Corpus-based Syntactic Typological Methods for Dependency Parsing Improvement

Diego Alves
Faculty of Humanities and
Social Sciences - University of Zagreb
dfvalio@ffzg.hr

Božo Bekavac
Faculty of Humanities and
Social Sciences - University of Zagreb
bbekavac@ffzg.hr

Daniel Zeman
Faculty of Mathematics and Physics
Charles University
zeman@ufal.mff.cuni.cz

Marko Tadić
Faculty of Humanities and
Social Sciences - University of Zagreb
marko.tadic@ffzg.hr

Abstract

This article presents a comparative analysis of four different syntactic typological approaches applied to 20 different languages to determine the most effective one to be used for the improvement of dependency parsing results via corpora combination. We evaluated these strategies by calculating the correlation between the language distances and the empirical LAS results obtained when languages were combined in pairs. From the results, it was possible to observe that the best method is based on the extraction of word order patterns which happen inside subtrees of the syntactic structure of the sentences.

1 Introduction

Dependency parsing is a Natural Processing Processing (NLP) task that concerns the process of determining the grammatical structure of a sentence by examining the syntactic relations between its linguistic units. In other words, it consists of the identification of heads and dependents as well as the type of relationship between them (Jurafsky and Martin, 2009).

From 2015 onward, the usage of deep learning techniques has been dominant in studies regarding the dependency parsing task. Although it has provided a great improvement in overall results even for under-resourced languages (Otter et al., 2018), it requires a large amount of annotated data which can be problematic, particularly in terms of cost (Guillaume et al., 2016).

To overcome the problem of lack of data, cross-lingual parsing strategies using typological methods have been proposed to determine which languages can be combined for effective improvement of dependency parsing results (Ponti et al., 2019b). Most of these studies rely on the usage of information provided by typological databases such as WALS (Dryer and Haspelmath, 2013) sometimes combined with n-grams analysis extracted from corpora. On the other hand, the usage of corpus-based typology for this aim is still incipient.

Moreover, most studies focus on the obtained improvement, without analyzing the existence of a proper correlation between the typological features involved in the process with the overall synergy regarding the impact on the dependency parsing results.

Therefore, our aim in this paper is to propose an examination of several corpus-based typological methods in terms of correlation between language distances and dependency parsing scores. The paper is composed as follows: Section 2 presents an overview of the related work to this topic. In Section 3, we describe the campaign design: language and data-sets selection, corpus-based typological characterization, dependency parsing experiments, and correlation measures; Section 4 presents 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 WALS database is one of the most used typological resources in NLP studies (Ponti et al., 2019a). It contains phylogenetic, phonological, morphosyntactic, and lexical information for a large number of languages that can be used for a large variety of linguistic studies (Dryer and Haspelmath, 2013). Along with that, the URIEL Typological Compendium was conceived as a meta-repository that is composed of numerous databases (WALS included) and is the base of the lang2vec tool (Littell et al., 2017). This tool is a powerful resource that allows languages to be characterized as vectors composed of typological features associated with specific values. Users can choose the type of features (i.e.: genealogical, phonological, syntactic, etc) according to their precise needs. While proposing an effective way to compare languages typologically, this tool does not characterize all languages homogeneously as it depends on the availability of linguistic descriptions provided by its sources. Thus, low-resourced languages usually have less information. For example, it is not possible to compare all 24 European Union languages as there are no common features with valid values for all of them. Furthermore, typological databases usually fail to illustrate the variations that can occur within a single language (i.e.: in general, only the most frequent phenomena are reported in the literature, and not all attested ones).

In terms of corpus-based typological studies, a broad survey was provided by Levshina (2022). The author showed that while several authors quantitively analyzed specific word-order patterns (e.g.: subject, verb, and object position (Östling, 2015), and verb and locative phrases (Wälchli, 2009)), other researchers have focused on quantitative analyses regarding language complexity (e.g.: (Hawkins, 2003) and (Sinnemäki, 2014)). On the other hand, the concept of Typometrics was introduced by Gerdes et al. (2021). The focus of their research was to extract rich details from corpora for testing typological implicational universals and explored new kinds of universals, named quantitative ones. Thus, different word-order phenomena were analyzed quantitatively (i.e.: the distribution of their occurrences in annotated corpora) to identify the ones present in all or most languages.

Thus, it is possible to notice that most studies regarding quantitative typology focus either on the analysis of specific linguistic phenomena or on the identification of universals. Our approach differs from theirs as our aim is to compare languages (i.e.: language vectors) using quantitative information concerning all syntactic structures extracted from corpora to obtain a more general syntactic overview of the elements in our language set and use the results as strategies to improve dependency parsing results.

An interesting method concerning the extraction and comparison of syntactic information from tree-banks was developed by Blache et al. (2016a). The MarsaGram tool is a resource that allows syntactic information (together with its statistics) to be extracted from annotated corpora by inferring context-free grammars from the syntactic structures. MarsaGram allows the extraction of linear patterns (i.e.: if a specific part-of-speech precedes another one inside the same subtree ruled by a determined head). The authors conducted a cluster analysis comparing 10 different languages and showed the potential in terms of typological analysis of this resource. However, the results were only compared to the genealogical classification of the selected languages and did not provide any comparison to other corpus-based methods. Moreover, the authors did not use the obtained classification with the perspective of improvement of dependency parsing systems via corpora-combination.

One example of effective usage of typological features (from URIEL database) to improve results of NLP methods was presented by Üstün et al. (2020). The authors developed the UDapter tool that uses a mix of automatically curated and predicted typological features as direct input to a neural parser. The results showed that this method allows the improvement of the dependency parsing accuracy for low-resourced languages. A similar study, using a different deep-learning architecture was conducted by Ammar et al. (2016), however, in both cases, there is no detailed analysis of which features were the most relevant.

Furthermore, Lynn et al. (2014) proposed a study concerning the Irish language using delexicalized corpora. The authors performed a series of cross-lingual direct transfer parsing for the Irish language and the best results were achieved with a model trained with the Indonesian corpus, a language from the Austronesian language family. The authors proposed some analysis considering similarities between the treebanks of both languages in terms of dependency parsing labels, however, a detailed statistical analysis of corpora and a complete comparison of specific typological features were not carried out.

