Analysis of Corpus-based Word-Order Typological Methods

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. We compared three specific quantitative methods, using parallel CoNLL-U corpora, to the classification obtained via syntactic features provided by a typological database (lang2vec). First, we analyzed the Marsagram linear approach which consists of extracting the frequency word-order patterns regarding the position of components inside syntactic nodes. The second approach considers the relative position of heads and dependents, and the third is based simply on the relative position of verbs and objects. From the results, it was possible to observe that each method provides different language clusters which can be compared to the classic genealogical classification (the lang2vec and the head and dependent methods being the closest). As different word-order phenomena are considered in these specific typological strategies, each one provides a different angle of analysis to be applied according to the precise needs of the researchers.

1 Introduction

Typology is usually described as language classification regarding structural types. Its scope can be defined as the quest for answers about how languages differ from each other, and about the explanation for the attested differences and similarities.

In terms of syntactic typology, one possible linguistic aspect that is analyzed concerns word-order patterns. These phenomena are commonly used to define sets of typological universals in terms of implications, correlations, and universals.

Most studies in this field rely on the identification of the most frequent word-order phenomena in different languages. Although based on attested syntactic constructions, what is extracted from the available linguistic data concerns only the most common syntactic structures. Thus, possible word-order patterns which are not the standard ones are usually ignored in these analyses. It is the case of the syntactic information provided by standard typological databases such as WALS (Dryer and Haspelmath, 2013). Although limited, these databases provide valuable information for theoretical typological analyses and can be used to improve the effectiveness of Natural Language Processing (NLP) tools, as shown by (Ponti et al., 2019).

On the other hand, corpus-based typological studies can provide a more precise description in terms of possible syntactic phenomena, thus, allowing languages to be compared in a more detailed way, as presented by (Levshina, 2022). Quantitative methods can be used in the analysis of numerous linguistic phenomena, and, even though they can present some bias regarding the corpora selection and annotation, they provide new insights that can challenge and/or complement classic theoretical approaches.

The aim of this article is to propose three different corpus-based quantitative methods concerning word-order typology and compare the obtained language classifications to the one provided by the comparison of the syntactic features provided from a typological database. The objective is to show that different approaches provide valuable but diverse contributions in terms of word-order structures attested in annotated corpora.

The paper is composed as follows: Section 2 presents an overview of the related work to this topic. Section 3 describes the campaign design: the language and data-set selection and the syntactic typological approaches; Section 4 present the obtained results which are discussed in Section 5. In Section 6 we provide conclusions and possible future directions for the research.

2 Related Work

According to (Ponti et al., 2019), the WALS database is one of the most used typological resources in NLP studies as it contains phonolog-

Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023), pages 36 - 46
March 9-12, 2023 ©2023 Association for Computational Linguistics

ical, morphosyntactic, and lexical information for a large number of languages. Besides that, the URIEL Typological Compendium is a meta-repository composed of several databases (WALS included) and is the base of the lang2vec tool (Littell et al., 2017). This specific resource provides typological information about languages in the format of feature and value pairs. Thus languages can be represented by vectors which are composed of the selected linguistic information required by the user (e.g.: genealogical, phonological, syntactic, etc). One problem usually observed in these databases is the fact that they suffer from discrepancies that are caused by their variety of sources. Therefore, comparisons can only be made if the selected languages have values for the ensemble of chosen features. Furthermore, there are many gaps as not all languages have the same amount of descriptive literature. Moreover, as previously mentioned, most databases fail to illustrate the variations that can occur within a single language (i.e.: only the most frequent phenomena are reported, and not all possible ones). On the other hand, quantitative methods, such as the ones proposed in this article, provide precise information regarding the frequency of all attested word-order phenomena inside the analyzed corpora.

An extended survey of corpora-based typological studies was provided by (Levshina, 2022). While certain authors quantitively analyzed specific word-order patterns (e.g.: subject, verb, and object position (Östling, 2015), and verb and locative phrases (Wälchli, 2009)), other authors have focused on quantitative analyses regarding language complexity (e.g.: (Hawkins, 2003) and (Sinnemäki, 2014)).

