is presented in the tables 2 and 3.

Language	Code	Family	Genus
Arabic	arb	Afro-Asiatic	Semitic
Chinese	cmn	Sino-Tibetan	Chinese
Czech	ces	Indo-European	Slavic
English	eng	Indo-European	Germanic
Finnish	fin	Uralic	Finnic
French	fra	Indo-European	Romance
German	deu	Indo-European	Germanic
Hindi	hin	Indo-European	Indic
Icelandic	isl	Indo-European	Germanic
Indonesian	ind	Austronesian	Malayo-Sumbawn
Italian	ita	Indo-European	Romance
Japanese	jpn	Japanese	Japanese
Korean	kor	Korean	Korean
Polish	pol	Indo-European	Slavic
Portuguese	por	Indo-European	Romance
Russian	rus	Indo-European	Slavic
Spanish	spa	Indo-European	Romance
Swedish	swe	Indo-European	Germanic
Thai	tha	Tai-Kadai	Kam-Tai
Turkish	tur	Altaic	Turkic

Table 1: List of languages, the respective ISO-639-3 code and the genealogical information

Languages	Number of UPOS labels
kor	13
cmn, tur	15
arb, ces, fin, hin, jpn, spa, swe	16
eng, fra, deu, ita, pol, por, rus, tha	17
isl, ind	18

Table 2: Distribution of the number of UPOS labels (core part-of-speech) for each language in PUD data-set

The CoNLL-U format also presents a column for

[1] https://github.com/UniversalDependencies/UD_English-PUD

[2] https://wals.info/

Languages	Nb. of deprel types	Nb. of deprel sub-types
arb	34	8
cmn	32	12
ces	31	12
eng	36	12
fin	30	14
fra	31	14
deu	33	12
hin	28	10
isl	31	5
ind	33	14
ita	33	7
jpn	25	0
kor	26	8
pol	28	31
por	33	9
rus	31	8
spa	32	9
swe	33	9
tha	33	10
tur	34	7

Table 3: Distribution of the number of deprel labels for each language in PUD data-set

language-specific part-of-speech tag (XPOS). For this feature, not all languages follow the same labeling system. Arabic, Chinese, English, French, German, Hindi, Italian, Korean, Portuguese, Spanish, Turkish and Thai use the same tags to characterize language specific POS, the other languages in PUD either present different sets of tags, or, as it is the case of Finnish and Indonesian, no information at all is provided concerning this feature.

3.2. Dependency Parsing Annotations

UDify tool (Kondratyuk and Straka, 2019) proposes an architecture aimed for PoS-MSD and dependency parsing tagging integrating Multilingual BERT language model (104 languages). It can be fine-tuned using specific corpora (mono or multilingual) to enhance overall results.

We have selected this tool as it presents the state-of-the-art algorithms concerning the specific task of dependency parsing Annotation.

Training parameters were defined as:

- Number of epochs: 80
- Warmup: 500

Other parameters remained the same as proposed by the authors in their configuration file.

As previously mentioned, the size of the PUD corpora is relatively small (1.000 sentences), therefore, a 10-fold-cross-validation was employed. For each experiment, 600 sentences were used for training, 200 for validation and 200 for testing.

We have considered the LAS (labelled attachment score) value, which is the percentage of words that are assigned both the correct syntactic head and the correct dependency label, as the main dependency parsing metric metric.

Since UDify uses Multilingual BERT, and knowing that languages are not equally represented in this model, it is important to present the data distribution of the selected languages inside it (table 4), as it may have an impact on the final dependency parsing results.

Language	Size Range (GB)
eng	[11.314, 22.627]
deu, fra, spa, rus	[2.828, 5.657]
cmn, ita, jpn, pol, por	[1.414, 2.828]
arb, ces, swe	[0.707, 1.414]
fin, ind, kor, tur	[0.354, 0.707]
tha	[0.177, 0.354]
hin	[0.088, 0.177]
isl	[0.022, 0.044]

Table 4: List of languages we consider in mBERT and its pre-training corpus size (Wu and Dredze, 2020)

It is possible to notice that there is a huge discrepancy regarding the amount of data from different languages used to generate multilingual BERT language model.

