Computational typological analysis of syntactic structures in European languages

Válio Antunes Alves, Diego Fernando

Doctoral thesis / Disertacija

2023

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Humanities and Social Sciences / Sveučilište u Zagrebu, Filozofski fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:131:663977

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-08-05**

Repository / Repozitorij:

ODRAZ - open repository of the University of Zagreb Faculty of Humanities and Social Sciences

Sveučilište u Zagrebu

Faculty of Humanities and Social Sciences

Diego Fernando Válio Antunes Alves

Computational Typological Analysis of Syntactic Structures in European Languages

DOCTORAL DISSERTATION

Zagreb, 2023

Sveučilište u Zagrebu

Faculty of Humanities and Social Sciences

Diego Fernando Válio Antunes Alves

Computational Typological Analysis of Syntactic Structures in European Languages

DOCTORAL DISSERTATION

Supervisors:

dr. sc. Božo Bekavac, Assistant Professor

RNDr. Daniel Zeman, Ph.D., Senior Researcher

Zagreb, 2023

Sveučilište u Zagrebu

Filozofski Fakultet

Diego Fernando Válio Antunes Alves

Računalna Tipološka Analiza Sintaktičkih Struktura u Europskim Jezicima

DOKTORSKI RAD

Mentori:
doc. dr. sc. Božo Bekavac
RNDr. Daniel Zeman, Ph.D.

Zagreb, 2023.

Dr. Sc. Božo Bekavac

Prof. Ph.D. Božo Bekavac is an assistant professor at the Department of Linguistics at the Faculty of Humanities and Social Sciences, University of Zagreb. In April 1997, he obtained a B.A. degree in General Linguistics and Information science at the Faculty of Humanities and Social Sciences of the University of Zagreb. He began to work as a research fellow at the Institute of Linguistics, Faculty of Philosophy in Zagreb in September 1997 on the project Machine Processing of the Croatian Language. He received his Ph.D. degree (2005) at the Faculty of Humanities and Social Sciences, University of Zagreb with the dissertation "Automatic Named Entities Recognition in Croatian Texts" and became assistant professor at this faculty in 2007.

His scientific work focuses on linguistics formalism, computational tools and language corpora for processing of the Croatian language. He participated in several international and Croatian conferences and he have around twenty published papers in the fields of Computational Linguistics, Linguistic Language Modelling, Corpus Linguistics, Named Entity Recognition and Classification (NERC), Linguistic tools and Mark-up languages. He participated as a research scientist in several nationally and EU funded projects (e.g.: ACCURAT, XLike, LetsMT, and CESAR).

RNDr. Daniel Zeman, Ph.D.

Prof. RNDr. Daniel Zeman, Ph.D., is currently a professor of Morphological and Syntactic Analysis and Computers and Natural Language at Faculty of Mathematics and Physics of the Charles University in Prague. At this faculty, he obtained in 1997 his MSc. degree in Computer Science, and, in 2005, he earned his Ph.D. degree in Mathematical Linguistics and the RNDr. Title, with the thesis entitled “Parsing with a Statistical Dependency Model”. In 2006, he received the award “Fullbright-Masaryk Fellowship” from the University of Maryland. And, between 2014 and 2016, he was a member of the Scientific Council of the Czech National Corpus project.

His research interests are statistical parsing of Czech, dependency modelling, morphological analysis, machine translation, and low-resourced languages. He is one of the creators of the Universal Dependencies framework and has helped in the organization of different shared-tasks involving dependency syntax (e.g.: Conference on Computational Natural Language Learning shared task, 2018). Furthermore, he participated as a research scientist in some funded projects (e.g.: GAAV, CZECHMATE, and MANYLA).

Acknowledgments

The work presented in this thesis would not be possible without the constant and sincere support of many incredible and inspiring people.

First of all, I would like to express my deepest gratitude to my mentor prof. Ph.D. Božo Bekavac and my co-mentor RNDr. Daniel Zeman, Ph.D. Their guidance throughout the process of writing this thesis was crucial for the development of this work. Their valuable feedbacks I received allowed me to finetune my research for a better understanding about the linguistic phenomena described here.

Secondly, I could not have undertaken this journey without the help of prof. Ph.D. Marko Tadić who gave me the opportunity to work at the Institute of Linguistics for the CLEOPATRA project. I am extremely grateful for his support and for his constructive guidance throughout these years in all the activities related to my Ph.D.

Moreover, I would like to express my gratitude to the other members of the commission for the Ph.D. defense, prof. Ph.D. Jan Šnajder and prof. Ph.D. Ranko Matasović, regarding the helpful and valuable comments. I am also grateful to prof. Ph.D. Ida Raffaelli for her help with the administrative procedures and for all the treasured conversations with valuable advices that helped me conducting a better research.

I sincerely thank Ph.D. Matea Filko for all the assistance provided to me since my arrival in Croatia. I am also thankful to Vanja Štefanac, Daša Farkaš and all the other colleagues from the Faculty of Humanities and Social Sciences of Zagreb. Special thanks to Ph.D. Gaurish Thakkar with whom I had the privilege to work as a member of the CLEOPATRA project.

I would like to extend my sincere thanks to all my friends all over the world who were always by my side (in person or virtually) during this endeavour. I dedicate a special thanks to Sérgio Mascate Pires whose moral support helped me a lot in the process of writing this thesis.

Lastly, I would be remiss in not mentioning all the support I received from my family, specially from my mother, sister, aunts, and grandmother Eneid. Thank you from the bottom of my heart for always believing in me and being there for me.

The work presented in this thesis 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.

Abstract

This thesis aims to propose new corpus-based syntactic typological methods for the extraction of syntactic features from annotated corpora (first hypothesis) and to examine the potential of these quantitative strategies for dependency parsing improvement via corpora association (second hypothesis).

In the first part, we presented the obtained corpus-based typological classifications of the 20 languages present in the Parallel Universal Dependencies collection and compared them to the classic phylogenetic classification and the typological ones built with syntactic classification provided by typological databases. We showed that although the corpus-based approaches present results with some similarities with the standard ones, each method provides a classification from a different angle, thus, allowing languages to be classified differently.

In the second part, we examined the improvement in terms of dependency parsing results obtained with the UDify tool when models were trained with two different languages in comparison with the monolingual models. Then, these results were correlated with the different typological approaches to identify the most efficient strategies to select the best language-pairs for dependency parsing improvement.

Additionally, in the third part, we applied the selected corpus-based methods to all 24 European-Union languages with corpora provided by the Universal Dependencies collection and analysed the obtained classifications. Furthermore, we conducted experiments to improve the parsing results for 4 European Union low-resourced languages and Croatian. Maltese and Hungarian were the languages with the best significant improvement for both parsing scores, showing the potential of the strategies for the languages with small training-sets, while Croatian and Lithuanian presented a positive delta for only one of the evaluation metrics. Irish, on the other hand, did not present any improvement. We showed that from the selected typological methods, MarsaGram linear patterns (cosine) and MarsaGram all properties were the ones that generated the best improvements and that longer sentences tend to benefit the most in terms of parsing enhancement when languages are combined.

Extended Abstract

The objective of this thesis was to propose new corpus-based syntactic typological methods characterized by the extraction of syntactic features from annotated corpora and examine the potential of these quantitative strategies for dependency parsing improvement via corpora association.

For this aim, we first analysed, based on typological theories, different ways of mining syntactic information from parallel corpora of 20 worldwide languages, then, we proceeded with the dependency parsing experiments to quantify the synergy obtained when languages were combined in pairs in terms of parsing evaluation metrics. Afterwards, we checked how well each classification, built with the quantitative typological methods, correlated with the parsing results which enabled us to identify the most optimized strategies for dependency parsing improvement. With the selected typological methods, we extended the analysis to all European Union languages by proposing a detailed corpus-based typological syntactic characterization of them. Finally, we conducted a series of parsing experiments with 4 low-resourced European Union languages and Croatian.

Our first hypothesis is that new ways of classifying languages can be achieved by determining the syntactic typological distance between languages using statistical information obtained from annotated corpora. Thus, in the first part of this thesis, we presented the four potential corpus-based typological methods which were considered for the multilingual syntactic characterization.

Two methods are based on the syntactic patterns extracted using the MarsaGram tool which is able to identify and quantify the patterns by composing context-free grammar from texts annotated with part-of-speech and dependency parsing information. One method considers all MarsaGram properties (i.e.: linear, exclude, require, and unicity), and the other one, just the linear property which takes into consideration the word order of elements inside syntactic subtrees. The third method quantifies the features regarding the relative position of heads and dependents, and the fourth one consists of the analysis of the attested verb and object positions in the corpora. For each method, the identified features and their frequency were used to build language vectors which were compared via Euclidean and cosine distances, thus, generating dissimilarity matrices which were the base for creating dendrograms.

The obtained language classifications of the 20 languages present in the Parallel Universal Dependencies collection were compared to the classic phylogenetic one and to the typological classification built with syntactic features provided by typological databases. We showed that although the corpus-based approaches present results with some similarities in comparison with the standard ones, each method provides a classification from a different angle of syntactic perspective, thus, allowing languages to be classified differently.

The second part of this thesis was dedicated to the analysis of the synergy in terms of dependency parsing improvement when corpora from 2 different languages are combined, and comparison of the results with the language classifications obtained in the first step. For this aim, we used the UDify software which is a deep-learning tool built with multilingual BERT language model.