While some papers focus on genealogical features, others consider syntactic ones. For example, Alzetta et al. (2020) presented a study whose aim 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) which detects patterns of syntactic structures in tree-banks. However, only four Indo-European languages were scrutinized but some interesting insights concerning language peculiarities were observed.

Thus, studies regarding corpus-based typology and dependency parsing are usually presented without a specific comparison to other existing approaches or to the classic one concerning typological databases. That is why in this article the idea is to analyse possible quantitative typological methods in terms of correlation with the improvement obtained regarding dependency parsing results when corpora from different languages are combined.

3 Campaign Design

In this section, a brief overview of the selected data-sets is provided, followed by a description of selected the corpus-based syntactic typological approaches. Moreover, we detail the dependency parsing experiments and the correlation measures that were chosen for the analysis of the results.

3.1 Parallel Corpora

For the ensemble of experiments presented in this paper, we decided to use the Parallel Universal Dependencies (PUD) compilation that was created for the CoNLL 2017 shared task on Multilingual Parsing from Raw Text to Universal Dependencies (Zeman et al., 2018).

Levshina (2022) showed the benefit of using parallel corpora in typological studies, as the bias regarding size and content is avoided. Especially in this case, the usage of parallel sentences allows the focus to be on the syntactic strategies that are used by each language to express the same meaning.

The PUD collection provides 1,000 parallel sentences from news sources and Wikipedia annotated following Universal Dependencies guidelines (De Marneffe et al., 2021) in the CoNLL-U format for twenty languages[1]: Arabic, Chinese, Czech, English, Finnish, French, German, Hindi, Icelandic, Indonesian, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish, Thai, and Turkish. The PUD corpora are composed of translations from English (750 sentences), German (100), French (50), Spanish (50), and Italian (50). Although avoiding some biases linked to size and genre, these data-sets may contain some "translationese" ones, phenomena described by Volansky et al. (2015). Dependency parsing annotations were done automatically and, then, verified manually.

The list of PUD languages together with their ISO 639-3 codes and their genealogical information[2] is provided in Table 1. Although the total number of languages is limited to 20, the PUD collection provides, at least, some variety in terms of genealogy (i.e.: most languages belong to the Indo-European family, but 8 other different linguistic families are also present in this data-set).

The PUD Collection used in this article corresponds to the one available in the Universal Dependencies[3] data-set v.2.7 (November 2020).

3.2 Corpus-based Typological Approaches

Four different quantitative approaches were selected:

- MarsaGram all properties
- MarsaGram linear properties
- Head and dependent relative order
- Verb and object relative order

Each method is fully described in the subsections below. In the results section, these strategies are compared to the typological classification obtained with lang2vec tool (Littell et al., 2017): PUD languages are represented as language vectors composed of 41 syntactic features with valid values (i.e.: 0.0, 0.33, 0.66, and 1.0). The total number of syntactic features in this tool is 103, but only 41 are common to all PUD languages.

For each typological method, first, we generated the language vectors by extracting the syntactic information from the data-sets. Then, dissimilarity matrices were calculated using Euclidean and cosine distances (using R scripts). Thus, for each strategy, two matrices were obtained. The distance information between the languages is one of the inputs for the correlation analysis.

3.2.1 MarsaGram all properties

MarsaGram is a tool for exploring treebanks, it extracts context-free grammars (CFG) from annotated data-sets that can be used for statistical comparison between languages as proposed by Blache et al. (2016b). We have used the latest release of this software downloaded from the ORTOLANG platform of linguistic tools and resources[4].

This software identifies four types of properties from the corpora:

[1] Originally it was composed of fewer languages. Polish and Icelandic were added after the shared task, for example.

[2] Although the existence of the Altaic family has been challenged by some experts as detailed by Norman (2009), WALS database consider it in its genealogical classification.

[3] https://universaldependencies.org/

[4] https://www.ortolang.fr/market/tools/ortolang-000917

Language	ISO 639-3	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-Sumbawan
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 inside PUD collection, their respective ISO 639-3 three-character code, and their genealogical information according to WALS.

- Precede or Linear: It describes the relative position of two elements (A precedes B) inside a subtree governed by a specific head. Each element is described by its part-of-speech (POS) and dependency relation (deprel) in the syntactic tree. Although being part of the same subtree, elements A and B are not necessarily syntactically linked. An example of a sentence with this property is presented in the Annex section (Figure 1).

- Require: This property describes the cases where the presence of an element A requires the existence of an element B inside the subtree. An example of a sentence with this property is presented in the Annex section (Figure 2).

- Unicity: an element A has this property if inside the subtree it occurs only once (i.e.: no other element with the same part-of-speech and dependency label is attested). In the Annex section, one example of a sentence with this property is presented (Figure 3).

- Exclude: In this case, the presence of element A excludes the occurrence of element B inside the subtree.

Property	Number of Patterns	%
Linear	21,242	13.38
Require	6,189	3.90
Unicity	2,144	1.35
Exclude	129,180	81.37

Table 2: Distribution of extracted features using MarsaGram in terms of properties.

Of the four properties described above, only the linear one is directly linked to word-order patterns on the surface level of the sentence. In total 158,755 patterns were extracted from the PUD corpora. The distribution in terms of types of property is presented in table 2.

Each language vector regarding the MarsaGram all properties strategy is composed of these features associated with the value corresponding to its frequency of occurrence inside the corpus.

3.2.2 MarsaGram linear properties

As previously explained, the patterns with the linear property extracted with the MarsaGram tool are the ones that correspond to word-order phenomena inside subtrees. Thus, it seems pertinent to analyze them separately from the patterns regarding other properties, especially because when all phenomena are considered, the large majority correspond to the "exclude" property as presented in Table 2.

Thus, by extracting just linear patterns from PUD corpora, we generated language vectors composed of 21,242 features.

3.2.3 Head and dependent relative order

Besides the typological analysis provided from the data extracted using the MarsaGram tool, we also propose a quantitative approach concerning syntax, more specifically the head directionality parameter (i.e.: whether the heads precede the dependents (right-branching) or follow them (left-branching) in the surface-level of the sentence (Fábregas et al., 2015).

Hence, the attested head and dependent relative position patterns (and their frequency) in the different PUD corpora were extracted using a Python script. All observed features extracted from the PUD corpora (2,890 in total) have been included in the language vectors. From this total, 1,374 features (47.5%) correspond to cases where the dependent precedes the head, and 1,516 (52.5%) to right-branching patterns. In the cases where a feature was not observed in a determined language, the value 0 was attributed to it.