As expected, English is the language which has the largest pre-training corpus size, followed by German, French, Spanish and Russian. It is possible to observe that the largest mBERT pre-training corpora come from Indo-European languages, only Chinese and Japanese languages are also quite well represented. Icelandic is the one with the smaller pre-training corpus, therefore, not as well represented in this language model as the other languages from PUD corpora.

Thus, even though we use parallel data to understand the influence of the position of head and dependent feature, by using a system based on mBERT introduce a bias regarding the discrepancy of the training data used in this language model. The importance of this bias will be analysed further in this article. We could have chosen a tool which does not depend on language models to conduct our experiments, however, as these models are part of the state-of-the-art concerning dependency parsing, we decided to keep our initial choice to verify how the chosen syntactic feature influences the results of parsing, if it plays an important role or if it is completely minimized.

3.3. Syntactic Typological Characterisation

To analyse the dependency parsing results obtained from different languages using parallel corpora, we propose a quantitative typological approach concerning syntax, more specifically the head directionality parameter, whether the head precedes the dependent (right-

branching) or is after it (left-branching) in the sentence (Fábregas et al., 2015). The extraction of parameters reflect the directionality observed at the surface level (position of head and dependent observed at the sentence level).
The corpora being parallel, therefore containing the same semantic information, allows us to focus on the syntax differences among the selected languages.
Using a python script, we have extracted for each language the existing patterns concerning the relative position in the sentence of the heads and the dependents, as well as the frequency of occurrence of each pattern. All observed patterns concerning the relative position of head and in PUD corpora have been considered.
All observed patterns extracted from the PUD corpora (2,890 in total) have been included in the language vectors. Our aim is to verify the relevance of this quantitative method to predict LAS results.
An example of extracted pattern is the following:

- ADV_aux_precedes_ADJ - head-final or left-branching - It means that the dependent, which is an adverb (ADV) precedes the head which is an adjective (ADJ) and has the syntactic function of an auxiliary (aux). The dependent can be in any position of the sentence previous to the head, not necessarily right before.

- CCONJ_cc_follows_NOUN - head-initial or right-branching - In this case, the dependent, a coordinating conjunction (CCONJ), comes after the head, which is a noun (NOUN), and has the function of coordination (cc). The dependent can be in any position after the head, not necessarily being right next to it.

Therefore, for each language, we have obtained a vector containing all the existing patterns and their frequency. The distances between languages were calculated using R dist() function (Euclidean) and from the obtained distance matrix, we generated a plot with language clusters using R hclust() function, which uses the complete linkage method for hierarchical clustering by default. This particular clustering method defines the cluster distance between two clusters to be the maximum distance between their individual components.

4. Results

In this section, first, we present the LAS results obtained using UDify trained with the different parallel corpora from PUD data-set. Then, we display the results of the typological analysis (clusters in the format of a dendogram).

4.1. Dependency Parsing Results using UDify

Using UDify with 10-fold cross-validation, we were able to obtain LAS results for all PUD languages. LAS scores and the respective standard deviation values are presented in the table 5.

Language	LAS	Std. Dev.
cmn	72.98	2.08
tur	75.34	2.11
hin	76,12	1.12
isl	77.80	2.56
fin	81.15	1.88
arb	82.37	0.70
kor	84.55	1.33
swe	85.13	1.53
ind	85.51	1.26
pol	86.08	1.59
ces	86.34	1.00
eng	87.39	1.28
deu	88.22	0.85
rus	88.22	0.97
por	88.88	0.85
ita	89.74	0.86
spa	90.23	1.20
jpn	90.75	2.11
fra	90.84	1.36

Table 5: LAS and standard deviation results obtained for each language of PUD data-set using UDify and 10-fold cross-validation. Results are presented from lowest to highest LAS score.

Even though parallel corpora were used to train UDify tool, LAS results vary considerably among PUD languages. The lowest LAS score was obtained for Chinese language (72.98) and the highest for French language (90.82), difference of 15.38 points which is much higher than the calculated standard deviation values.
LAS results are higher than 85 for 11 out of the 16 PUD languages considered in this part of the study, which can be considered as relatively satisfying scores considering the small size of the training corpora.
Analysing Indo-European languages, Romance languages tend to have better LAS scores (higher than 90 for French and Spanish), followed by Germanic and Slavic languages, the exception being Icelandic which has the second lowest LAS value (78.12) among the considered languages.
Indonesian and Korean have scores comparable to other Indo-European languages such as Swedish, Polish and Czech (around 85). Finnish and Arabic have lower scores than Indo-European languages but higher than 80 and, therefore, better than Icelandic and Turkish languages.
When we analyse these LAS results together with the training size of mBERT (mean value of the size range), it is possible to calculate the following correlation coefficients:

- Pearson's correlation = 0.37

- Spearman's correlation: 0.73

Thus, it seems that these two variables are strongly correlated following a non-linear monotonic function (as it is attested by the value obtained for Spearman's coefficient).