The second hypothesis of this thesis is that typological classifications measuring quantitative syntactic typological distance between languages are efficient in identifying typologically similar languages whose corpora can be combined to improve the performance of deep learning tools in terms of automatic syntactic annotation. Thus, we first defined the baseline of our experiments which consisted of the dependency parsing evaluation scores obtained with models trained with monolingual training-sets. We observed that although the corpora were parallel, the parsing scores varied considerably, being correlated with the size of the language representation in the multilingual language model (mBERT). When languages were combined, we observed that some of them are more prone to present positive deltas and that the obtained improvements were not directly linked to proximity in terms of genealogical features.

Moreover, when these results were compared with the language-distances from the dissimilarity matrices obtained with the different corpus-based approaches, it was possible to identify that the strategy based on the MarsaGram linear patterns (cosine distance) was the one with the highest number of moderate or strong correlation. Additionally, we analysed the potential of each method to select the best language-pair for dependency parsing improvement and found out that three different strategies provided the most promising results: the one considering all MarsaGram properties (cosine), the one based on MarsaGram linear patterns (cosine), and a specific combination obtained via linear regression of the MarsGram all properties and the head and dependent position methods (both with Euclidean distances). Besides them, the verb and object relative position method was also selected for being the one proposing the highest number of right-choices when the other methods failed to do so. The

standard syntactic typological classification presented slightly better scores for a specific dependency parsing evaluation metric, however, its usage was not possible for the analysis of all European Union languages as not all of them had enough information in the typological databases.

In the third part, we applied the selected corpus-based methods to all 24 European-Union languages with corpora provided by the Universal Dependencies collection. We could verify that the usage of non-parallel corpora did not have an impact on the methods and were able to present a syntactic typological description of these languages together with 10 other worldwide ones. Moreover, by analysing the availability of annotated corpora and the literature results in terms of dependency parsing, we could identify that Hungarian, Irish, Lithuanian, and Maltese are the languages with the lowest resources in our language-set. Thus, we conducted a series of experiments regarding corpora-combination to improve their parsing results. Besides them, we also decided to analyse the potential of the typological methods for Croatian which has relatively more resources than these 4 languages, but still presents some room for improvement.

We showed that from the selected typological methods, MarsaGram linear patterns (cosine) and MarsaGram all properties (cosine) were the ones allowing the best improvements. Moreover, from the analysed languages, Hungarian and Maltese (i.e.: the ones with the smallest training-corpora) presented the best positive deltas for both dependency parsing metrics. Irish did not show any improvement. Furthermore, we also noticed that language combinations improve especially results concerning most complex sentences.

Therefore, we showed that the corpus-based methods can bring new light to the syntactic typological analysis of languages and that these strategies are useful for the improvement of dependency parsing results for low-resourced languages.

Extended Abstract (Croatian)

Cilj je ove disertacije predložiti nove korpusno utemeljene sintaktičke tipološke metode koje karakterizira izvlačenje sintaktičkih značajki iz obilježenoga korpusa, te istražiti mogućnosti takvih kvantitativnih strategija za poboljšanje rezultata ovisnosnoga parsanja s pomoću kombiniranja korpusnih podataka.

U tu smo svrhu prvo, na temelju tipoloških teorija, analizirali različite načine izvlačenja sintaktičkih informacija iz usporednih korpusa 20 jezika svijeta, a potom proveli eksperimente s ovisnosnim parsanjem kako bismo, korištenjem metrika za evaluaciju parsanja, kvantificirali sinergiju dobivenu kombiniranjem jezičnih parova. Nakon toga smo provjerili kako svaka od klasifikacija temeljena na kvantitativnim tipološkim metodama korelira s rezultatom evaluacije parsanja, na temelju čega smo pronašli optimalne strategije za poboljšanje ovisnosnoga parsanja. Koristeći se odabranim tipološkim metodama, proširili smo analizu na sve jezike Europske unije pružajući detaljnu tipološku sintaktičku karakterizaciju svakoga od njih. Naposljetku, proveli smo niz eksperimenata s parsanjem nad četirima jezicima Europske unije s malo računalnih podatkovnih izvora i hrvatskim.

Prva nam je hipoteza da se određivanjem sintaktičke tipološke udaljenosti među jezicima koristeći se statističkim podacima iz označenih korpusa mogu iznaći novi načini klasifikacije jezika. Slijedom toga, u prvom smo dijelu ove disertacije predstavili četiri potencijalne tipološke metode temeljene na korpusu za višejezičnu sintaktičku karakterizaciju.

Dvije se metode temelje na sintaktičkim obrascima izvučenima s pomoću alata MarsaGram. Taj alat iz tekstova u kojima su označene vrste riječi i ovisnosna sintaksa izvlači i kvantificira sintaktičke obrasce koristeći se beskontekstnim gramatikama. Jedna metoda uzima u obzir sve značajke koje MarsaGram izvlači (linear, exclude, require i unicity) (metoda MARSAGRAM SVE), dok druga gleda samo značajku linear koja opisuje red riječi u sintaktičkom podstablu (metoda MARSAGRAM LINEAR). Treća metoda kvantificira značajke prema relativnoj poziciji glava i dependenata (metoda GLAVA-DEPENDENT), dok četvrtu sačinjava analiza pozicija glagola i objekta potvrđenih u korpusu (metoda GLAGOL-OBJEKT). Identificirane su značajke i njihova frekvencija korištene za izradu vektora koji su potom uspoređivani s pomoću euklidske i kosinusne udaljenosti, tj. izradom matrice različitosti na osnovu kojih su izrađeni dendrogrami.

Dobivene klasifikacije 20 jezika iz zbirke Parallel Universal Dependencies uspoređene su s klasičnom filogenetskom klasifikacijom te s tipološkom klasifikacijom izgrađenom s pomoću sintaktičkih značajki iz tipoloških baza podataka. Pokazali smo da, iako korpusno utemeljeni pristupi daju rezultate koji su usporedivi sa standardnim pristupima, svaka metoda pruža mogućnost za klasifikaciju jezika iz malo drukčije sintaktičke perspektive.

Drugi dio ove disertacije posvećen je analizi sinergije koja se očituje u poboljšanju rezultata ovisnosnoga parsanja kada se kombiniraju korpusi dvaju jezika te usporedbi tih rezultata s klasifikacijama jezika dobivenima u prvome dijelu. U tu je svrhu korišten UDify softver, alat baziran na dubokom strojnom učenju s višejezičnim BERT jezičnim modelom.

Druga je hipoteza ove disertacije da su tipološke klasifikacije nastale mjerenjem kvantitativne sintaktičke tipološke udaljenosti među jezicima efikasan način identifikacije tipološki sličnih jezika čiji se korpusi mogu kombinirati u svrhu poboljšanja rezultata automatske sintaktičke anotacije provedene alatom dubokog strojnog učenja. Prvo smo od rezultata evaluacije ovisnosnoga parsanja s pomoću modela treniranih na jednojezičnim skupovima za učenje definirali referentne vrijednosti naših eksperimenata. Uočili smo da su, iako su korpusi usporedni, rezultati prilično varirali i korelirali s veličinom reprezentacije jezika u višejezičnom modelu mBERT. Kad smo jezike kombinirali, uočili smo da su neki od njih skloniji uzrokovati poboljšanje rezultata, a ta se poboljšanja ne mogu objasniti blizinom u smislu genealoških značajki.

Nadalje, kad se ti rezultati usporede s vrijednostima udaljenosti iz matrice različitosti dobivene s pomoću različitih metoda baziranih na korpusu, može se uočiti da je strategija bazirana na linearnim obrascima dobivenim iz MarsaGrama (kosinusna udaljenost) ona s najviše umjerenih i jakih korelacija. Uz to, analizirali smo potencijal svake od metoda da odabere najbolji jezični par za poboljšanje rezultata ovisnosnoga parsanja i zaključili da tri strategije donose najbolje rezultate: ona koja uzima u obzir sve značajke MarsaGrama (kosinusna udaljenost), ona koja uzima u obzir samo linearne obrasce (kosinusna udaljenost) te specifična strategija dobivena linearnom regresijom koja kombinira metode MARSAGRAM SVE i GLAVA-DEPENDENT (euklidska udaljenost). Osim njih, prepoznali smo i metodu GLAGOL-OBJEKT kao onu koja daje najveći broj točnih odabira jezičnih parova kad druge metode podbace. Standardna sintaktička tipološka klasifikacija davala je nešto bolje rezultate za određene metrike evaluacije ovisnosnoga parsanja, no njome se nismo mogli koristiti za analizu svih jezika Europske unije s obzirom na to da za neke od njih nema dovoljno podataka u tipološkim bazama podataka.

U trećem smo dijelu primijenili odabrane metode na sva 24 jezika Europske unije na korpusima iz zbirke Universal Dependencies. Utvrdili smo da upotreba neusporednih korpusa nije negativno utjecala na metode i predstavili smo sintaktički tipološki opis tih jezika s 10 drugih svjetskih jezika. Nadalje, analizirajući dostupnost označenih korpusa i postojanja znanstvene literature na temu ovisnosnoga parsanja, zaključili smo da su mađarski, irski, litavski i malteški jezici s najmanje računalnih podatkovnih izvora u našem uzorku. Stoga smo proveli niz eksperimenata s ciljem kombiniranja korpusa u svrhu poboljšanja rezultata ovisnosnoga parsanja. Osim toga, analizirali smo korisnost primjene predstavljenih metoda na hrvatski, za koji bi se, iako ima više računalnih podatkovnih izvora od spomenutih četiriju jezika, rezultati automatskoga parsanja mogli značajno poboljšati.