With the aim of examining diachronic syntactic changes that characterize the evolution from Latin to Romance languages, (Liu and Xu, 2012) proposed a quantitative approach to analyze the distributions of dependency directions. In total, 15 modern languages (8 Romance languages and 7 from other families) and 2 ancient ones (Latin and Ancient Greek) were scrutinized by the extraction of syntactic information from annotated corpora. The attested dependency syntactic networks for each language were analyzed with the calculation of certain syntactic parameters extracted from each corpus (i.e.: the mean sentential length, the percentage of the head-final dependencies, the head-initial dependencies, the dependencies between adjacent words, and of dependencies between non-adjacent words, the mean distance of all head-final dependencies, and the mean distance of all head-initial dependencies). It has been shown that the dependency syntactic networks arising from the selected data-sets reflect the degree of inflectional variation of each language. The adopted clustering approach also allowed Romance languages to be differentiated from Latin diachronically and between each other synchronically. However, the authors used data from the shared tasks of CoNLL 2006 (Buchholz and Marsi, 2006) and CoNLL 2007 (Nivre et al., 2007), however the dependency annotation schemes differed substantially from each other, so any studies based on those treebanks were problematic.

Another method concerning the extraction and comparison of syntactic information from treebanks was proposed by (Blache et al., 2016a). They developed the Marsagram tool, a resource that allows typological syntactic information (together with its statistics) to be obtained by inferring context-free grammars from syntactic structures inside annotated corpora. In terms of word-order, this tool allows the extraction of linear patterns (i.e.: if a specific part-of-speech precedes another one inside the same node of the syntactic tree governed 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 quantitative methods. Thus, one of the corpus-based typological approaches to be examined and compared in this article concerns the linear patterns provided by Marsagram tool.

The concept of Typometrics was introduced by (Gerdes et al., 2021). The authors extracted rich details for testing typological implicational universals and explored new kinds of universals, named quantitative universals. In their study, different word-order phenomena were analyzed quantitatively (i.e.: the distribution of their occurrences in annotated corpora) to identify universals (i.e.: present in all or most languages). Our approach differs from theirs as our aim is not to identify these implications or correlations but to compare languages (i.e.: language vectors) using all syntactic structures identified in the corpora to obtain a more general syntactic overview of the elements in our language set.

What is possible to observe in many studies regarding corpus-based typology is that usually a method is presented without a specific comparison to the existing approaches or to the classic one concerning the typological databases. Moreover, usually, the selected corpora are not completely homogeneous in terms of size or genre. Thus, in this study, the idea is to compare 20 different languages by using parallel corpora. (Levshina, 2022) showed the benefit of using this type of data, as the bias regarding size and content is avoided. Especially in this case, where syntactic patterns are the center of the analysis, 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. Our objective is not to determine which is the best corpus-based approach, but to show how data can be explored from different angles, allowing typological nuances to be analyzed in detail.

3 Campaign Design

In this section, a brief overview of the selected datasets is provided, followed by a complete description of the syntactic typological approaches which were selected to conduct the corpus-based word-order analyses.

3.1 Parallel Corpora

The Parallel Universal Dependencies (PUD) compilation is an ensemble of tree-banks (parallel annotated corpora following Universal Dependencies guidelines (De Marneffe et al., 2021)) that was developed for the CoNLL 2017 shared task on Multilingual Parsing from Raw Text to Universal Dependencies (Zeman et al., 2018). It provides 1,000 parallel sentences from news sources and Wikipedia annotated 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. As previously explained, we decided to conduct the experiments with parallel annotated corpora to avoid biases regarding semantic content and size. However, as the PUD corpora are composed of translations from English (750 sentences), German (100), French (50), Spanish (50), and Italian (50), they may contain some "translationese" biases as 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 and geographical information[2] is provided in Table 1.