Two examples of head and dependent relative position patterns are presented below:

- ADV_advmod_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 adverbial modifier (advmod). The dependent can be in any position of the sentence previous to the head, not necessarily right before. An example of a sentence with this pattern is presented in the Appendix section (Figure 4).

- NOUN_obl_follows_VERB - head-initial or right-branching - In this case, the dependent (NOUN), comes after the head, which is a verb, and has the function of oblique nominal (obl). The dependent can be in any position after the head, not necessarily being right next to it. An example of a sentence representing this pattern is presented in the Appendix (Figure 5).

This specific analysis of the head and dependent relative position corresponds to a quantitative interpretation of the Head and Dependent theory (Hawkins, 1983) which considers that there is a tendency of organizing head and dependents in homogeneous word ordering. This author proposed a set of language types according to attested word-order phenomena concerning a limited list of elements as heads and dependents. In this article, we decided to consider all possible head and dependent pairs to conduct our analysis to have a more global overview of these ordering phenomena.

3.2.4 Verb and object relative order

Inside the ensemble of features extracted for the analysis of the head and dependent relative position, it is possible to extract the ones regarding verbs and direct objects (deprel: "obj") for a specific analysis of these phenomena. We decided to examine the position of these two elements in detail as they are key in typological studies such as the one proposed by Dryer (1992) where correlations are defined according to whether the verb comes before or after the object.

Thus, to compose the language vectors we extracted the head and dependent patterns which concern verbs and objects only (not only nominal but all other possible ones). We have decided to consider all the direct objects as if only nominal ones were analysed, the obtained classification would be similar to the general one available in databases (VO or OV languages), thus, not allowing us to differentiate in detail all PUD languages. In total, 13 OV and 12 VO features were attested in the PUD collection, allowing us to generate a 25-dimension language vector for each language.

3.3 Dependency parsing experiments

For the ensemble of experiments regarding dependency parsing, we used the UDify tool (Kondratyuk and Straka, 2019) which proposes an architecture aimed at PoS-MSD and dependency parsing tagging of tokenized texts integrating Multilingual BERT language model (104 languages) (Pires et al., 2019). It can be fine-tuned using specific corpora (mono or multilingual) to enhance overall results. This tool was selected as it presents 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. To calculate the statistical significance of the results, for each training corpus, we conducted 6 experiments with different values of random seeds, allowing us to calculate the mean value of the labeled attachment score (LAS) and its standard deviation.

The baseline regarding dependency parsing results consists of LAS values obtained with monolingual-trained models of PUD languages. For each experiment, 600 sentences were used for training, 200 for validation, and 200 for testing. Regarding the multilingual experiments, we combined PUD languages in pairs (concatenation of the training corpora). Thus, a total of 380 models were trained. Validation and test sets were the same ones as those used for the baseline experiments (monolingual ones).

With the baseline scores and the results obtained with the multilingual language pairs, we were able to calculate deltas to quantify the existing synergy between languages when corpora are combined for dependency parsing improvement. The deltas were obtained with:

$$Delta = LAS_{lang\_1\_and\_2} - LAS_{lang\_1} \quad (1)$$

The deltas were considered statistically significant if the p-value calculated between the two LAS scores was lower than 0.01.

3.4 Correlation calculation

The main focus of this study is to check whether the language distances obtained from the corpus-based typological approaches correlate with the LAS deltas (i.e., with the synergy between the languages when combined in dependency parsing experiments with deep-learning tools).

Two different correlation coefficients were chosen as they represent different ways that variables can correlate: Pearson's and Spearman's. The first one corresponds to the measure of linear correlation between two variables (Pearson, 1895), while the second determines how well the relationship between two variables can be defined as a monotonic function (Lehman, 2005).

Correlation values vary from -1 to 1. In our case, we expect negative values as we hypothesize that languages distances and deltas are inversely correlated (i.e.: the higher the distance between the languages, the lower will be the delta).

Language	LAS	Std. Dev.
tha	74.68	0.13
cmn	74.84	0.56
tur	76.68	0.21
hin	77.46	0.35
isl	78.90	0.16
fin	82.46	0.28
arb	83.34	0.24
swe	84.69	0.26
ind	85.72	0.19
kor	85.99	0.20
eng	86.63	0.15
ces	86.80	0.40
pol	86.88	0.21
rus	88.42	0.15
ita	89.48	0.14
deu	89.55	0.17
por	89.65	0.16
fra	91.20	0.21
spa	91.24	0.09
jpn	91.57	0.20

Table 3: LAS results obtained using UDify tool and PUD corpora using monolingual models.

4 Results

In the following subsections, we present the baseline results regarding the dependency parsing experiments together with an overview of the LAS values obtained when languages were associated. Then, the correlation analyses are displayed.

4.1 Dependency parsing baseline

As previously explained, the baseline consists of the LAS values obtained when monolingual training corpora were used to train the models using UDify tool. The PUD corpora were divided into train, development, and test sets (with 600, 200, and 200 sentences respectively). For each dataset, we conducted 6 experiments varying the random seed value for the calculation of the standard deviation and p-values. The results are presented in Table 3.

It is possible to notice that LAS results vary from 74.68 (for the Thai language) to 91.57 (for Japanese), almost 17 points of difference. Moreover, besides Japanese, all Romance languages also have rather high scores. The German language appears in between the ones of the Romance group, while other Germanic languages have lower scores (below Slavic languages). English and Swedish

Language	Positive deltas	Negative deltas
hin	0	0
jpn	0	6
kor	0	14
ind	1	1
tha	1	6
arb	2	0
fra	3	0
cmn	4	0
tur	4	1
deu	6	0
pol	9	0
ita	10	0
por	11	0
spa	11	0
ces	12	0
eng	14	0
isl	14	0
swe	14	0
rus	15	0
fin	16	0

Table 4: Number of positive and negative deltas concerning the LAS scores of the language combination experiments with the UDify tool (p-value < 0.01).

have quite similar results, however, Icelandic is positioned with the languages with the lowest scores (below 80) which are: Thai, Chinese, Turkish, and Hindi.

It has been shown by Alves et al. (2022) that these results are moderately correlated with the size of the language representation inside the language model (mBERT) present in the UDify architecture. However, it does not mean that this is the only parameter with a major influence on the results. Languages with more strict word order configurations tend to have higher LAS.