4.2. Quantitative Syntactic Language Classification

As explained previously in this article, languages were compared and classified considering quantitative information of the patterns of the position of heads and dependents. The figure 1 presents the clusters of languages generated using R' hclust function.

It is possible to observe in this dendogram the main central cluster corresponding to most of Indo-European languages (except for Hindi). Romance languages are grouped in a sub-cluster of the Indo-European one. We can also notice the proximity of English and Swedish (both Germanic) and Russian and Czech (both Slavic). Icelandic, although being a Germanic language, is closer to Polish language when this specific syntactic feature is analysed. Icelandic is presented in the dendogram grouped with the other Slavic languages. German, also a Germanic language, is grouped closer to the Romance group (specially with Italian and French) and not with the other Germanic languages from PUD corpora.

Close to the Germanic/Slavic cluster, it is possible to notice the group containing Thai, Arabic and Indonesian which have no genealogical relation. The two extremes groups are composed, on the left, by Hindi and Japanese, and, on the left by two sub-clusters: Finnish and Turkish (which is expected as similarities between these languages have been previously observed) and Chinese and Korean.

Beside the classification presented in the figure 1, with the syntactic information extracted for each language, it is also possible to analyse the overall tendency of left-branching or right-branching. The table 6 presents the percentage of cases inside each language corpus where the dependent comes before the head in the sentence (left-branching).

With the results presented in the table 6 and in the Figure 2, it is possible to check whether PUD languages are head-initial or head-final. Arabic, Thai and Indonesian are head-initial languages, Japanese also tends towards being head-initial. Oppositely, Turkish and Korean are distinctly head-final languages. Chinese, Romance and Germanic languages, except Icelandic, have a tendency of being head-final (percentage superior to 55 in the table 6). Slavic languages present different patterns, Polish does not present any tendency, Russian and Czech tend to be head-final because of more relaxed word order in Slavic languages.

The correlation coefficients (Spearman's and Pearson's) were calculated using the percentage of heads preceeding dependents and the delta concerning a balanced distribution of directionality (50/50). The obtained results are lower than 0.1, thus, no correlation can be stated concerning this feature.

Language	%
arb	36.33
tha	39.05
ind	41.91
jpn	45.85
pol	49.21
isl	51.08
rus	54.50
fin	55.85
hin	56.10
ita	57.09
ces	57.13
spa	57.79
por	57.94
fra	58.28
swe	58.75
cmn	60.06
eng	63.77
deu	66.81
tur	69.96
kor	79.86

Table 6: Total percentage of occurrences where the dependent precedes the head (left-branching / head-final) in each selected language corpus

5. Discussion

Comparing the results obtained in our campaign to the scores presented by the developers of UDify (Kondratyuk and Straka, 2019), it is possible to notice that, in general, our LAS values for PUD corpora are higher. It may be due to the fact of using different strategies for using PUD as training set. Also, as expected, LAS scores using PUD are not as high as compared to results from other models trained with larger corpora.

We observed that, as expected, the size of the corpus used to train mBERT has a strong positive correlation with the LAS scores, however, it does not explain the ensemble of the results as English has the biggest training corpus but do not provide the best score concerning UDify.

Analysing Indo-European languages results, it is possible to see that, overall, Romance languages are the ones with the highest LAS values. In terms of multilingual BERT, all of them have large pre-training mBERT corpora. French and Spanish have larger pre-training corpora when compared to Portuguese and Italian and UDify performs better for these two languages. Romance languages are grouped in the figure 1, showing similar distribution of patterns concerning head and dependents position when compared to other PUD languages.