The number of languages in this study is limited to 20 as we decided to focus on parallel data analysis. However, PUD collection provides, at least, some variety in terms of genealogy (i.e.: the great majority belongs to the Indo-European family, but 8 other different linguistic families are also present in this data-set). In terms of geographical areas, most languages are from the Eurasia region, the exceptions are Arabic (Africa), Chinese, Indonesian, and Thai (these 3 being from Southeast and Oceania region). The geographical areas presented in this article correspond to the ones described by (Dryer, 1992) and contain some discrepancies when compared to the ones proposed by WALS (Dryer and Haspelmath, 2013) (e.g.: while (Dryer, 1992) considers Arabic as an African language, in WALS, it is associated to Eurasia geographical area).

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

3.2 Typological Approaches

The main idea is to generate, for each method, language vectors whose features correspond to specific word-order features and the values, to the frequency of the syntactic phenomenon in each corpus. With these vectors, languages are compared using Euclidean distance measures, generating dissimilarity matrices that can be, later, visually analyzed using a clustering algorithm.

The obtained classifications using the quantitative strategies are compared to the one provided by the clustering analysis conducted with typological information (syntactic features) provided by lang2vec tool (i.e.: the lang2vec classification is considered as our baseline).

Three typological approaches were chosen:

- Marsagram linear patterns
- Head and Dependent relative position
- Verb and Object relative position

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

[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.

Language	ISO 639-3	Family	Genus	Geographical Area
Arabic	arb	Afro-Asiatic	Semitic	Africa
Chinese	cmn	Sino-Tibetan	Chinese	Southeast Asia and Oceania
Czech	ces	Indo-European	Slavic	Eurasia
English	eng	Indo-European	Germanic	Eurasia
Finnish	fin	Uralic	Finnic	Eurasia
French	fra	Indo-European	Romance	Eurasia
German	deu	Indo-European	Germanic	Eurasia
Hindi	hin	Indo-European	Indic	Eurasia
Icelandic	isl	Indo-European	Germanic	Eurasia
Indonesian	ind	Austronesian	Malayo-Sumbawan	Southeast Asia and Oceania
Italian	ita	Indo-European	Romance	Eurasia
Japanese	jpn	Japanese	Japanese	Eurasia
Korean	kor	Korean	Korean	Eurasia
Polish	pol	Indo-European	Slavic	Eurasia
Portuguese	por	Indo-European	Romance	Eurasia
Russian	rus	Indo-European	Slavic	Eurasia
Spanish	spa	Indo-European	Romance	Eurasia
Swedish	swe	Indo-European	Germanic	Eurasia
Thai	tha	Tai-Kadai	Kam-Tai	Southeast Asia and Oceania
Turkish	tur	Altaic	Turkic	Eurasia

Table 1: List of languages inside PUD collection, their respective ISO 639-3 three-character code, their genealogical information according to WALS, and the Geographical Area provided by (Dryer, 1992)

More details regarding the lang2vec analysis and each one of the new approaches are provided in the following sub-sections.

Thus, for each method, we first generate the 20 language vectors relative to the ensemble of PUD languages. Then, using the dist() R function, we obtain the dissimilarity matrices which are used for the clustering analysis.

In terms of hierarchical clustering methods, the Ward linkage method (Ward Jr, 1963) is applied to the obtained dissimilarity matrices. This strategy, instead of minimizing possible distances between pairs of clusters, minimizes the sum of squared differences within all clusters, thus, being a variance-minimizing approach. This agglomeration strategy has been chosen as its efficiency has been proven in many studies in the field of corpus-based linguistics and related disciplines (Eder, 2017). With the programming language R, it is possible to generate language clusters using the chosen linkage method with the function hclust() and the specific argument (method="ward.D2").

In the Results section, the different clustering classifications are presented, analyzed, and compared.

3.2.1 Lang2vec

As mentioned before, the lang2vec tool (Littell et al., 2017) is a valuable resource that provides typological information in the format of language vectors. In our case, lang2vec syntactic vectors are used. They describe languages morphosyntactically with information coming from the WALS database (Dryer and Haspelmath, 2013), the Syntactic Structures of World Languages (SSWL)[3], and Ethnologue [4].