4.2 Dependency parsing multilingual results

In Table 4, we present the overall synergy results regarding the association of PUD corpora in terms of the number of cases, per language, where the combination of corpora provided statistically positive and negative deltas. For these experiments, each PUD language was combined in pairs with all the others (i.e.: the training sets were merged, a total of 1.200 sentences, and the development and test sets remained monolingual).

It is possible to observe that the group of languages with more than 10 cases of language combination with positive deltas is composed of Finnish, some Slavic, Germanic, and Romance languages. Nevertheless, not all PUD languages from these genera have the same positive tendency: it is the case of Polish, German, and French, all of them with less than 10 positive deltas. The Finnish language is the most favored one in terms of LAS when combined with other languages (i.e.: statistically relevant positive delta in 84% of the cases).

On the other hand, Japanese, Korean, and Thai do not obtain considerable improvement when combined with other PUD languages in terms of LAS but present many combinations which implicate a decrease in this score when compared to the baseline. Other non-Indo-European languages, such as Indonesian, Chinese, Thai, and Arabic do not benefit much from the language combinations but, at least, do not present negative synergies.

4.3 Correlations

As previously described, we calculated Pearson's and Spearman's correlation for each PUD language and for each typological strategy using the language distances from the dissimilarity matrices and the LAS deltas obtained when the languages were combined. All the correlation coefficients are displayed in the Appendix section (Tables 7 and 8)

When the obtained correlation value was between -0.7 and -0.5, it was considered a moderate inverse correlation, and a strong one for values below -0.7. In Tables 5 and 6, we present the overall results concerning the number of cases presenting either moderate or strong inverse correlation per typological strategy (Pearson's and Spearman's correlations respectively).

From the results displayed in table 5, the typological approach which provides the language classification which correlates the most with the empirical improvement in terms of LAS is the MarsaGram linear one concerning cosine distances. This approach presents a moderate or strong correlation for half of all PUD languages. It indicates that the linear order of components inside the same subtree is one of the relevant factors that may affect deep-learning systems. However, since the correlation is not observed for all languages, further research is necessary to verify the extent of this influence.

The classic classification using lang2vec syntactic features only shows a strong or moderate correlation for 7 out of the 20 PUD languages. This score is even lower than other new methods such as Head and Dependent (cosine) and MarsaGram

	Msg.	**Msg.**	**Msg.**	**Msg.**	**HD**	**HD**	**VO**	**VO**	**L2v**	**L2v**
	all	**all**	**lin.**	**lin.**						
	Euc.	**cos**	**Euc.**	**cos**	**Euc.**	**cos**	**Euc.**	**cos**	**Euc.**	**cos**
Strong	0	0	0	0	0	1	1	2	1	1
Moderate	3	8	3	10	7	7	5	2	6	5
Total	3	8	3	**10**	7	8	6	4	7	6

Table 5: Number of Pearson's correlations (moderate and strong) regarding all 20 PUD languages. In bold is highlighted the highest value regarding the total number.

	Msg.	**Msg.**	**Msg.**	**Msg.**	**HD**	**HD**	**VO**	**VO**	**L2v**	**L2v**
	all	**all**	**lin.**	**lin.**						
	Euc.	**cos**	**Euc.**	**cos**	**Euc.**	**cos**	**Euc.**	**cos**	**Euc.**	**cos**
Strong	0	1	0	0	1	2	2	0	1	1
Moderate	3	2	3	7	6	5	5	5	5	5
Total	3	3	3	**7**	**7**	**7**	**7**	5	6	6

Table 6: Number of Spearman's correlations (moderate and strong) regarding all 20 PUD languages. In bold is highlighted the highest value regarding the total number.

all properties (cosine).

5 Discussion

The results displayed in Table 4 show that as it is described in the literature, combining corpora is an effective way to improve dependency parsing scores. In our experiments, we showed that the simple association of corpora allowed us to improve significantly the LAS score for 17 out of the 20 selected languages. The ones which did not present any improvement are from linguistic families which are not well represented in the language sample. It is important to mention that all experiments were conducted in a low-resourced scenario (i.e.: corpora composed of 1,000 sentences) even though the majority of the selected languages have other annotated corpora. The idea was to find the best typological method which could be used for under-resourced languages which are the ones with the lowest LAS scores in the literature.

Moreover, from tables 5 and 6, it is possible to notice that the method with the highest number of inverse correlations is the MarsaGram linear one with language distances calculated with the cosine measure. The scores were either moderate or strong for half of the languages in the PUD collection. This specific corpus-based approach seems to be more effective than the state-of-the-art one (i.e.: using features from the lang2vec tool). Moreover, since the highest values were obtained with Pearson's correlations, it is possible to say that what is observed is a linear inverse correlation between the distances and the deltas.

However, even though the MarsaGram linear (cosine) strategy provides the most optimized results, it fails to explain the LAS values for 10 PUD languages. For Icelandic, Indonesian, and Turkish, the Pearson's correlation coefficient of this strategy is lower than -0.2, which indicates, at least, a low correlation, while for Italian, this coefficient is lower than -0.10 but higher than -0.20. On the other hand, for Chinese, Japanese, German, and Russian, this coefficient is very close to 0.00 (i.e.: no correlation). And, for Korean and Hindi, values are positive.

With the values from the dissimilarity matrix obtained using the MarsaGram linear method, it is possible to generate a dendrogram with the hclust() function using R. The classification in clusters is presented in the Annex (Figure 6). It is possible to notice some similarities with the languages' genealogy (e.g.: Romance languages in the same cluster) and with other typological classifications (e.g. OV languages on the same side of the dendrogram), however not all languages are classed following these expected configurations.

6 Conclusion and Perspectives

In this paper, we presented four corpus-based typological approaches and evaluated them in comparison with the state-of-the-art method consisting of using syntactic information from databases. First, we described these new strategies followed by the results of the dependency parsing experiments via corpora association.

We showed that the combination of corpora is an effective way to improve LAS results in low-resourced scenarios and that the typological approach concerning the order of elements inside subtrees (MarsaGram linear) is the one with the highest number of moderate and strong correlations for the languages in the PUD collection. In the future, we aim to analyze in detail the languages for which this method was not effective. Moreover, we intend to increase the number of languages to have a more homogeneous language-set in terms of the number of languages per linguistic family as well as conduct tests with non-parallel corpora. Another perspective for future work is to optimize Marsagram linear method defining weights for the features as the extracted patterns may influence the results differently.