Pokazali smo da su od odabranih četiriju tipoloških metoda metode MARSAGRAM LINEAR i MARSAGRAM SVE one koje su dovele do najvećeg poboljšanja rezultata. Od svih analiziranih jezika za mađarski i malteški (dakle, za jezike s najmanjim korpusima za učenje) pokazala su se najveća poboljšanja rezultata parsanja u objema evaluacijskim metrikama. Na primjeru irskog nisu se pokazala značajna poboljšanja. Također smo primijetili da kombiniranje jezika najviše poboljšava rezultate parsanja na složenim rečenicama.

Naposljetku, pokazali smo da korpusno utemeljene metode mogu dati novu perspektivu sintaktičkoj tipološkoj analizi jezika te da su te metode korisne za poboljšanje rezultata automatskoga ovisnosnog parsanja jezika s malo računalnih podatkovnih izvora.

Key words

typology, corpus-based typology, multilingualism, dependency syntax, dependency parsing, deep-learning, low-resourced languages

List of Abbreviations

A/Adj	Adjectives
AP	Adjectival Phrase
APiCS	Atlas of Pidgin and Creole Language Structures
ASJP	Automated Similarity Judgement Program
BDT	Branching-Direction Theory (BDT)
BERT	Bidirectional Encoder Representations from Transformers
CCH	Cross-Category Harmony
CFG	Context-Free Grammar
CNP	Common Noun Phrases
CoNLL	Conference on Computational Natural Language Learning
Dem	Determiner
Deprel/DEPREL	Dependency Relation
DMorphS	Deep Morphological Structure
DNP	Determined Noun Phrases
DP	Dissimilation Principle
DSyntRel	Deep Syntatic Relation
DSyntS	Deep Syntactic Structure
EU	European Union
FDG	Functional Generative Description
FEATS	Morphological features
G/Gen	Genitive
HDT	Head-Dependent Theory
HSP	Heaviness Serialization Principle

ID/id	Word Index
IDS	Intercontinental Dictionary Series
ISO	International Organization for Standardization
L	Language
LAPSyD	Lyon-Albuquerque Phonological Systems Database
LAS	Labelled Attachment Score
LFG	Lexical Function Grammar
LHS	Left-Hand Side
LSTM	Long Short-Term Memory
mBERT	Multililngual Bidirectional Encoder Representations from Transformers
MHIP	Mobility and Heaviness Interaction Principle
MLAS	Morphology-Aware Labelled Attachment Score
MLP	Multi-Layered Perceptron
Morph-D	Morphological Dependency
MP	Mobility Principle
MSD	Morphosyntactic description
MTT	Meaning-Text Theory
NLP	Natural Language Processing
NP	Nominal Phrase
NSP	Natural Serialization Principle
Num	Numeral
O	Object
ORTOLANG	Outils et Ressources pour un Traitement Optimisé de la Langue
POS	Part-of-Speech

Postp/Po	Postposition
PP	Prepositional Phrase
Prep/Pr	Preposition
PrNMH	Prepositional Noun Modifier Hierarchy
PUD	Parallel Universal Dependencies
Rel	Relative Clause
RHS	Right-Hand Side
RRG	Role-and-Referencial Grammar
S	Subject
SemS	Semantic Structure
SSWL	Syntactic Structures of the World’s Languages
SSyntRel	Surface Syntactic Relation
SSyntS	Surface Syntactic Structure
SUD	Surface-Syntax Universal Dependencies
Synt-D	Syntactic Dependency
SyntS	Syntactic Structure
UAS	Unlabelled Attachment Score
UD	Universal Dependencies
UPOS	Universal Part-of-Speech
V	Verb
V-1	Verb-first sentences
VP	Verbal Phrase
WALS	World Atlas of Language Structures
WOLD	World Loanword Database

Contents

1. Introduction 1
2. Theoretical background and related work review 5
  2.1. Typology 5
    2.1.1. Historical overview 5
    2.1.2. Main principles 7
  2.2. Syntactic Typology 9
    2.2.1. Greenberg's contribution to syntactic typology 14
    2.2.2. Hawkins' contribution to syntactic typology 16
    2.2.3. Dryer's contribution to syntactic typology 33
  2.3. Corpora-based Typology 37
  2.4. Typology and Natural Languages Processing 42
  2.5. Typology and Dependency Parsing 48
  2.6. Dependency Syntax 55
  2.7. Dependency Parsing 67
    2.7.1. Historical background 67
    2.7.2. Dependency parsing formalisms 68
    2.7.3. Dependency parsers 71
    2.7.4. Dependency parsing evaluation 76
3. Objective and Hypotheses of Research 80
4. Syntactic Typological Classifications 84
  4.1 Language Resources and Tools 84
    4.1.1 Universal Dependencies 84
    4.1.2 Parallel Universal Dependencies (PUD) corpora 94
    4.1.3 URIEL and lang2vec 106
    4.1.4 MarsaGram 108
  4.2 Methods 114
  4.3 Genealogical Classification of PUD languages 118
  4.4 Classification of PUD Languages From lang2vec Syntactic Vectors 122
  4.5 Quantitative Typological Classification Using MarsaGram 129
  4.6 Quantitative Typological Classification Using Head and Dependents Ordering 148
  4.7 Quantitative Typological Classification Using Verb and Object Ordering 159
  4.8 General Discussion 164
5. Dependency Parsing Improvement with Typological Strategies 165

5.1 Tools 165
  5.1.1 UDify 165
  5.1.2 Multilingual BERT (mBERT) 167
5.2 Methodology 169
  5.2.1 Definition of the baseline in terms of parsing results 169
  5.2.2 Language combination experiments using UDify 172
  5.2.3 Typological strategies evaluation in relation to parsing results 173
5.3 Results and Analysis 176
  5.3.1 Definition of the baseline in terms of parsing results 176
  5.3.2 Language combination experiments using UDify 181
  5.3.3 Typological strategies evaluation in relation to parsing results 189
5.4 Overall Discussion 212
6. Typological Analysis and Dependency Parsing Improvement of EU Languages 215
6.1. European Union Languages Characterization 216
6.2. European Union Low-resourced Languages 221
6.3. Corpus-based Typological Classification of EU Languages 225
  6.3.1. MarsaGram linear properties (cosine) typology 232
  6.3.2. Combination of MarsaGram all properties and Head and Dependent strategies (Euclidean) 240
  6.3.3. Verb and Object relative position strategy (cosine) 249
  6.3.4. MarsaGram all properties (cosine) 253
  6.3.5. Discussion and language-pairs selection 257
  6.3.6. Dependency parsing experiments 259
6.4. General Discussion 284
7. Conclusion 288
8. References 300

1. Introduction

Natural language processing (NLP) is a field of linguistics and computer science encompassing a variety of areas that involve computational processing of human languages. Its core sub-areas concern solving fundamental issues such as segmenting text and sentences, morphological processing, part-of-speech identification, syntactic (parsing), and semantic processing.

Many available NLP programs propose Language Processing Chains that encompass most of the sub-areas listed above and are composed of tokenisation, sentence splitting, part-of-speech (POS) and morphosyntactic description (MSD) tagging, lemmatisation, and dependency parsing modules.

Since the 1980s, “the NLP field has increasingly relied on statistics, probability, and machine learning methods which require large amounts of linguistic data. Furthermore, from 2015 onwards, the usage of deep learning techniques has been dominant in this field” (Otter et al., 2019). These approaches require a large amount of annotated data which can be problematic for some languages considered low-resourced.

Linguistic manual annotation of texts can be very costly, especially for tasks requiring specific linguistic knowledge as is the case of dependency relations (Fort et al. 2014). Therefore, “different solutions for improving dependency parsing scores have been proposed involving a great variety of strategies. One way to overcome this issue is to combine data from similar languages according to established typological classifications. However, in general, these studies do not present deep analyses of typological features which may play a significant role when corpora are combined, and do not consider statistics concerning possible (or impossible) syntactic constructions inside the available data as possible typological classifications” (Alves et al., 2021).

An interesting example of “the usage of typological features to improve the results of NLP tools was presented by Üstün et al. (2020). They proposed UDapter, a tool that uses a mix of automatically selected typological features (phonological, morphological, and syntactical) obtained via URIEL language typology database (Littell et al., 2017). Results showed that this strategy was crucial for the improvement of the dependency parsing accuracy for under-resourced languages” (Alves et al., 2021).

Furthermore, Lynn, T. et al. (2014) conducted a series of cross-linguistic experiments with the Irish language concerning automatic syntactic text annotation and showed that Indonesian

(Austronesian linguistic family) presented the best results in terms of parsing Irish texts when compared to Indo-European languages, showing that the usage of phylogenetic classification does not always guarantee the most optimized scores.

As it was the case of UDapter (Üstün et al., 2020), many other studies are based on different typological databases such as WALS (Dryer, M. S. & Haspelmath, M., 2013), PHOIBLE (Moran, S. et al., 2014), Ethnologue (Lewis, M. P. et al., 2015), and Glottolog (Hammarström, H. et al., 2015). The lang2vec tool "provides uniform, consistent and standardized information about languages drawn from the resources listed above" (Little et al., 2017). These studies have shown the potential of language combinations for the improvement of parsing results concerning low-resourced languages. Nevertheless, most studies are conducted with delexicalized corpora (i.e.: data-sets without word-forms or lemmas, only containing part-of-speech or dependency parsing labels) and use systems that do not correspond to the state-of-the-art in terms of dependency parsing.