In terms of syntactic features, the average vector (i.e.: compiling all possible features from the different databases) is composed of 103 features. The number of valid features (i.e.: with a specific value associated with it) varies from language to language. Each feature can receive the following values:

- 0.00 – the absence of the phenomenon
- 0.33 – the phenomenon can be observed but is not common
- 0.50 – the phenomenon is commonly observed together with other possible word-orders

[3] http://sswl.railsplayground.net/

[4] https://www.ethnologue.com/

- 0.67 – the phenomenon is relatively common.
- 1.00 – the phenomenon is normally encountered in the language.

There is a great discrepancy in terms of the availability of syntactic information regarding lang2vec syntactic features among PUD languages. It varies from 66 valid features for Arabic to 103 for English. Moreover, when checking the number of common valid features of all PUD languages, the final amount is 41 (i.e.: lang2vec PUD language vectors have 41 dimensions).

In terms of word-order phenomena described by the 41 common features composing the lang2vec PUD vectors, they correspond to:

- Subject, verb, and object (e.g.: SVO, SOV, SUBJECT_BEFORE_VERB)
- Adposition and noun (e.g.: ADPOSITION_BEFORE_NOUN)
- Possessor and noun (e.g.: POSSESSOR_AFTER_NOUN)
- Adjective and noun (e.g.: ADJECTIVE_AFTER_NOUN)
- Demonstrative and noun (e.g.: DEMONSTRATIVE_WORD_BEFORE_NOUN)
- Numeral and noun (e.g.: NUMERAL_AFTER_NOUN)
- Negative word and verb (e.g.: NEGATIVE_WORD_BEFORE_VERB)
- Degree word and adjective (e.g.: DEGREE_WORD_BEFORE_ADJECTIVE)
- Subordinator word and clause (e.g.: SUBORDINATOR_WORD_AFTER_CLAUSE)
- Polar question particle position: initial or final (e.g.: POLARQ_MARK_INITIAL)
- Existence of demonstrative prefix or suffix (e.g.: DEMONSTRATIVE_PREFIX)
- Existence of negative prefix or suffix (e.g.: NEGATIVE_PREFIX)
- Existence of TEND prefix or suffix (e.g.: TEND_SUFFIX)
- Existence of case mark, enclitic, proclitic, prefix, and suffix (e.g.: CASE_ENCLITIC)

We decided to use all the syntactic features available in lang2vec which are common to all PUD languages even if some of them are not directly related to word-order phenomena because when lang2vec vectors are used for experiments regarding the improvement of Natural Language Processing results, the whole set of lang2vec features is used.

3.2.2 Marsagram Linear Patterns

Marsagram is a tool for exploring treebanks, it extracts context-free grammars (CFG) from annotated data-sets that allow statistical comparison between languages as proposed by (Blache et al., 2016b). We have used the latest release of this software[5] available in the ORTOLANG platform of linguistic tools and resources.

This software identifies four types of properties: precede, require, exclude, and unicity. However, since the focus of this study is on word-order patterns, only "precede" property (linear) is considered. The extracted syntactic patterns contain information concerning part-of-speech and dependency parsing labels as well as the associated property type.

For example, *NOUN_precede_DET-det_NOUN-nmod* which means that a *DET* which has the dependency relation *det* precedes a *NOUN* with *nmod* as dependency label in the context of a node having *NOUN* as the head. An example of a sentence with this pattern is presented in the Appendix section (Figure 5). For each identified word-order phenomenon, Marsagram also indicates its frequency inside the corpus.

As expected, some patterns are common to all languages and some of them appear only in one or a few corpora. Therefore, the typological classification provided here concerns all possible identified rules (with an associated frequency value equal to zero for languages in which the pattern does not appear). In total, 21,242 linear patterns are extracted from the PUD collection (i.e.: the union of all patterns identified in PUD languages). The average amount of patterns with a frequency different from 0 is 15,790. However, even though only parallel corpora are considered, the number of extracted properties occurring in the corpora varies considerably among different languages: less than 10,000 for Japanese and Korean and more than 20,000 for English, Hindi, and Icelandic. The other PUD languages have a number of properties closer to the

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

average.