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).

References

Diego Alves, Marko Tadić, and Božo Bekavac. 2022. Multilingual comparative analysis of deep-learning dependency parsing results using parallel corpora. In *Proceedings of the BUCC Workshop within LREC 2022*, pages 33–42.

Chiara Alzetta, Felice Dell'Orletta, Simonetta Montemagni, Petya Osenova, Kiril Simov, and Giulia Venturi. 2020. Quantitative linguistic investigations across universal dependencies treebanks. In *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. 2016a. Marsagram: an excursion in the forests of parsing trees. In *Language Resources and Evaluation Conference*, 10, page 7.

Philippe Blache, Stéphane Rauzy, and Grégoire Montcheuil. 2016b. 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. European Language Resources Association (ELRA).

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

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. 1992. The greenbergian word order correlations. *Language*, 68(1):81–138.

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

Antonio Fábregas, Jaume Mateu, and Michael T. Putnam. 2015. *Contemporary Linguistic Parameters: Contemporary Studies in Linguistics*. Bloomsbury Academic, London.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2021. Starting a new treebank? go SUD! In *Proceedings of the Sixth International Conference on Dependency Linguistics (Depling, SyntaxFest 2021)*, pages 35–46, Sofia, Bulgaria. Association for Computational Linguistics.

Bruno Guillaume, Karën Fort, and Nicolas Lefèbvre. 2016. Crowdsourcing Complex Language Resources: Playing to Annotate Dependency Syntax. In *International Conference on Computational Linguistics (COLING)*, Proceedings of the 26th International Conference on Computational Linguistics (COLING), Osaka, Japan.

John A Hawkins. 1983. *Word order universals*, volume 3. Elsevier.

John A Hawkins. 2003. Efficiency and complexity in grammars: Three general principles. *The nature of explanation in linguistic theory*, 121:152.

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

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

A Lehman. 2005. Jmp for basic univariate and multivariate statistics: a step-by-step guide. 481p.

Natalia Levshina. 2022. Corpus-based typology: applications, challenges and some solutions. *Linguistic Typology*, 26(1):129–160.

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.

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. Association for Computational Linguistics and Dublin City University.

Jerry Norman. 2009. A new look at altaic. *Journal of the American Oriental Society*, 129(1):83–89.

Robert Östling. 2015. Word order typology through multilingual word alignment. 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 205–211.

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.

Karl Pearson. 1895. Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58(347-352):240–242.

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. Association for Computational Linguistics.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019a. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Edoardo Maria Ponti, Helen O'horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019b. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Kaius Sinnemäki. 2014. Complexity trade-offs: A case study. In *Measuring grammatical complexity*, pages 179–201. Oxford University Press.

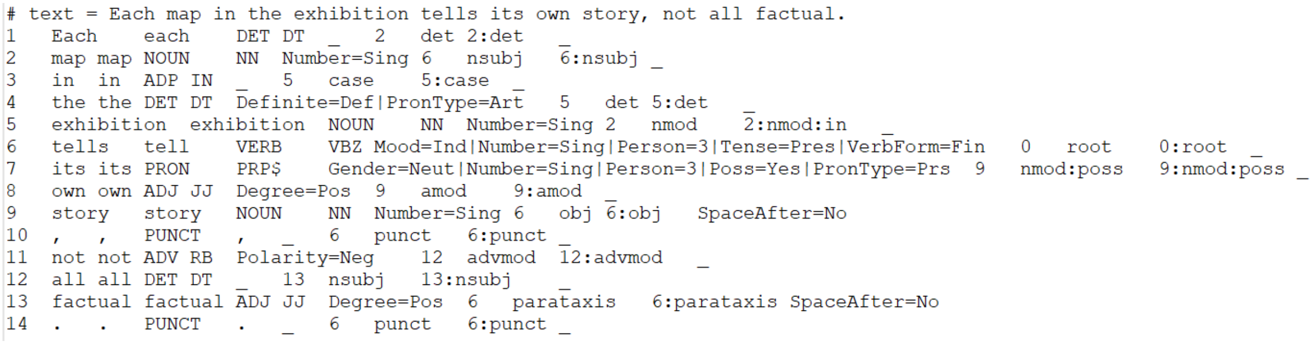
Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. UDapter: Language adaptation for truly Universal Dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2302–2315, Online. Association for Computational Linguistics.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Bernhard Wälchli. 2009. Data reduction typology and the bimodal distribution bias. 13(1):77–94.

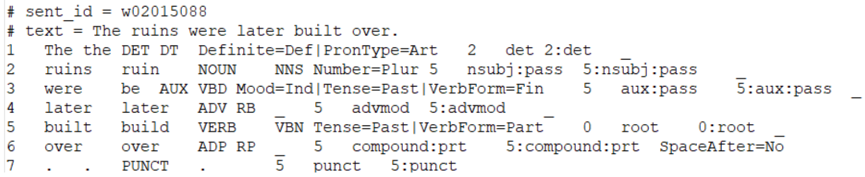
Daniel Zeman, Jan Hajic, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. Conll 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies*, pages 1–21.