English language, which is the one with the largest pre-training mBERT corpus, does not have the highest LAS

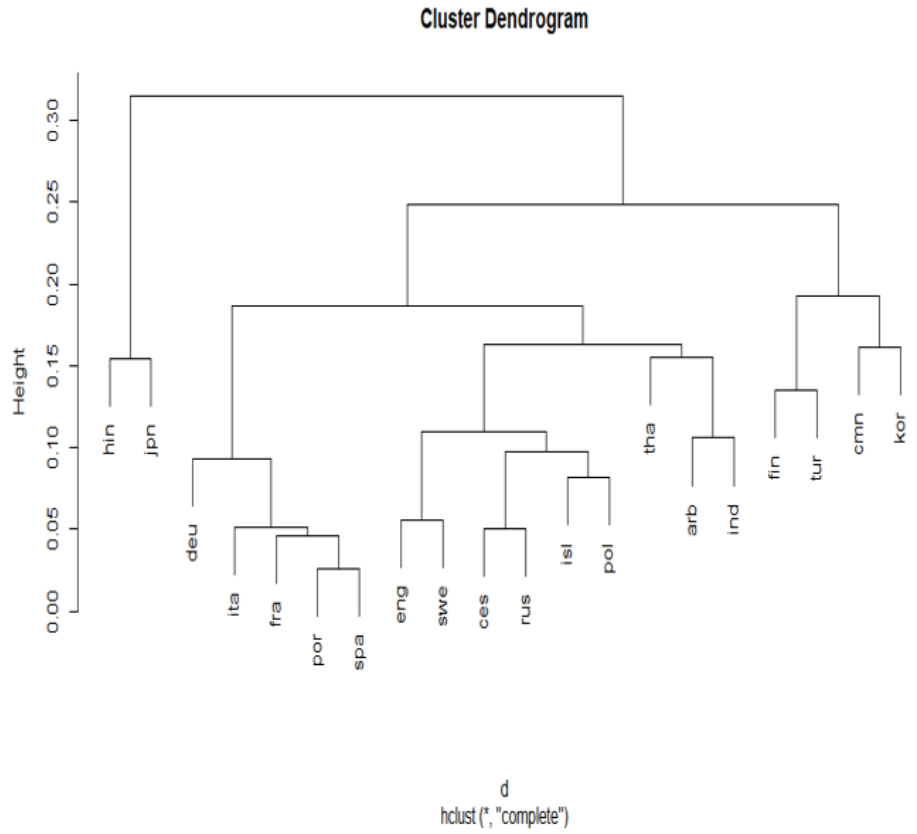

Figure 1: PUD language clusters generated using quantitative analysis of head and dependent position features

score. Its result can be compared to other languages with much smaller mBERT corpus such as Russian, German, Czech and Polish. Thus, it seems that size of the representation of a language in mBERT may play a role only to a certain point when it is used for fine-tuning in parsing tools.

When we observe, more precisely, Germanic and Slavic languages, it is possible to notice that although English and Swedish languages form a sub-cluster, their LAS scores are slightly different. In this case, it may be caused by the discrepancy of the representation of the languages in mBERT. It is the same when we consider Russian and Czech languages.

It is also interesting to observe the sub-cluster formed by Polish and Icelandic inside the group of Indo-European languages. Polish language has a mBERT representation size comparable to Portuguese and Italian, however, its LAS value is much lower. It may be due to its specific syntactic structures as well as to elevated number of deprel labels (59) which is much higher than all the other PUD languages. Icelandic has the second lowest LAS score among PUD languages. Although being a Germanic language, it is not similar in its syntactic structure of heads and dependents when compared to English, Swedish nor German. In addition, Icelandic has the smallest mBERT pre-training corpus which has probably strongly contributed to the low LAS value obtained using UDify.

On the left of the main cluster of Indo-European languages in the Figure 1, we have the sub-cluster formed by Arabic and Indonesian. Both languages have lower LAS scores when compared to Indo-European ones, Indonesian having a better performance even though its mBERT representation is smaller and its number of deprel is higher (47 for Indonesian and 42 for Arabic).

Considering the cluster on the right side of the figure 1, formed by Finnish, Turkish, Chinese and Korean, these languages tend to present lower LAS scores. Finnish, Turkish and Korean have similar size of mBERT representations, but smaller than Indo-European languages. However, their size is comparable to Indonesian which presents better LAS value and, in the figure 1, this language is clustered closer to the Indo-European group. As seen in the table 4, Turkish is a head-final language, it may influence the results. However, Korean language is even more head-final when compared to Turkish and has a better LAS score, however, Korean presents only 34 deprel, while Turkish has 41.