Furthermore, in the abovementioned databases, syntactic phenomena are usually described in a generic way that considers only the most frequently observed structures. Another disadvantage concerns the lack of sufficient typological information for languages with low resources, thus making it difficult to precisely compare all languages.

Classical syntactic typology regarding word-order usually focuses on universals and correlations with the analysis of specific groups of syntactic features separately. On the other hand, corpus-based typological studies scrutinize in a quantitative way all observed phenomena extracted from annotated corpora. Liu and Xu (2012) proposed a quantitative syntactic typological analysis of Romance languages using information from texts annotated with syntactic information. They have examined the overall distribution of dependency directions which enabled them to correlate it with the degree of inflectional variation of a language and to classify them diachronically (compared to Latin) and synchronically.

A different approach to "extracting and comparing syntactic information from treebanks is proposed by Blache, P. et al. (2016) by inferring context-free grammars (together with its statistics) from syntactic structures inside annotated corpora. Their analysis comparing 10 different languages showed the potential of the proposed tool (MarsaGram) in terms of quantitative typological analysis. However, so far, corpus-based typology has not been deeply examined in terms of its potential for dependency parsing improvement" (Alves et al., 2021).

Overall, it is possible to observe that combining languages is a pertinent strategy to improve natural language processing tools. Nevertheless, researchers tend to follow either classical genealogical classifications or the selection of certain typological aspects, or even just random combinations of languages in their experiments. Obtained results are scarcely investigated in terms of specific correlation to the selected typological features (morphological, syntactical, phonological, etc.). Thus, there is little understanding of the role of different linguistic aspects in language combinations for machine and deep-earning applications.

The concept of Typometrics was introduced by Gerdes et al. (2021). The authors extracted rich details for testing typological implicational universals and also explored new kinds of universals, classified as quantitative. In their study, different word-order phenomena were analysed quantitatively (i.e.: the distribution of their occurrences in annotated corpora) to identify universals (i.e.: presence 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 using all syntactic structures identified in the corpora to obtain a more general syntactic overview of the elements in our language set. Thus, our study differs considerably from the classic typological studies as our main objective is to test how different ways of comparing and classifying languages quantitatively can be applied for the improvement of dependency parsing tools (i.e.: the different approaches are assessed according to the extrinsic evaluation regarding the dependency parsing results).

The goal is to propose new typological classifications of the official European Union (EU) languages by using corpus-based typological strategies based on the quantification of different syntactic phenomena occurring in corpora annotated in terms of syntactic relations. These methods will be the base for corpora associations to improve dependency parsing results for low-resourced EU languages.

Our first hypothesis is that new ways of classifying European languages can be achieved by determining the syntactic typological distance between languages using statistical information obtained from annotated corpora with the identification of syntactic features that have not been considered so far in qualitative typological analysis. The second hypothesis is that the obtained typological classifications can be used to identify related languages whose corpora can be combined to optimize the performance of deep-learning tools in terms of automatic syntactic annotation.

The second section of this thesis describes the theoretical background and related work regarding typology with a specific focus on syntactic typology, corpus-based studies, and how typological strategies have been used in NLP applications. This section also presents an overview of dependency syntax and dependency parsing. In the third section, the objectives of this study are detailed with a precise definition of the research questions and hypothesis. It is followed by the fourth section which presents the different corpus-based approaches regarding syntactic typology and the obtained language classifications (i.e.: the first hypothesis).

The fifth section is a complete analysis of the synergy in terms of dependency parsing results using a deep-learning tool regarding language association. The obtained scores are examined in terms of the correlation with the obtained language classifications in section 4 to determine the most accurate corpus-based methods for parsing improvement (i.e.: the second hypothesis). Additionally, in section 6, we describe a complete analysis of all EU languages and present the language combination experiments for parsing improvement regarding the identified EU low-resourced languages (i.e.: Hungarian, Irish, Lithuanian, Maltese) and Croatian. It is followed by conclusions and perspectives for future work (section 7).

2. Theoretical background and related work review

2.1. Typology

2.1.1. Historical overview

According to Shibatani (2015), typology is nowadays considered as a particular method in linguistic investigation and has evolved from the nineteenth-century attempts to classify languages and to verify variation among them (therefore involving entire languages) via the observation and characterization of specific linguistic phenomena or individual constructions. Usually, restrictions regarding possible dissimilarities are expressed as implicational universals which are deducted from observed distributional patterns of existing and non-existing types. Typological methods are functionally oriented, thus, what is analysed is the typology of the relations between function and form, with the objective of finding external explanation of the observed phenomena.

Typology is generally understood as language classification in terms of structural types. Since its beginning (i.e.: in the beginning of the nineteenth century), the efforts concerning language classification were a way to provide information concerning what are possible variations in human languages. On the other hand, modern typology concerns methods for recognizing and explaining constraints ruling languages and possible grammatical constructions.

It is also usually compared to the genealogical classification of languages as, in the beginning, genetic and comparative approaches were not distinguished from the typological one. It is the case of the work proposed by Friedrich Schlegel who developed a typological framework in his book “Über die Sprache und Weisheit der Indier” (“On the Language and Wisdom of the Indians”) (1808). The author proposes a 2-type classification, attesting that languages can be either flexional (or inflexional) or affixing (or agglutinative), and that the opposition between them represents the two main categories of languages. Other authors from the nineteenth century such as August Wilhelm Schlegel and Wilhelm von Humboldt elaborated Friedrich Schlegel’s classification, focusing mainly on morphological typology.

In the second half of the nineteenth century, the genealogical approach was the mainstream method vis-à-vis the scientific comparative studies of languages, and different languages were seen as different stages of development within the different language families (i.e.: the inflectional type being considered the most evolved and sophisticated stage) (Shibatani, 2015).

Many typological manuals from the nineteenth century divide languages into four types: inflectional (or fusional), agglutinative, isolating, and incorporating (or polysynthetic). The scenario changed in the twentieth century into a more synchronic outlook aligned with the structuralist context. The 4-type classification has deficiencies as languages may present morphological characteristics of different types simultaneously. As an answer to that, Edward Sapir proposed in his book “Language” (1921) a multidimensional typological approach, accommodating gradient characterization of linguistic types in terms of observed tendencies. And Vladimír Skalička, from the Prague School of Typology, introduced later a fifth possible type of language, the introflexive one.

Even if typology has considerably evolved throughout the twentieth century from this initial morphological approach, many developments introduced by classical typologist authors are still relevant for contemporary studies (Shibatani, 2015). For example, August Schleicher documented the possible connections between morphological features and the way in which grammatical relations are expressed. This observation is related to the current point of view which considers that a language type should be defined as a combination of properties showing correlative patterns. Furthermore, the previous evolutionary approach is now seen, in more recent studies of grammaticalization, as cyclic developments (e.g.: Bybee, 1985 and Hagège, 1988).

Since the middle of the twentieth century, typology has been changing its focus, preferring the holistic approach of entire language characterization grounded on a limited set of typological features and where specific phenomena or constructions are examined according to cross-linguistic distributional patterns.

Roman Jakobson is considered by many as the founding father of typology thanks to his works describing the existence dependencies between two elements in language systems (element X in a system implies the existence of Y, but not vice-versa) (Shibatani, 2015). Following Jakobson’s work, Joseph H. Greenberg presented a new perspective in terms of typological studies with a method of universals research. His study on word-order typology (Greenberg, 1963), which will be detailed later in this section, presents a great number of implicational universals in both morphological and syntactic levels. Thus, Greenberg was able to make extensive predictions based only on several simple basic language features.

Typology contrasts with the Chomskyan generative approach which deals with the abstract formal skeleton of syntactic structures independently of semantic or functional factors explaining it in terms of the system internal categories (Chomsky, 1976). The typological approach is functional and assumes that the main function of the language is to represent cognition and to communicate. Therefore, the objective of typology is to determine the range of variation that a language can exhibit when this central function is achieved. The term "functional typology" has been adopted lately to emphasize this aspect of current studies (Shibatani, 2015).

Greenberg also contributed to bringing typology to the centre of diachronic studies by defining what he calls the "state-process model" which considers attested language types as possible linguistic stages which can be attained by human languages, whilst non-attested types are unattainable stages (Greenberg 1969 and 1995). In Greenberg's approach of diachronic typology, typological and transitional states are described by two independent factors: stability and frequency. The first one represents the probability for a language in a specific state to move to another one, and the second factor is the probability that a language will reach a certain state. Hence, the state-process model can be described as a probabilistic one, considering the gradual characteristic of the conversion from one state to another.

This gradual aspect regarding language changes has an impact on the synchronic account of a language as, synchronically, many observed linguistic phenomena are in a transition phase, thus, questioning the usual division between purely synchronic studies and diachronic ones.

2.1.2. Main principles

In the book "Introducing Language Typology" (2012), Moravcsik says that the scope of typology can be defined by the quest for answers concerning how languages differ from each other, and what are the possible explanations for the encountered differences and similitudes. The first question concerns, more specifically, the distribution of linguistic features among languages, whereas the second deals with the elucidation of the distributional observations (Moravcsik, 2012).