A Appendix



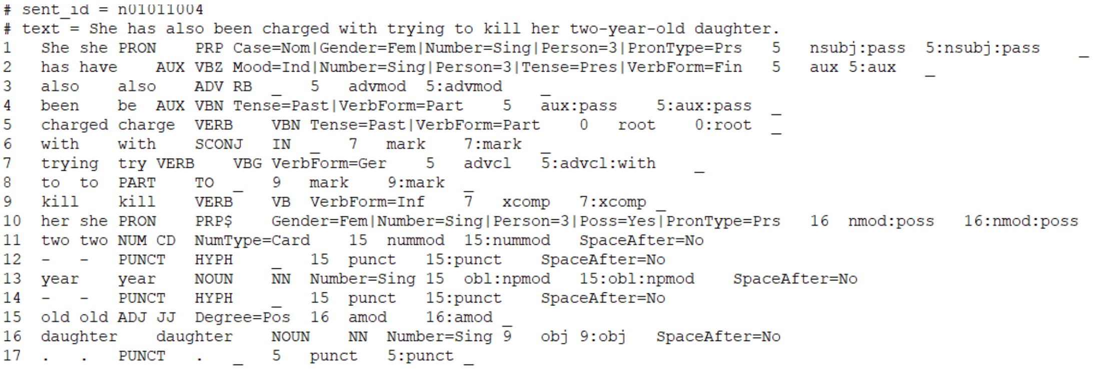
```
# text = Each map in the exhibition tells its own story, not all factual.
1   Each    each    DET DT  _   2   det 2:det   _
2   map map NOUN    NN  Number=Sing 6   nsubj   6:nsubj _
3   in  in  ADP IN  _   5   case    5:case  _
4   the the DET DT  Definite=Def|PronType=Art   5   det 5:det   _
5   exhibition  exhibition  NOUN    NN  Number=Sing 2   nmod    2:nmod:in   _
6   tells   tell    VERB    VBZ Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   0   root    0:root  _
7   its its PRON    PRP$    Gender=Neut|Number=Sing|Person=3|Poss=Yes|PronType=Prs  9   nmod:poss   9:nmod:poss _
8   own own ADJ JJ  Degree=Pos  9   amod    9:amod  _
9   story   story   NOUN    NN  Number=Sing 6   obj 6:obj   SpaceAfter=No
10  ,   ,   PUNCT   ,   _   6   punct   6:punct _
11  not not ADV RB  Polarity=Neg    12  advmod  12:advmod   _
12  all all DET DT  _   13  nsubj   13:nsubj    _
13  factual factual ADJ JJ  Degree=Pos  6   parataxis   6:parataxis SpaceAfter=No
14  .   .   PUNCT   .   _   6   punct   6:punct _
```

Figure 1: Example of a sentence with the pattern NOUN_precede_DET-det_NOUN-nmod rom the PUD English corpus. The determiner (DET) on line 4 has the incoming relation det. It precedes the noun (NOUN) on line 5, which has the incoming relation nmod. Both appear in the subtree headed by a NOUN (the first tag in the pattern description); in this case, it is again the noun on line 5.



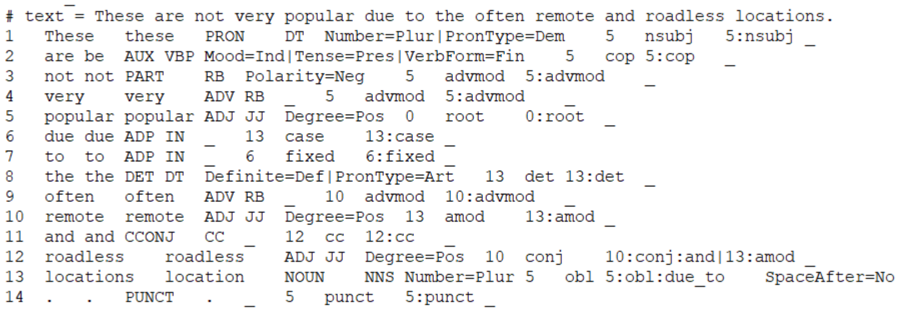
```
# sent_id = w02015088
# text = The ruins were later built over.
1   The the DET DT  Definite=Def|PronType=Art   2   det 2:det   _
2   ruins   ruin    NOUN    NNS Number=Plur 5   nsubj:pass  5:nsubj:pass    _
3   were    be  AUX VBD Mood=Ind|Tense=Past|VerbForm=Fin    5   aux:pass    5:aux:pass  _
4   later   later   ADV RB  _   5   advmod  5:advmod    _
5   built   build   VERB    VBN Tense=Past|VerbForm=Part    0   root    0:root  _
6   over    over    ADP RP  _   5   compound:prt    5:compound:prt  SpaceAfter=No
7   .   .   PUNCT   .   _   5   punct   5:punct
```

Figure 2: Example of a sentence with the pattern VERB_require_NOUN-nsubj:pass_AUX-aux:pass from the PUD English corpus. The noun (NOUN) on line 2 has the incoming relation nsubj:pass. It requires the auxilary (AUX) on line 3, which has the incoming relation aux:pass. Both appear in the subtree headed by a VERB (token "built" on line 5).



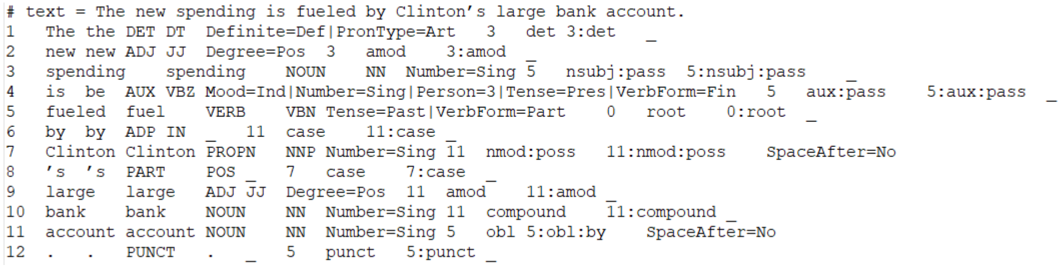
```
# sent_id = n01011004
# text = She has also been charged with trying to kill her two-year-old daughter.
1   She she PRON    PRP Case=Nom|Gender=Fem|Number=Sing|Person=3|PronType=Prs   5   nsubj:pass  5:nsubj:pass    _
2   has have    AUX VBZ Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   5   aux 5:aux   _
3   also    also    ADV RB  _   5   advmod  5:advmod    _
4   been    be  AUX VBN Tense=Past|VerbForm=Part    5   aux:pass    5:aux:pass  _
5   charged charge  VERB    VBN Tense=Past|VerbForm=Part    0   root    0:root  _
6   with    with    SCONJ   IN  _   7   mark    7:mark  _
7   trying  try VERB    VBG VerbForm=Ger    5   advcl   5:advcl:with    _
8   to  to  PART    TO  _   9   mark    9:mark  _
9   kill    kill    VERB    VB  VerbForm=Inf    7   xcomp   7:xcomp _
10  her she PRON    PRP$    Gender=Fem|Number=Sing|Person=3|Poss=Yes|PronType=Prs   16  nmod:poss   16:nmod:poss    _
11  two two NUM CD  NumType=Card    15  nummod  15:nummod   SpaceAfter=No
12  -   -   PUNCT   HYPH    _   15  punct   15:punct    SpaceAfter=No
13  year    year    NOUN    NN  Number=Sing 15  obl:npmod   15:obl:npmod    SpaceAfter=No
14  -   -   PUNCT   HYPH    _   15  punct   15:punct    SpaceAfter=No
15  old old ADJ JJ  Degree=Pos  16  amod    16:amod _
16  daughter    daughter    NOUN    NN  Number=Sing 9   obj 9:obj   SpaceAfter=No
17  .   .   PUNCT   .   _   5   punct   5:punct _
```