All the linear patterns that were identified with the Marsagram tool were considered when building the language vectors, even if they do not represent real dependency structures (e.g.: coordination phenomena). The main focus of the research is to obtain different quantitative typological classifications which can be used for dependency parsing improvement, thus, it is relevant to keep all the identified patterns.

3.2.3 Head and Dependent Relative Position

To analyze 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 reflects the directionality observed at the surface level (position of head and dependent observed at the sentence level).

Thus, using a python script, the attested head and dependent relative position patterns are extracted together with their frequency of occurrence in each corpus. All observed features extracted from the PUD corpora (2,890 in total) have been included in the language vectors. In the cases where a feature is not observed in a determined corpus, the value 0 is 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 6).

- 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 7).

The analysis of these patterns corresponds to a quantitative approach of the Head and Dependent theory (Hawkins, 1983) which considers that there is a tendency of organizing head and dependents in homogeneous word ordering. (Hawkins, 1983) proposed a set of language types according to specific word-order phenomena concerning a limited list of heads and dependents. In this study, we consider all possible head and dependent pairs to compare the languages and classify them.

3.2.4 Verb and Object relative position

The verb (V) and direct object (O) relative position is part of the analysis regarding the heads and dependents ordering. We decided to analyze specifically the position of these two elements as they are key in typological studies such as the one proposed by (Dryer, 1992) where the correlations are defined according to whether in a language the verb comes before or after the object (i.e.: dependency relation "obj").

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). The idea is to go beyond the classical approaches which usually consider only nominal objects (e.g.: (Dryer, 1992)) to see how languages are classified if all possible direct objects are analyzed. 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.

4 Results

As explained previously, each one of the presented typological methods generates a cluster dendrogram which is displayed in this section (Figures 1 to 4).

Starting with the lang2vec dendrogram (1), it is possible to notice that the central cluster is divided into two sub-groups (one composed of Chinese, English, and Swedish, and the other of Finnish, German, Icelandic, and the Slavic PUD languages). Arabic is classified in the same sub-group as Indonesian and Romance languages.

It is also noticeable that Hindi, Korean, Japanese, and Turkish form an isolated cluster. Moreover, Germanic languages are split into two sub-clusters, one formed by English and Swedish, together with Chinese, and the other composed of German and Icelandic (grouped with Slavic languages). Regarding this specific genus, although Polish and

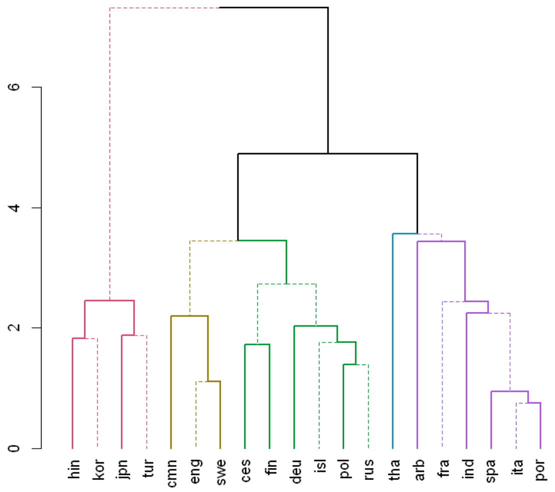


Figure 1: Lang2vec Clustering Dendrogram

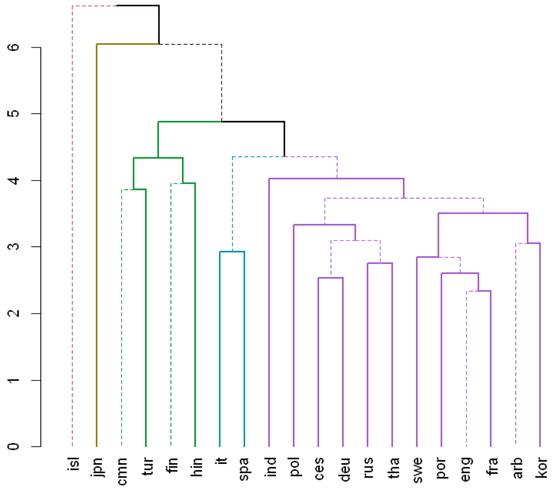


Figure 2: Marsagram Linear Properties Clustering Dendrogram

Russian are closer in both dendrograms, Czech is positioned closer to Finnish. Furthermore, as previously mentioned, when considering only lang2vec syntactic features, Thai and Arabic are classified as closer to Romance languages when compared to the others in the PUD collection.