Languages can be similar due to a shared historical origin (genetic relatedness). For example, the lexical similarities observed when German, English, and Swedish are compared as they are all derived from the Proto-Germanic language. In addition, another possible explanation of similarities is language contact. And, the third reason is shared cultural environment (e.g.: vocabulary similarities due to socially-conditioned distinctions which can be observed both in Japanese and in Guugu Yimidhirr, an aboriginal language from Australia).

However, the explanations presented in the previous paragraph are not enough to explain all the encountered similarities. Thus, two additional reasons are needed: types and universals. Languages are considered to belong to a type when similar characteristics are shared (a fact that can or cannot be explained by genealogy, contact or shared environment). It is the case of classification of languages in terms of the linear relative position of the subject (S), verb (V) and object (O) in utterances. For example, Hindi and Turkish are unrelated but they are both "SOV" languages. On the other hand, universals are similarities shared by all languages (e.g.: the existence of personal pronouns in all known languages).

Typological studies concern both universals and types and confront the possibility of occurrences of determined phenomena with reality. Thus, in typological studies, crosslinguistic states may be existential or universals, as detailed below:

- Existential statements: "In some language, there is X".
- Universal statements: "In all languages, there is X".

Moreover, universal statements can be further differentiated regarding their broadness (unrestricted universals or implicational ones) and modality (absolute or statistical) (Moravcsik, 2012):

- Unrestricted and absolute universals: "In all languages, there is X".
- Implicational universals: "In all languages, if there is Y, there is also X".
- Statistical universals: "In most/in some languages or in a defined % of languages, there is X".

Moravcsik (2012) details that implicational universals also present some variation vis-à-vis the relationship between the terms based on which they are defined and their complexity. Thus, implications can be classified in 5 different types:

- Paradigmatic implications: "In all languages, if there is Y, there is also X, where Y and X are different constructions".
- Syntagmatic implications: "In all languages, if there is Y, there is also X, where Y and X are parts of the same constructions".
- Reflexive implications: "In all languages, if there is Y, there is also X, where Y and X are features of the same constituent within a construction".
- Single implicans and/or implicatum: "In all languages, if there is Y, there is also X".
- Complex implicans and/or implicatum: "In all languages, if there is Y (and/or W), there is also X (and/or Z).

The main objective of typologists is to find generalities that can be observed either in all human languages, or for the majority of them, or even for a specific subset of them. This is done by proposing and testing these language-universal statements.

Therefore, the typological scope concerns not only existing languages but also those who have disappeared as well as future languages. However, it is evidently impossible to consider every single human language when conducting typological studies. Daniel Nettle (2000) estimated that around 230,000 languages have disappeared throughout human history, and according to Bakker (2011), from the around 7,000 languages that exist today in the world, only one third is well described.

The amount of information available for each language varies a lot, thus, limiting the type of possible typological analysis. Consequently, typologists frequently deal with selected and well-defined samples of languages (Moravcsik, 2012). As the main objective of typology is to find resemblances and variances among languages which are independent, the chosen sub-set of selected languages should, at least, represent all language families, all geographical areas, and cultures. It is the case of the study published by Matthew Dryer (1989) concerning universal word-order tendencies: all languages were assigned to genetic groups (genera) and each genus assigned to a geographical area (one of the five continental areas of the world). Thus, a pattern was considered a significant universal inclination if, in all areas, it was present in the majority of genera.

As it is impossible to examine all existing human languages (and the ones that have existed throughout human history), universal statements are, therefore, merely hypothetical and can never be totally attested. Moravcsik (2012) defines them as best-possible guesses, as it involves extrapolations from what is known from some languages.

2.2. Syntactic Typology

To express a specific meaning, languages vary in terms of choice of words, word forms and order of words. Syntax concerns all of these aspects. Thus, syntactic typology involves identifying, and possibly explaining, observed crosslinguistic differences and similarities within syntactic structures.

Considering the choice of words, variations can be overserved in how available word categories are distributed across sentence-types, and across languages, whereas, regarding word forms, what is important is the grammatical agreement which implies government of one term over another. And, finally, concerning the word order, variations reflect how different terms are positioned inside the sentences or phrases.

In the book “Introducing Language Typology”, Moravcsik (2012) presents several examples of typological differences concerning syntax. Concerning the choice of words, the author mentions the existence or non-appearance of copula, phenomenon which can vary both inside the language internal distribution (e.g.: in Hungarian, the copula is only observed in cases where the subject is first or second person, and never attested in the third person when the verb is in the present tense) or cross-linguistically (e.g.: presence in English and Croatian but absence in Arabic and Russian for sentences in the present tense).

Furthermore, Moravcsik (2012) details cross-linguistic differences and similarities concerning resumptive pronouns and classifiers. About resumptive pronouns, Moravcsik (2013) presents the analysis conducted by Keenan and Comrie (1977) who identified 26 languages which require this kind of pronouns in a set of specific relative clauses. For example, in Persian:

mardi	ke	man	shir-râ	be	u	dadâm
man	that	I	milk(ACC)	to	him	gave:1S

“the man that I gave milk to”

Where ACC stands for accusative and 1S for first-person, singular. In Persian, “u” is a required resumptive pronoun (phenomenon not attested in English). Keenan and Comrie (1977) showed that the occurrence of these pronouns follows a precise pattern across several sorts of relative clauses, and proposed the following scheme concerning an accessibility hierarchy:

Subject > Direct Object > Indirect Object > Oblique Object > Genitive > Object of Comparison

Thus, the tendency for the presence of resumptive pronouns increases for the elements on the right side of the hierarchy. This observation allowed the authors to provide two generalizations:

1) If in a language, a resumptive pronoun is compulsory at any point on the hierarchy, it is also obligatory for the other points to the right.
2) If in a language, a resumptive pronoun is non-compulsory at a point on the hierarchy, it is not obligatory either to points on the left.

The identification of this hierarchical pattern allows languages to be classified in seven different types (varying in what level of the hierarchy the phenomenon starts to be observed). One possible explanation of that is the sparsity of relativization down the mentioned hierarchy (Keenan and Comrie, 1977).

The other example provided by Moravcsik (2012) about choice of words deals with classifiers. Languages differ immensely regarding constructions containing numeral classifiers: if classifiers are present or absent, and in the noun-classes defined by them. Rijkhoff (2002) displayed that there is a great discrepancy in the quantity of classifiers across languages (1 for Cebuano, more than 200 for Vietnamese and Burmese). Numeral classifiers may be compulsory or optional and may also occur with other satellites such as demonstratives.

Still about classifiers, David Gil (2005) noticed that the existence of numeral classifiers is observed in all geographical areas of the Earth (with a specific concentration in East and South-East parts of Asia), however, this sort of word is absent in 260 out of the 400 analysed languages.

Regarding the choice of word forms, syntactic typology deals with the concepts of agreement and government, meaning that a word form may be dependent on a different one of the same sentence. Languages also differ in terms of the presence or absence of different types of agreements, and on how these governing phenomena occur.

Moravcsik (2012) defines that the agreement-target is the constituent that agrees with the one so-called agreement-controller. Additionally, the agreement-features are the properties that are duplicated from the controller by the target-word from (e.g.: person, number, and gender). Languages present variations in terms of type of agreements and agreement-features, however, these differences occur in a controlled way.

One example of generalization that can be stated concerning agreements is the following (Moravcsik, 2012):

In most languages,

a) if the verb agrees with the indirect object, it also agrees with the direct object, and,
b) if the verb agrees with the direct object, it also agrees with the subject

This generalization is observable in most languages but not in all as shown by Siewierska and Bakker (1996). Cross-linguistically preferred verb-agreement follows the simple hierarchy presented by the generalization above which focuses on agreement-controllers. However, an agreement hierarchy can also be determined regarding the agreement-features exhibited by many agreement-targets (Corbett, 2012).

The Agreement Hierarchy:

Attributive > Predicate > Relative Pronoun > Personal Pronoun

Furthermore, Comrie (2005) proposed the following generalization concerning language alignment systems after analysing different languages in that aspect: “In most languages, case marking follows either the accusative or the ergative alignment, with the accusative one being more frequent”.

The other important syntactic aspect which is a common object of study in syntactic typology is word order. In this type of surveys, usually frequencies are mentioned: “Order pattern X is frequent/not frequent across languages”. Also, phenomena regarding word order are commonly defined by sets of implicational universals, thus, the distribution of the different word order clusters relative to each other is not random (Moravcsik, 2012).

When analysing the frequency of linear word order phenomena, the attested possibilities are excrutinated via the examination of the terms within the ordering patterns and the existing relations among them. Moravcsik (2012) presents a list of nine possible cases, the first four linked to the ordering terms and the other five concerning their relations:

1. Classes of words: for example, generalizations concerning adpositions and demonstratives.
2. Classes of phrases and clauses: words inside phrases and clauses follow certain rules but these clusters also have specific positions inside sentences (e.g.: in English, the relative clauses are positioned after the noun that they modify).
3. Individual lexemes: in some sporadic cases, word-order can be defined by individual words. For example, the adposition “ago” in English language which is not preposed as other adpositions (such as “before”), but postposed (e.g.: “two years ago”).
4. Numerical position: for example, the order pattern which requires that some constituents must be situated at the second position of the sentence (known as the Wackernagel’s Law, which is a generalization about the position of enclitics and other small weak emphasized post positive words in the sentence) (Wackernagel, 1892).
5. Precedence regardless of adjacency (“A is placed before B regardless of distance”): it defines a relation of linear precedence between terms (but not necessarily one immediately after the other).