Figure 3: Example of a sentence with the pattern ADJ_unicity_NOUN-obl:npmod from the PUD English corpus. The head of the subtree is the token "old" (ADJ) on line 15. The element on line 13 ("year") has the part-of-speech of noun (NOUN) and the dependency relation of obl:npmod and no other element with the same characteristics can be found inside the same subtree.



```
# text = These are not very popular due to the often remote and roadless locations.
1   These   these   PRON    DT  Number=Plur|PronType=Dem    5   nsubj   5:nsubj _
2   are be  AUX VBP Mood=Ind|Tense=Pres|VerbForm=Fin    5   cop 5:cop   _
3   not not PART    RB  Polarity=Neg    5   advmod  5:advmod    _
4   very    very    ADV RB  _   5   advmod  5:advmod    _
5   popular popular ADJ JJ  Degree=Pos  0   root    0:root  _
6   due due ADP IN  _   13  case    13:case _
7   to  to  ADP IN  _   6   fixed   6:fixed _
8   the the DET DT  Definite=Def|PronType=Art   13  det 13:det  _
9   often   often   ADV RB  _   10  advmod  10:advmod   _
10  remote  remote  ADJ JJ  Degree=Pos  13  amod    13:amod _
11  and and CCONJ   CC  _   12  cc  12:cc   _
12  roadless    roadless    ADJ JJ  Degree=Pos  10  conj    10:conj:and|13:amod _
13  locations   location    NOUN    NNS Number=Plur 5   obl 5:obl:due_to    SpaceAfter=No
14  .   .   PUNCT   .   _   5   punct   5:punct _
```

Figure 4: Example of a sentence with two occurrences of the pattern ADV_advmod_precedes_ADJ. The adverb (ADV) on line 9 has the incoming relation advmod. It precedes the adjective (ADJ) on line 10. And, the adverb (ADV) on line 4 has the incoming relation advmod. It precedes the adjective (ADJ) on line 5.



```
# text = The new spending is fueled by Clinton’s large bank account.
1   The the DET DT  Definite=Def|PronType=Art   3   det 3:det   _
2   new new ADJ JJ  Degree=Pos  3   amod    3:amod _
3   spending    spending    NOUN    NN  Number=Sing 5   nsubj:pass  5:nsubj:pass    _
4   is  be  AUX VBZ Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   5   aux:pass    5:aux:pass  _
5   fueled  fuel    VERB    VBN Tense=Past|VerbForm=Part    0   root    0:root  _
6   by  by  ADP IN  _   11  case    11:case _
7   Clinton Clinton PROPN   NNP Number=Sing 11  nmod:poss   11:nmod:poss    SpaceAfter=No
8   ’s  ’s  PART    POS _   7   case    7:case  _
9   large   large   ADJ JJ  Degree=Pos  11  amod    11:amod _
10  bank    bank    NOUN    NN  Number=Sing 11  compound    11:compound _
11  account account NOUN    NN  Number=Sing 5   obl 5:obl:by    SpaceAfter=No
12  .   .   PUNCT   .   _   5   punct   5:punct _
```

Figure 5: Example of a sentence with the pattern NOUN_obl_follows_VERB. The noun (NOUN) on line 11 has the incoming relation obl. It comes after the verb (VERB) on line 5.