If we consider the 3 main clusters provided by this dendrogram, it is possible to analyze which syntactic lang2vec features are shared by its languages. The isolated cluster formed by Hindi, Japanese, Korean and Turkish is composed of SOV languages with postpositions and with adjectives before nouns. The middle cluster (i.e.: Slavic and Germanic languages, plus Chinese and Finnish) has SVO languages with adjectives before nouns. And, finally, the cluster on the right side of the dendrograms is composed of VO (but not necessarily SVO) languages with prepositions and adjectives after the noun. Moreover, this cluster differs from the one located on the extreme left side of the dendrograms by ordering the negative word before the verb.

The analysis of the dendrogram concerning Marsagram linear patterns (2) shows Icelandic as an isolated language inside PUD collection. Japanese is also quite isolated from the other languages, however with lower distance values than Icelandic. Chinese, Turkish, Finnish, and Hindi form one small central cluster, as well as Italian and Spanish, and the other languages are grouped altogether. For this specific representation, languages from the same family or genus are not always clustered together (e.g.: Portuguese and Spanish, which formed a sub-cluster in lang2vec dendrograms in this case).

The large purple cluster is composed of Romance, Slavic, and Germanic languages (except Icelandic), but it also includes Indonesian, Arabic, Thai and Korean. Subject, Verb, and Object positions are not relevant criteria in this type of language classification. Marsagram extracts word-order patterns between elements that are part of the same syntactic node, thus, these components are not necessarily syntactically related.

When considering the classification provided by the dendrogram obtained with the head and dependent patterns (3), we observe that the Romance languages form one single isolated cluster positioned on the left side of the figure. On the other hand, the Germanic sub-group is closer to the Slavic one (with Icelandic being positioned inside the Slavic cluster and not with the other Germanic PUD languages). It is also noticeable that Thai, Arabic, and Indonesian form a specific sub-group closer to the Germanic and Slavic languages.

Japanese is clustered with Hindi and closer to other OV languages (i.e.: Turkish and Korean). The large cluster containing all OV languages from PUD also includes Finnish and Chinese which are not OV. When compared to the genealogical clas-