6. Adjacency regardless of precedence (“A is placed next to B regardless of whether before or after”): it concerns terms which may occur in different orders but that are always in an immediate sequence in a sentence.
7. Both precedence and adjacency (“A is place immediately before B”): no universal patterns have been observed regarding immediate precedence.
8. Neither precedence nor adjacency – free order (“A is placed anywhere”): in languages showing this pattern, constituents can appear in all possible orders without any of them being most frequent. Order may be observed but due to other factors such as focalization.
9. Interlocking order (“A is placed between parts of B”): this type of pattern is not frequent, it involves the cases where a phrase is dissembled into two parts surrounding another one. An example is the negation in French (e.g.: “Je ne sais pas”, the negative particles “ne” and “pas” surround the verb “sais”).

One major work concerning the relation types described above (5 to 9) was developed by Dryer (2005) who analysed 1,228 languages regarding the 6 possible orders of subject (S), object(O) and verb (V). The author presented the following observed distribution:

- SOV: 497 languages;
- SVO: 435 languages;
- VSO: 85 languages;
- VOS: 26 languages;
- OVS: 9 languages;
- OSV: 4 languages.

Also, 172 of them did not presented any dominant order (thus, being free-order languages).

From this distribution, it is noticeable that SOV and SVO are the most common ordering patterns (the subject preceding both verb and object), followed by the type VSO (subject preceding object). Dryer (2005) also observed that, in general, the object and the verb occur next to each other (which is related with the Behaghel law proposed by Ruszkowski, 2003: elements presenting semantic relations are located together).

Analysing word order patterns also involves the search of typological implications (conditions that favour one specific ordering pattern over another).

Beside the quantitative distribution analysis described previously, Dryer (2005) also defined generalizations regarding the observed implications, such as:

> "S, V, O and possessive constructions:
>
> a) If a language has OV order, it usually has he possessor before the possessum.
> b) If a language has VO order (other than SVO), it usually has the possessor after the possessum".

These generalizations allow us to identify similarities vis-à-vis the behaviour between different types of constituents. One possible theory to explain these similarities is the proposed division of syntactic constituents into two classes: heads and dependents (Moravcsik, 2012). According to Vennemann (1973), it is the existence of heads and dependents that explains the similar behaviours that can be observed within different constituents, implying that dependents are positioned before heads in OV languages and heads are positioned before dependents in VO languages. Vennemann's work will be further described later in this section.

2.2.1. Greenberg's contribution to syntactic typology

As previously mentioned, Shibatani (2015) considers Greenberg as a pioneer in the way of analysing language universals in word order typology. The impact of Greenberg's work "Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements" (1963) is also recognized by Hawkins in the preface of his book "Word Order Universals" (1983) where he affirms that by expressing word order universals with implicational statements, Greenberg made possible the discovery of other sets of implicational universals, thus, developing a sort of universal grammar (labelled as "Typological Universal Grammar" by Hawkins, 1983). Greenberg's work has influenced research in generative grammar (which started to include setting of parameters on cross-language variation), as well as in descriptive grammar (in terms of properties to be analysed).

Greenberg's paper from 1963, mentioned by Hawkins, is a compilation of observed universals concerning mostly morphemes and word order in a 30-language sample: the languages are from 5 different geographical zones and were selected due to the availability of adequate grammars. The majority of the proposed universals are implicational: "Given X in a particular language, we always find Y" (but not necessarily vice-versa). Greenberg states that propositions which are valid for all 30 languages have a fair likelihood of complete (or almost complete) universal validity.

First, Greenberg (1963) established universals according to the empirical evidence from the selected languages, then, he proceeded with an exploratory analysis of possible generalizations. For him, phenomena concerning word order reveal that some features are closely related to each other, while others are relatively independent. In this way, Greenberg (1963) defined a basic word order typology, considering specifically certain basic factors: the existence of prepositions (Pr) against postpositions (Po); the relative position of subject (S), verb (V) and object (O) in declarative sentences; and the ordering patterns of qualifying adjectives (A) with respect to nouns (N).

Greenberg (1963) observed that, considering the relative order of subject, object and verb, only three combinations normally occur: SVO, SOV, VSO (first Greenberg universal). Therefore, by combining the three word-order features which compose the basic order typology, Greenberg was able to classify the languages in twelve different classes described in the Table 2.1.

	VSO	SVO	SOV
Po-AN	0	1	6
Po-NA	0	2	5
Pr-AN	0	4	0
Pr-NA	6	6	0

Table 2.1. Language distribution according to Greenberg basic order typology (Greenberg, 1963) regarding his 30-language sample.

The genitive (G) order was not considered as a basic factor as it is highly correlated with the position of the adposition (Greenberg, 1963). Using the classification showed in table 1.1, Greenberg (1963) proposed 6 other universals, such as: “Languages with dominant VSO order are always prepositional” (Universal 3).

Following this first set of universals, Greenberg (1963) proposed other ones using the order of nominal subject, nominal object and verb in declarative sentences in relation with the position of interrogative particle or affix in interrogative sentences, and, also, in relation with the place of interrogative words. He also analysed verbal subordination: introductory words (conjunctions), verbal inflections, conditional statements, and expressions of volition and purpose.

Still considering syntactic elements, Greenberg (1963) proposed other universals related to: the position of auxiliary verbs; the position of attributive adjectives in relation to the modified noun; the relative order of demonstratives and numerals; the order of adverbial qualifiers of adjectives in relation to the adjective; the relative order of adjective, comparison marker and standard in comparisons of superiority; the position of common and proper nouns in appositions; the order of relative clauses which modify a noun; and, finally, the ordering of pronominal objects.

In total, Greenberg (1963) established 13 universals considering the syntactic elements presented in the paragraphs above. Following the typological syntactic analysis, he suggested a set of 20 morphological universals considering inflectional and derivational affixes in terms of the existence or absence of these elements, their relative position, and their relation with agreements, number and gender categories, case systems, adjectives, relative order of subject and object, and pronouns.

Greenberg (1963) also introduced two basic notions: dominance of a particular order over its alternative and the harmonic or disharmonic relations among distinct rules of order. In Greenberg's words: "A dominant order may always occur, but its opposite, the recessive, occurs only when a harmonic construction is likewise present" (Greenberg, 1963).

To exemplify these notions, Greenberg (1963) explains in details the universal number 25: "if the pronominal object follows the verb, the nominal object also does". In other words, the nominal object may follow the verb if the pronoun precedes or follows it, meanwhile, the nominal object may precede the verb only if the pronoun also precedes it. Therefore, VO is dominant over OV (as the latter only occurs under specific conditions). Moreover, the order noun object-verb is harmonic with pronoun object-verb but disharmonic with verb-pronoun object. Harmonic and disharmonic relations are examples of generalizations and, in similar constructions, the corresponding members usually follow the same order.

In short, Greenberg (1963) concludes that prepositions, NG, VS, VO, NA are directly or indirectly harmonic with each other. While the opposite type of harmonic relation is: postpositions, GN, SV, OV and AN.

2.2.2. Hawkins' contribution to syntactic typology

Twenty years after the appearance of Greenberg's article, Hawkins published the book "Word Order Universals" (1983) which he described as a major contribution to the descriptive data-driven work proposed by Greenberg (1963), as well as a response to the theoretical problems that have appeared since Greenberg's publication. While Greenberg (1963) focused on the empirical description of universals, Hawkins' intention was to review Greenberg's conclusions, to find explanations for the observed patterns, and to describe the implications of universals within theories of diachronic changes.

Hawkins' language data-set is composed with 350 different languages (an extended version compared to the language ensemble used by Greenberg), thus, with more material, he was able to suggest a substitute ensemble of descriptive word order statements. These propositions were tentatively explained mainly by using formal syntax (mostly generative grammar). Hawkins' book is, therefore, an attempt to join typological universals and formal syntax concepts. Even though focusing on formal syntax, Hawkins (1983) also considered other linguistic aspects for his theoretical approaches concerning word order universals, what can be described as a multifactored method: descriptive universals are a result of many different demands interacting in linguistic systems. Hawkins refused the innateness factor to explain the observed phenomena, only using it when other explanatory principles (e.g.: language function, semantics, language processing, etc.) have been excluded.

Hawkins (1983) also abandoned the Greenbergian trichotomy VSO/SVO/SOV. For him, a type is defined as an ensemble of variant subtypes, and each subtype respects some regularities and shares at least one "typological indicator".

Hawkins starts his work (1983) by presenting his considerations vis-à-vis the goal of finding a theory of word order universals. He acknowledges that languages present differences regarding word order, however, affirming that some clear patterns can be identified. In his opinion, the idea previously suggested that languages have the tendency to place modifying elements (dependents) either constantly before or after modified elements (heads) is too strong and many forecasts resulting from this principle are not correct.