In light of these results, it is possible to notice that differences in the syntactic structures concerning head and dependents may play a role in dependency parsing tools overall results. The size of the language representation

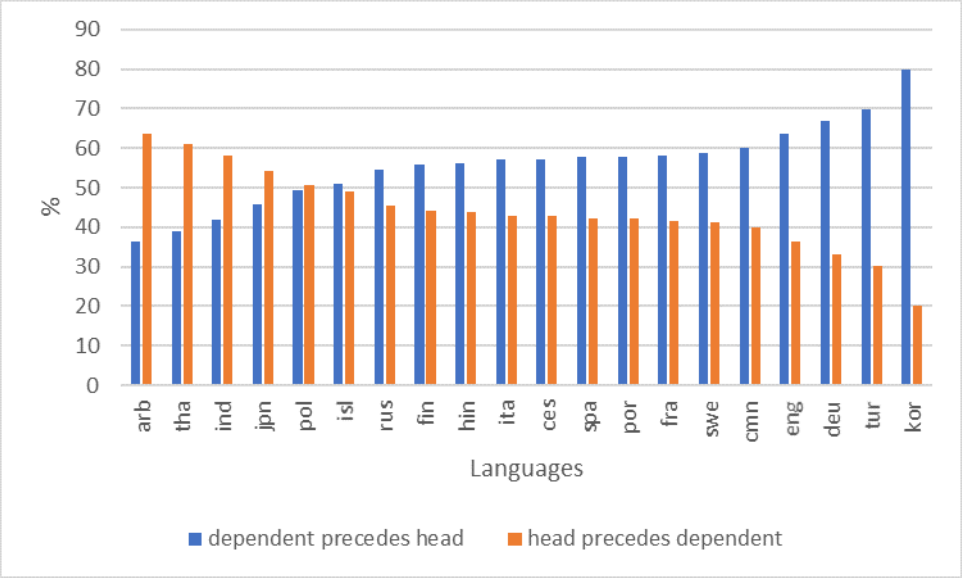

Figure 2: Distribution of the directionality of head and dependent in PUD language corpora

in the language model, however, plays a major role as it helps improving final scores (as it is the case for French and Spanish). Nevertheless, it may not be sufficient to guarantee satisfying LAS result (as it is the case for Turkish).

Also, the parameter of head and dependent position is not the only syntactic feature playing an important role in the observed LAS results, the complexity of the language, represented by the number of deprel labels should be considered as it may have caused the lower LAS value for Polish language (which has a high number of deprel subtypes).

Furthermore, it is important to mention that morphological aspects (whether the language has synthetic or analytical morphology), can also influence the efficacy of the parser. As it can be observed for Finnish and Turkish (both synthetic languages), LAS results are low. However, it is also the case for Chinese, which is an analytical language.

6. Conclusions and Perspectives

In this article we have presented a detailed analysis of dependency parsing results obtained for 16 languages using parallel corpora to understand the differences in the obtained scores considering the specific syntactic feature of head directionality parameter with which we have conducted a quantitative syntactic typological classification.

Thus, we have conducted a series of experiments using UDify tool, using a 10-fold cross-validation to obtain LAS metric for the selected languages. In parallel, we have extracted patterns concerning the position of head and dependents (left or right branching) to generate vectors which were used to compare and classify languages into clusters.

We have observed that, even though parallel corpora were used, different languages present considerably different LAS results. Indo-European languages tend to present better LAS scores, inside this group, Romance languages are the ones that performed the best.

UDify tool uses multilingual mBERT and as the sizes of each language inside this language model are not homogeneous, it was possible to notice that this discrepancy plays a major role in the LAS scores. As expected, languages with larger mBERT representation tend to perform better.

However, the size of the language in mBERT is not the only parameter playing a role in the overall results. English has, by far, the largest size in mBERT and still has a lower LAS score when compared to Romance languages which were all classified the same cluster in our typological study. It is also the case for Russian, which has a mBERT size comparable to French and Spanish for which LAS values are comparable to Portuguese and Italian with smaller mBERT size.