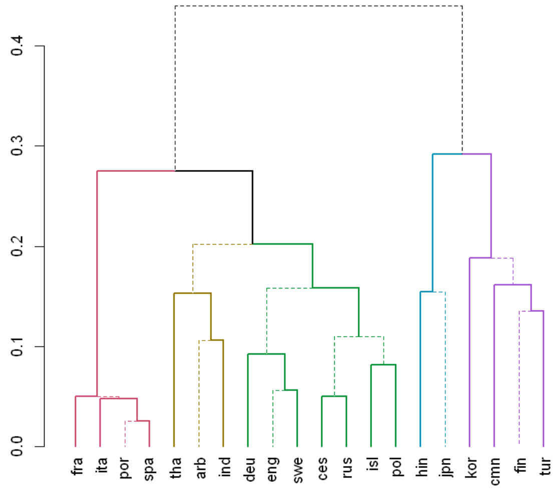


Figure 3: Head and Dependent relative position Clustering Dendrogram

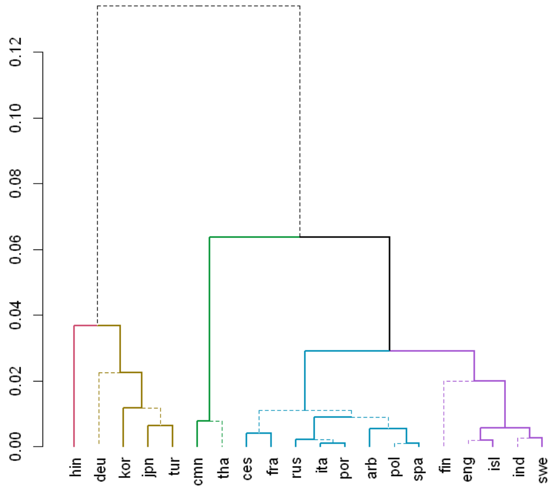


Figure 4: Verb and Object relative position Clustering Dendrogram

sification of PUD languages, it is possible to see that the proximity between Spanish and Portuguese and their relation to French and Italian is also present when the head and dependent orderings are examined. Icelandic is genealogically closer to Swedish, however, in terms of head directionality it is closer to Slavic languages, this classification is closer to the one proposed by (Hawkins, 1983): Icelandic, Czech, and Russian are all considered as type 10. Nevertheless, still according to (Hawkins, 1983), Indonesian and Thai are from the same language type as Romance languages (type 9), but in these dendrograms, although these two languages are grouped together, they are not classed among Romance ones. Moreover, although not genealogically related, the syntactic proximity between Finnish and Turkish is similarly attested with the head directionality analysis.

As expected, when VO and OV patterns are used to generate a dendrogram (4), there is a clear split of PUD languages into two clusters: one contains all OV languages and German (with no dominant order, according to WALS database), and the other, all the VO languages. When analyzing VO languages in detail, it is noticeable that French and Czech are closer in the Verb and Object relative position dendrogram. Finnish is placed together with Germanic languages (except for German) and Indonesian. Slavic languages (except for Czech) are clustered with Romance languages (except for French) in a sub-group that also contains Arabic. The Thai language forms a small sub-cluster with Chinese.

As not only nominal objects are considered for the construction of this dendrogram, it also provides also insights into how other types of objects are ordered (e.g.: pronominal). Thus, this classification cannot be compared to the one provided by (Hawkins, 1983) where only nominal objects were analyzed.

The overall analysis of all obtained dendrograms shows that both lang2vec and head and dependent position figures have more similarities to the classical genealogical classification of languages. Marsagram dendrogram clearly presents a specific typological classification that considers word-order phenomena not contemplated by the other analysis. The verb and object classification provides a particular typological overview that can be interesting for studying focusing on how these two elements are positioned.

In comparison with the language types proposed by (Hawkins, 1983), the typological classifications presented in this article present the advantage of allowing languages to be compared in terms of a larger number of word-order structures, thus, being more precise for NLP applications where the objective is to find the closest languages. For example, as previously mentioned, Indonesian and Thai are classified as type 9 by (Hawkins, 1983), the same group as the PUD Romance languages. However, using the described quantitative methods it is possible to determine how close these two languages are to the Romance ones in a more detailed way.

5 Conclusion and Perspectives

In this paper, we presented three new typological approaches regarding word-order phenomena applied to 20 different languages using parallel corpora. The new methods were compared to the standard one which considers syntactic features provided by a typological database (lang2vec).

Each approach provided a syntactic typological classification of languages in the format of a dendrogram which was obtained via dissimilarity matrices composed of Euclidean distances between language vectors.

We showed that each different approach has its own particularities. The aim of this study was not to state which is the best typological method but to show in which way they provide different angles for typological analysis. However, it is possible to notice that the lang2vec and the Head and Dependent relative position dendrograms are more coherent with the genealogical classification of languages. The Marsagram approach provides interesting aspects regarding specific word-order phenomena of elements that are not syntactically related, while the Verb and Object relative position approach provides a specific analysis of all attested phenomena regarding these elements.

The usability of each method depends on which particular syntactic features are of interest and the purpose of further linguistic processing. Preliminary experiments showed that the language distances obtained using the described quantitative typological methods present moderate or strong correlations with the improvement of dependency parsing results when different languages are combined to train deep-learning models. Thus, in the future, we aim to analyze precisely how each method provides valuable information concerning the improvement of the dependency parsing results to determine the best corpus-based typological strategy for this aim.