The author's main objective was to explain the difference between the mathematically possible and the really observed word order associations across languages. He observed that a "large amount of languages shares only a small set of word order co-occurrences, moreover, the large majority of possible types related to word ordering patterns are not attested" in his sample (Hawkins, 1983). He defines a set of 6 questions to be answered:

a) "What are the relatively few word order co-occurrences that languages select from among the mathematical possibilities?"

b) “Why do languages select these rather than other possible co-occurrences?”
c) “What are the relative frequencies of languages among the attested word order types?”
d) “Why do the attested word order types exhibit the varying frequencies that they do?”
e) “What can historical principles contribute to the explanation of current synchronic word order variation?”
f) “How and why do languages change from a specific word order type to another?”
g) “What predictions do the synchronic universals of word order make for word order change?”
h) “What use can be made of the synchronic universals of word order in linguistic reconstruction?”

Greenberg’s work (1963) was an attempt to answer question “a” with the proposal of word order universals, while Lehman (1978) and Vennemann (1981) worked on theories to explain these phenomena (question “b”). Questions “c” and “d” concern the synchronic variation while the final 4 questions are related to diachrony.

Regarding the theories of Universal Grammar, Hawkins (1983) distinguishes two major traditions: the theory related with generative grammar, and what he called “typological universal grammar”. The first one is grounded on Chomsky’s work (1965) whose main objective was the extraction of formal and substantive universals[1] of language, thus, developing a universal grammar (of all possible human languages). This generative approach suggests that by bringing into light the characteristics of the grammar of a particular language, it is possible to establish the universals of language in general. Therefore, allowing the obtention of a general and descriptive model which can be used to analyse any human language. In the generative grammar, universals are related to innate properties of human mind.

On the other hand, the so-called “typological universal grammar” (associated to names such as Greenberg, Keenan, Comrie, Thompson) was designed to reflect the existing variations which are observed when different languages are compared. The concept of universals (common properties present in all languages) is not totally rejected, still, the main focus is the search for patterns regarding the manners by which languages vary, and the identification of constraints underlying the attested variations. In conclusion, a language can only be attested if its properties respect the parameters allowed by the variation-defining language universals. These properties outline clusters which allow language classification (typologies). For example, Keenan (1978) defined a set of syntactic and semantic primitive terms which enables the construction of individual language grammars and corresponds to an ensemble of descriptive statements describing attested and non-attested patterns of cross-language variation.

---

[1] Chomsky defines formal universals as those “dealing with the form and shape of the grammars of all languages (components, rule types, principles of rule interaction, etc.) and substantive universals as those concerning the content of rules” (Chomsky, 1965).

For Hawkins (1983), the typological universal grammar presents a series of advantages: it is a descriptive formulation using typological statements considering cross-lingual syntactic and semantic properties and providing explanatory debates (pragmatic, semantic and psycholinguistic) regarding the observed variations.

As previously mentioned, Hawkins' study (1983) is grounded on a sample of 350 languages from all the largest linguistic families of the world. He focused on approximately a dozen word order patterns which were composed by specific pairs of head and modifier. However, only for five of them, data was available for all languages in his dataset: adjective/noun, genitive/noun, adposition/noun phrase, object and verb, and subject/verb. Furthermore, most of Hawkins' universals are from the implication kind (like Greenberg's ones).

Hawkins (1983) acknowledges that not all properties can be compared across all languages (the study being heavily dependent on semantic criteria), and some patterns present variant orders. However, his focus was on "basic word orders[2]" which are easily identified even when some variation is encountered. He also assumed that the selected word categories (subject, object, verb, etc.) are comparable across languages[3].

The cross-lingual approach allowed Hawkins (1983) to attest not possible combinations but also what is impossible in languages. Moreover, with the analysis of a larger language-set, frequency hierarchies were also possible to be defined via distributional universals. These cross-lingual frequencies provided an ensemble of relevant information which enabled the development of a theory of rule complexity (and markedness).

For Hawkins (1983), the main problem regarding the notion of basic word order patterns is "the position of the verb arguments at the sentence level: the predominant opinion being that sentence-level word-order freedom is influenced by pragmatic rules which are sensitive to old versus new information, topic, focus, and so on. However, the two arguments that present a quite fixed ordering (related to the verb) are the subject and the object" (Hawkins, 1983). These arguments are privileged in his study together with the ordering patterns of components of noun phrases and adpositional phrases, and this is due to the fact that the word order freedom is considerably less extensive in these cases.

---

[2] For example: "The basic verb position in English is SVO".

[3] "It does not imply sharing of all the relevant properties for the entity in question by all languages but a sharing of sufficient properties to allow the comparison to be made and, therefore, being recognizable across languages" (Hawkins, 1983).

Another possible bias identified by Hawkins (1983) is the presence of "doubling" which is the case where the same modifier category can be placed both before and after the head. He proposed three overlapping criteria to define which order is the most basic one (basicness decision):

1. "The more frequent doublet is the basic one".
2. "The more frequent doublet in the grammatical system is the basic one".
3. "If one doublet is grammatically unmarked and the other one is marked, the unmarked one is the basic one".

Some residual phenomena can still be problematic concerning basic word order definition. Hawkins (1983) give the example of the existence of both prenominal and postnominal genitives in English and the complexity of object and verb position in German. Nevertheless, these residual cases do not pose any problem to the implicational typological universals as the statements are based on the basic word order and are only valid for languages where they can be identified.

Concerning the classification of languages as VSO/SVO/SOV (Greenberg, 1963), Hawkins (1983) proposed some changes for his approach of typological universals:

1. "SVO is no longer a type indicator, thus, nothing correlates with SVO in a unique and principled way".
2. "VSO and SOV are type indicators, but limited ones".
3. The whole notion of "word order type" becomes more abstract. "The set of languages comprising a common type no longer share all of a given set of word order properties".

The decision of not considering SVO as a type comes from the observation that all the 15 Greenberg's universals regarding verb position involve only VSO or SOV types. Thus, it appears that SVO type does not correlate with other ordering properties (at least, not in a unique and explainable way).

Before presenting his own set of universals, Hawkins (1983) conducted a critic and detailed analysis of the work previously done by Greenberg and Vennemann due to the importance of this literature to the development of his research on synchronic word order universals.

Hawkins (1983) observed that 25 Greenberg's implicational universals involve word order (embodying 34 distinct claims) and that, in around half of these statements, the prior property involves verb position. Also, many further properties (deducted from these verbal universals) are then used to define new universals. Thus, "the relative ordering of the verb became a strong indicator of language types. Greenberg (1963) described three basic positions of the verb: VSO, SVO and SOV" (Hawkins, 1983), nevertheless, other possibilities concerning verb position have been attested since the publication of Greenberg's article. Moreover, as previously mentioned, SVO appears to correspond to a mixed type, thus, the verb position as whole is not a general or useful indicator for typological universals as it seemed to be in Greenberg's study (1963).

Another issue with Greenberg's universals (1963) identified by Hawkins (1983) is that although they are considered as statistical, in reality, only 11 (out of 34) are truly based on frequency. Also, Greenberg's work failed on identifying organizing principles related to his conclusions by relating disparate non-verbal-position word order claims to one-another without providing reasonable explanation for the obtained correlations (Hawkins, 1983). Greenberg's universals present a vast number of exceptions, and some evident patterns were missed. In many cases, Hawkins (1983) attested that more robust claims without attested exceptions could have been made just by reformulating Greenberg's universals.

Hawkins also criticized Greenberg's statistical implications stating that the kind of propositions generally used by Greenberg (i.e.: "P & Q is much more frequent than P & -Q") is not the most suitable way for describing language distributions as it only reflects cases relating high versus low frequencies of co-occurrences, thus, in this way, medium-sized, small and single-proposition types are excluded.

Vennemann's works (e.g.: 1972, 1973, 1974, 1981) were attempts to reformulate and elucidate Greenberg's universals while proposing a link between them and a diachronic theory of word order change. His first innovation was to reduce Greenberg 3-way typology (VSO/SVO/SOV) to two possibilities: VO or OV. Then, inspired by Lehmann's developments (1974) he related all Greenberg's word order properties to them, thus, proposing two major language types (VX and XV).

The main concern in Vennemann's theory was to deliver an organizing principle for Greenberg's universals and propose a theory to explain the observed correlations. Thus, he established categories of operators and operands[4] and serialized them in a consistent order: "OV (XV) languages have co-occurrences with the order operator before operand, while VO (VX) languages have operand before operator" (Vennemann, 1972). Operators and operands were defined as such: "the application of an operator results in a specification of the operand predicate (semantically speaking), while the application of an operator to an operand results in a constituent of the same general category of the operand" (Vennemann, 1972). Based on these concepts, Vennemann formulated his "Natural Serialization Principle" (NSP): "Languages serialize all the operator-operand pairs either operator before operand (OV languages), or operand before operator (VO languages).

Vennemman was aware that not all languages are consistent with the NSP, therefore, he proposed some historical reasons to explain the attested exceptions: inconsistent languages are those which are moving from one type to the other. For Vennemann, the NSP is a "theory of basic word order", and Greenberg's universals can be summarized and explained via this principle. Thus, Greenberg implications were transformed from unilateral and non-reversible to bilateral and reversible statements. It means that, while, in Greenberg's universals, "if P then Q" means that "P & Q", "-P & Q", and "-P & -Q" are possible constructions (only P & -Q being excluded), for Vennemann's implications of the type "P if and only if Q", the only two possible co-occurrences are: "P & Q" and "-P & -Q". Hence, any operator-operand order should ensure the co-occurrence of all the others, thus the position of the verb loses its special significance.