In addition to that, Arabic language has a mBERT representation comparable to Czech and Swedish but its LAS results are not as good as these two languages. Arabic language forms a sub-cluster with Indonesian, not as close to other languages with better performance as it is the case for Czech and Swedish. Also, it is possible to conclude that the size of the language in mBERT and the head and dependent position are not the only aspects influencing the results. Polish language is an example of that, and the reason for the lower LAS value obtained for this language may be the higher number of dependency parsing labels specific of this language.

For future research, it would be interesting to observe how these languages perform in systems which either use more homogeneous language models in terms of language representation or that do not depend on lan-

guage models at all. Furthermore, specific analysis could be done considering only subject-verb or object-verb directionality.

7. 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).

8. Bibliographical References

Agić, Ž. (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, May. Association for Computational Linguistics.

Alves, D., Thakkar, G., and Tadić, M. (2020). Evaluating language tools for fifteen EU-official under-resourced languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1866–1873, Marseille, France, May. European Language Resources Association.

Alzetta, C., Dell'Orletta, F., Montemagni, S., Osenova, P., Simov, K., and Venturi, G. (2020). Quantitative linguistic investigations across universal dependencies treebanks. In Johanna Monti, et al., 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.

Boroș, T., Dumitrescu, S. D., and Burtica, R. (2018). NLP-cube: End-to-end raw text processing with neural networks. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 171–179, Brussels, Belgium, October. Association for Computational Linguistics.

de Lhoneux, M., Bjerva, J., Augenstein, I., and Søgaard, A. (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.

Dell'Orletta, F., Venturi, G., and Montemagni, S. (2013). Linguistically–driven selection of correct arcs for dependency parsing. *Computación y Sistemas*, 17.

Matthew S. Dryer et al., editors. (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Fábregas, A., Mateu, J., and Putnam, M. T. (2015). *Contemporary Linguistic Parameters: Contemporary Studies in Linguistics*. Bloomsbury Academic, London.

Glavaš, G. and Vulić, I. (2021). Climbing the tower of treebanks: Improving low-resource dependency parsing via hierarchical source selection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4878–4888, Online, August. Association for Computational Linguistics.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Jurafsky, D. and Martin, J. H. (2009). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J.

Jurafsky, D. and Martin, J. H. (2021). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd. (draft) edition.

Kaalep, H.-J. and Veskis, K. (2007). Comparing parallel corpora and evaluating their quality. In *Proceedings of Machine Translation Summit XI: Papers*, Copenhagen, Denmark, September 10-14.

Kondratyuk, D. and Straka, M. (2019). 75 languages, 1 model: Parsing universal dependencies universally.

Kuhn, J. (2005). Parsing word-aligned parallel corpora in a grammar induction context. In Philipp Koehn, et al., editors, *Proceedings of the Workshop on Building and Using Parallel Texts@ACL 2005, Ann Arbor, Michigan, USA, June 29-30, 2005*, pages 17–24. Association for Computational Linguistics.

Litschko, R., Vulić, I., Agić, Ž., and Glavaš, G. (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), December. International Committee on Computational Linguistics.

Littell, P., Mortensen, D. R., Lin, K., Kairis, K., Turner, C., and Levin, L. (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.

Liu, H. and Xu, C. (2012). Quantitative typological analysis of romance languages. *Poznań Studies in Contemporary Linguistics*, 48(4):597–625.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, June. Association for Computational Linguistics.

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 Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Otter, D. W., Medina, J. R., and Kalita, J. K. (2018). A survey of the usages of deep learning in natural language processing. *CoRR*, abs/1807.10854.

Pires, T., Schlinger, E., and Garrette, D. (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.

Straka, M. and Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

Teubert, W. (1996). Comparable or Parallel Corpora? *International Journal of Lexicography*, 9(3):238–264, 09.

Wang, D. and Eisner, J. (2018). Surface statistics of an unknown language indicate how to parse it. *Transactions of the Association for Computational Linguistics*, 6:667–685.

Wu, S. and Dredze, M. (2020). Are all languages created equal in multilingual bert?

Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinková, S., Hajič jr., J., Hlaváčová, J., Kettnerová, V., Urešová, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C. D., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.-C., Sanguinetti, M., Simi, M., Kanayama, H., de Paiva, V., Droganova, K., Martínez Alonso, H., Çöltekin, Ç., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Alcalde, H. F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonça, G., Lando, T., Nitisaroj, R., and Li, J. (2017). CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada, August. Association for Computational Linguistics.