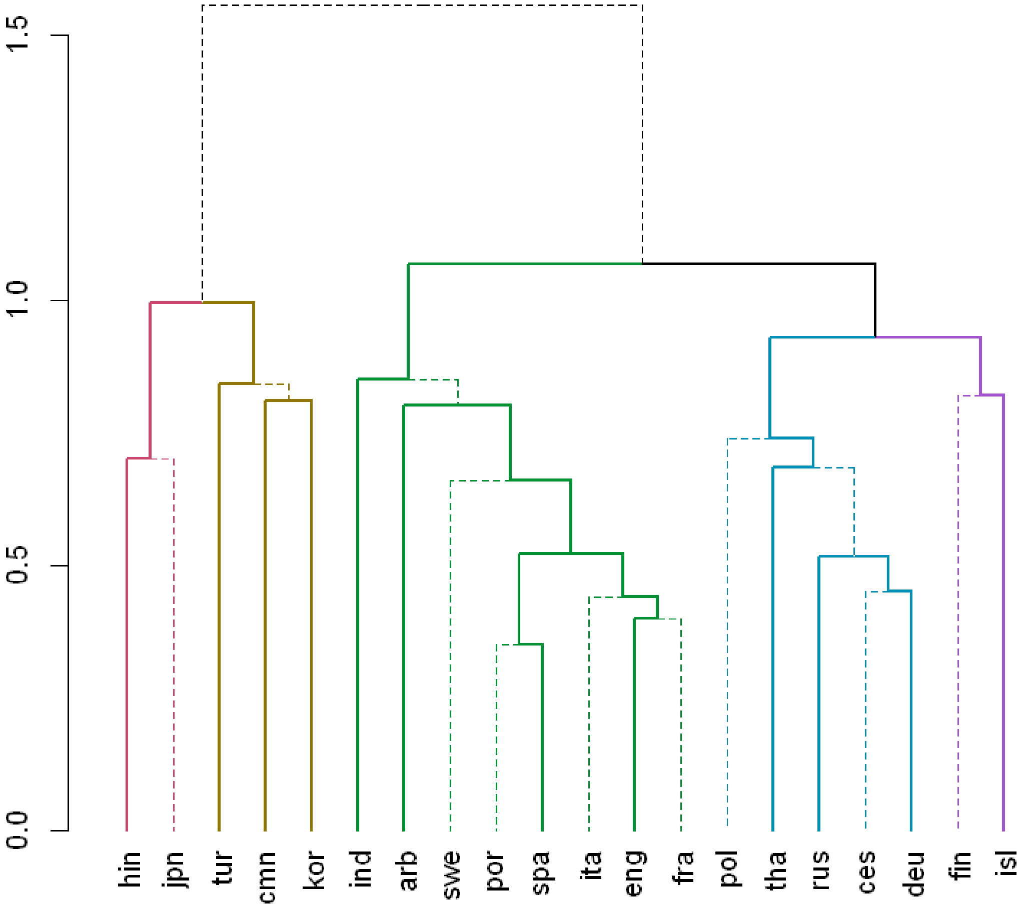


Figure 6: Marsagram Linear cosine Dendrogram

	Msg.	**Msg.**	**Msg.**	**Msg.**	**HD**	**HD**	**VO**	**VO**	**L2v**	**L2v**
	all	**all**	**lin.**	**lin.**						
	Euc.	**cos**	**Euc.**	**cos**	**Euc.**	**cos**	**Euc.**	**cos**	**Euc.**	**cos**
arb	-0.11	-0.52	-0.03	-0.57	-0.54	-0.65	-0.59	-0.47	-0.59	-0.55
cmn	0.19	-0.11	-0.26	0.00	0.25	0.15	-0.06	-0.21	0.01	-0.03
ces	-0.25	-0.60	-0.28	-0.57	-0.65	-0.67	-0.57	-0.57	-0.36	-0.28
eng	-0.34	-0.53	-0.21	-0.59	-0.41	-0.49	-0.35	-0.16	-0.36	-0.41
fin	-0.16	-0.52	-0.46	-0.63	-0.46	-0.44	-0.71	-0.72	-0.10	-0.01
fra	-0.50	-0.51	-0.52	-0.62	-0.62	-0.59	-0.38	-0.31	-0.50	-0.47
deu	-0.48	-0.11	-0.22	-0.03	-0.23	-0.22	0.03	0.46	-0.03	-0.02
hin	-0.36	-0.27	0.05	0.40	0.12	0.41	0.56	0.50	0.44	0.46
isl	0.18	-0.19	-0.26	-0.36	-0.12	-0.31	-0.49	-0.44	-0.40	-0.42
ind	0.23	-0.30	0.20	-0.21	0.12	0.05	0.00	0.05	-0.21	-0.11
ita	-0.21	-0.23	-0.02	-0.13	-0.14	-0.17	-0.30	-0.16	-0.10	-0.17
jpn	-0.18	0.06	-0.15	-0.05	0.38	0.35	0.02	0.07	0.40	0.50
kor	0.30	0.29	0.08	0.38	0.42	0.49	0.41	0.47	0.43	0.37
pol	-0.23	-0.37	-0.50	-0.62	-0.13	-0.34	-0.51	-0.40	-0.37	-0.34
por	-0.64	-0.52	-0.39	-0.61	-0.64	-0.53	-0.45	-0.40	-0.57	-0.50
rus	-0.16	-0.08	0.17	0.03	-0.27	-0.24	-0.46	-0.28	-0.15	-0.17
spa	-0.59	-0.45	-0.57	-0.51	-0.53	-0.50	-0.43	-0.38	-0.60	-0.55
swe	-0.48	-0.59	-0.31	-0.64	-0.58	-0.63	-0.59	-0.49	-0.70	-0.68
tha	0.26	-0.59	-0.22	-0.62	-0.64	-0.88	-0.60	-0.80	-0.76	-0.81
tur	-0.09	0.10	-0.34	-0.25	-0.45	-0.53	-0.42	-0.56	-0.61	-0.60

Table 7: Pearson's correlation values regarding all 20 PUD languages.

	Msg.	**Msg.**	**Msg.**	**Msg.**	**HD**	**HD**	**VO**	**VO**	**L2v**	**L2v**
	all	**all**	**lin.**	**lin.**						
	Euc.	**cos**	**Euc.**	**cos**	**Euc.**	**cos**	**Euc.**	**cos**	**Euc.**	**cos**
arb	-0.05	-0.33	-0.09	-0.53	-0.55	-0.66	-0.65	-0.52	-0.70	-0.69
cmn	0.24	-0.15	-0.12	-0.08	0.36	0.18	0.03	-0.12	-0.02	-0.03
ces	-0.14	-0.54	-0.31	-0.49	-0.51	-0.48	-0.52	-0.57	-0.31	-0.31
eng	-0.38	-0.59	-0.27	-0.48	-0.49	-0.52	-0.46	-0.02	-0.37	-0.38
fin	-0.20	-0.48	-0.41	-0.60	-0.35	-0.44	-0.74	-0.66	-0.09	-0.06
fra	-0.50	-0.48	-0.55	-0.59	-0.57	-0.56	-0.47	-0.26	-0.50	-0.53
deu	-0.52	-0.28	-0.22	-0.03	-0.30	-0.29	0.05	0.44	-0.09	-0.08
hin	-0.31	-0.23	0.06	0.32	-0.05	0.34	0.68	0.60	0.43	0.44
isl	0.24	-0.19	-0.21	-0.46	-0.03	-0.20	-0.50	-0.26	-0.43	-0.44
ind	0.13	-0.27	0.02	-0.22	0.04	0.01	-0.16	-0.24	-0.29	-0.23
ita	-0.23	-0.31	-0.02	-0.11	-0.16	-0.15	-0.24	0.12	-0.20	-0.20
jpn	0.08	0.16	-0.01	-0.26	0.45	0.52	-0.10	-0.16	0.50	0.49
kor	0.52	0.34	0.13	0.52	0.18	0.53	0.22	0.17	0.24	0.27
pol	-0.29	-0.44	-0.67	-0.62	-0.23	-0.42	-0.55	-0.48	-0.31	-0.31
por	-0.42	-0.29	-0.23	-0.37	-0.41	-0.42	-0.49	-0.47	-0.48	-0.47
rus	-0.01	-0.14	0.16	0.07	-0.08	-0.09	-0.46	-0.06	-0.08	-0.06
spa	-0.51	-0.45	-0.55	-0.55	-0.56	-0.53	-0.50	-0.55	-0.67	-0.66
swe	-0.46	-0.73	-0.38	-0.68	-0.70	-0.74	-0.80	-0.40	-0.64	-0.63
tha	0.25	-0.49	-0.19	-0.62	-0.51	-0.81	-0.36	-0.69	-0.68	-0.70
tur	0.09	-0.15	-0.26	-0.18	-0.59	-0.69	-0.15	-0.31	-0.59	-0.57

Table 8: Spearman's correlation values regarding all 20 PUD languages.