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

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).

Sabine Buchholz and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the tenth conference on computational natural language learning (CoNLL-X)*, pages 149–164.

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

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.

Maciej Eder. 2017. Visualization in stylometry: cluster analysis using networks. *Digital Scholarship in the Humanities*, 32(1):50–64.

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.

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.

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.

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

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The conll 2007 shared task on dependency parsing. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 915–932.

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.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. 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.

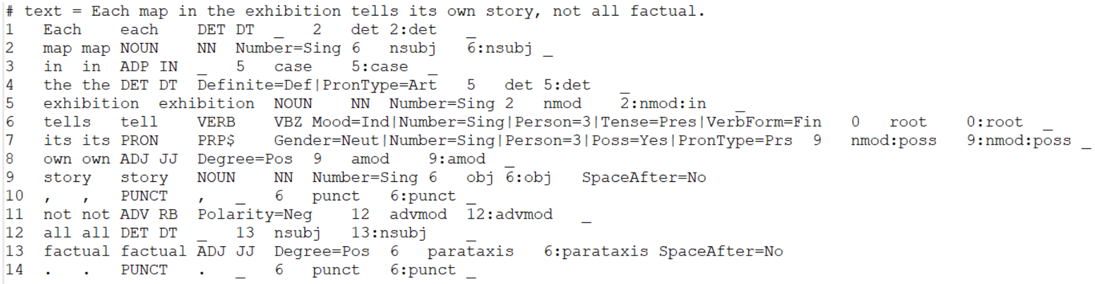
Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

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

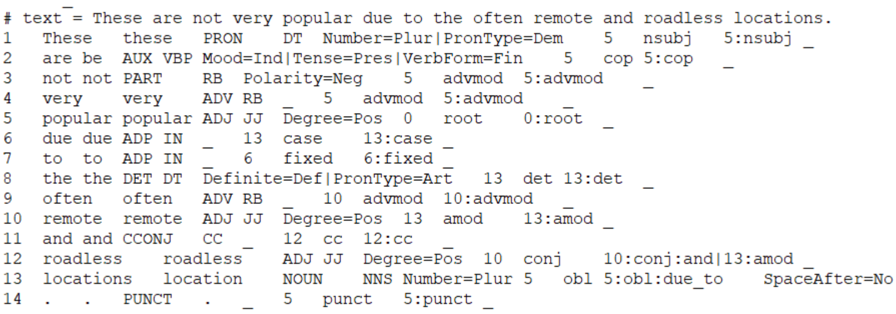
Daniel Zeman, Jan Hajic, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. Conll 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal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

A Appendix



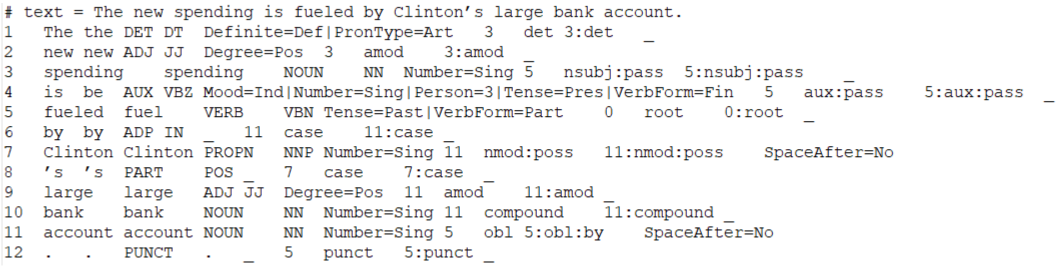
```
# text = Each map in the exhibition tells its own story, not all factual.
1   Each    each    DET DT  _   2   det 2:det   _
2   map map NOUN    NN  Number=Sing 6   nsubj   6:nsubj _
3   in  in  ADP IN  _   5   case    5:case  _
4   the the DET DT  Definite=Def|PronType=Art   5   det 5:det   _
5   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 5: Example of a sentence with the pattern NOUN_precede_DET-det_NOUN-nmod. 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.



```
# 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 6: 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 7: 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.