According to Hawkins (1983), despite the fact that Vennemann assumptions were less atomistic and more explanatory than Greenberg's universals, his proposition had lower descriptive power, which is noticeable by the higher number of exceptions emerging from his predictions as they are unable to distinguish "-P & Q" from "P & -Q". Remembering that "-P & Q" co-occurrences are abundantly confirmed by Greenberg's universals, whilst "P & -Q" are not. Furthermore, even though the verb position loses its central status in Vennemann's works (compared to Greenberg's universals), the verb position alone still maintains its implicational antecedent status, for example, in language acquisition and concerning language changes.

---

[4] Examples of operators (category II): Adjective, Relative Clause.
Examples of operand (category II): Noun.

Vennemann equated his operator-operand relation with the logical function-argument theory in his first works. However, Keenan (1979) showed that this association is not reasonable and even Vennemann in later works abandoned this idea. Keenan (1974, 1979) proposed a principle of serialization using a more accurate definition of function and argument, however, a "Dissimilation Principle" (DP) was also required: "Functional expressions taking Determined Noun Phrases (DNP) as arguments and functional expressions taking Common Noun Phrases (CNP) as arguments tend to serialize on the opposite side of their argument categories". This principle was criticized by Hawkins (1983) who preferred to establish a serialization principle explaining all word orders using only a single abstract distinction.

Instead of the correspondence between operators and operands with function and argument, a better connexion can be established with "modifiers" and "heads", which is aligned with X-bar theory of generative grammar (Chomsky 1970, Jackendoff 1977, Lightfoot 1979): operands being the head of phrases (N, V, A, and Prep), while operators are specifiers and complements. Thus, in summary, what Vennemann proposed is that languages order their modifier and head categories in a consistent way: "in OV languages the modifier is placed before the head within all phrasal categories (NP, VP, AP, PP), whereas VO languages locate the head before the modifier" (Vennemman, 1972).

Numerically, when the NSP was tested by Hawkins (1983) with his language-set, 68 languages had no inconsistency, 50 presented 25% of inconsistency, and 24 showed 50% of inconsistency. For Hawkins, the NSP predictions are both too strong (too many exceptions) and too weak (distinctions between attested and non-attested languages not being totally identified), thus, Vennemann's universals are less adequate than Greenberg's (Hawkins, 1983).

In 1976, Vennemann presented a redefinition of the operator-operand relation introducing two types of operators (or specifiers): attributes and complements. These two types are semantically different: attributes (e.g.: adjectives in relation to nouns) are functional categories that preserve the category constancy, while, complements (e.g.: direct object noun phrases in relation to verbs) are arguments of the corresponding functions. In this way, the natural serialization principle could be rewritten with a different terminology: languages consistently "prespecify" if all specifiers (operators) are positioned before the head, and "post-specify" if specifiers are located after it. However, following this principle, SVO languages should behave like VSO languages (post-specifying), and, as it is not the case, Vennemann (1976) simply claimed that SVO languages are inconsistent regarding this specifier.

Later on, in 1981, Vennemann affirmed that a complete consistency should be considered only as typological ideal and not as a universal requirement: languages may or may not achieve the ideal, or may attain it in different stages. This declaration introduced a clear partition between typologies and universals. Two types of universals were, then, defined (Vennemann, 1981):

1) "for all L: A(L)": for all languages (L) the predicate A is attested. This type of universal also permits the construction of implicational ones, such as: "for all L: if B(L), then C(L)", B and C being different predicates. Universals of this type can only occur as part of a general grammatical theory.
2) The second type correspond to Greenberg's statistical universals which are "near-universal", "universal preferences", "unmarked case", "natural case", and "instances of more than chance frequency".

For Vennemann (1981), the second type of universals does not belong to grammatical theory and should be accommodated in a separate one of linguistic preferences. He also suggested that typology should be divided in three types: classificatory, ideal, and graduating. The first one divides all the languages into different typological classes. The ideal typology is consisted of idealizations only: some languages may be partitioned into a finite number of types, while others may match none of the ideal classes, remaining outside the classification. His NSP principle became a part of this "ideal typology". For him, typology is a branch of applied theoretical linguistics and its purpose is purely practical, thus, with an orientative function.

On the other hand, Hawkins' point of view (1983) is not so radical. While agreeing that in some instances typologies and universals are distinct from each other, when implications are defined, mathematically possible language types can indeed be identified, thus, in these specific cases, the formulation of language universals and the definition of language types are the same thing.

One last weakness identified by Hawkins (1983) in Vennemann's later works was that, although being more coherent with function-argument categories, the concept of head was introduced without defining precisely what all heads have in common. Hawkins (1983) argued that the proposed criterion of "category constancy", which was given for defining modifier and head, is not offered by the function-argument theory, thus, leading to an internal contradiction.

Before reformulating Greenberg's word order universals, Hawkins (1983) defined a set of general properties which are necessary for the establishment of implicational universals (avoiding the problems encountered by him in Greenberg's and Vennemann's propositions):

- Implicational universals should preferably be non-statistical (no exception in the language set should exist).
- They must be unilateral ("if P then Q"), thus, defining three-way typologies (i.e.: P & Q, ¬P & Q, and ¬P & ¬Q), not two-way classifications emerging from using bilateral propositions (i.e.: "P if and only if Q") which does not allow ¬P & Q statement.
- An important number of the implicational universals are multi-termed rather than bi-termed (i.e.: defined in terms of at least 3 properties rather than just 2). Multi-termed statements are able to cover a larger range of regularities of co-occurrences when compared to bi-termed ones.

Hawkins (1983) also indicates that, even though the definition of possible and impossible human languages cannot be stated with absolute certainty, it is reasonable to be confident with the fact that, with the set of language information available today, there is sufficient data for extracting the universals of word order variation: the attested language-set has evident frequency around a small and specific common set of co-occurrences, furthermore, patterns which distinguish attested from the unattested co-occurrences were possible to be established, and, lastly, there exists a regular decrease in the amounts of languages possessing some of the many observed phenomena which are regulated by principles.

Moreover, Hawkins (1983) showed that "there is a large inconsistency between the mathematically possible and the truly existing co-occurrences concerning word order patterns". Many of the mathematically possible phenomena are unattested as some linguistic regularities constrain the languages in the selections they make. The method for explaining these regularities is done by identifying precisely the small number of attested co-occurrences, which can only be achieved with exceptionless implicational statements (avoiding distributional ones).

The first reformulation of Greenberg's universals proposed by Hawkins (1983) concerned the position of adjective and genitive:

I) "If a language has SOV (or simply OV) word order, then if the adjective precedes the noun, the genitive precedes the noun (i.e.: SOV ⊃ (AN ⊃ GN))"

This implicational universal allows the following co-occurrences: "AN & GN"; "NA & GN"; and "NA & NG". On the other hand, "AN & NG" is excluded. In Hawkins' language sample, only the allowed scenarios are attested.

VSO languages exhibit the precise mirror-image pattern (Hawkins, 1983):

II) “If a language has VSO word order, then if the adjective follows the noun, the genitive follows the noun (i.e.: VSO ⊃ (NA ⊃ NG))”

In this case, the universal cannot be generalized to all VO cases as some of the attested SVO languages also present the noun modifier co-occurrence which is not allowed by II: “NA & GN”. Nevertheless, Hawkins (1983) proposed that it can be transformed to a more generalized verb-first (V-1) classification universal (verb-first order instead of VSO). This V-1 class can be divided in three subclasses: VSO, VOS and V-initial (where no basic relative order of S and O can be defined).

Languages with preposition (Prep type) have the same conditioning effect on the co-occurrence of adjective and genitive order as V-first languages and the definition of the third non-statistical universal was made by considering the possible co-occurrences defined in universals I and II (Hawkins, 1983):

III) “If a language has Prep and any verb position other than SVO, then if the adjective follows the noun, the genitive follows the noun (Prep & -SVO ⊃ (NA ⊃ NG))”

Concerning languages with postpositions, as some exceptions were observed in the language sample, only a statistical universal was possible to be defined.

Hawkins (1983) followed the same strategy for the revision of other Greenberg’s universals concerning: “adposition order within the adposition phrase; noun modifier orders within the noun phrase; adjective modifier orders within the adjective phrase”. The main difference is that instead of using verb-based statements, he preferred to focus on adpositions (Prep and Post language-types) due to the typological uncertainty of SVO types.

The new established universals allowed Hawkins (1983) to join all the information into one single and more general implication and also defined the “Prepositional Noun Modifier Hierarchy” (PrNMH):

“Prep ⊃ ((NDem ∨ NNum ⊃ NA) & (NA ⊃ NG) & (NG ⊃ NRel))”

This hierarchy expresses the “relative instability of noun modifiers regarding the maintenance of the operator-operand serialization of the adposition phrase, thus, the demonstrative and numeral are less stable than the adjective, while the adjective is more unstable than the genitive, as well as the latter is more unstable than the relative clause” (Hawkins, 1983).

An example proposed by Hawkins (1983) concerns the doubling phenomenon observed in French concerning nouns and adjectives (NA and AN). It is noticeable that this phenomenon occurs exactly at the transition points between preposed and postposed modifiers. This observation has an impact on diachronic changes studies as it allows the prediction of the historical acquisition of doublets: if a language from the subtype 4 (DemN & NumN & AN & GN & NRel) acquires a doublet (e.g.: French language), this phenomenon is predicted to occur either for the adjective or for the genitive.

Concerning the postpositional languages and their mirror-image pattern when compared to prepositional ones, Hawkins (1983) established that:

- "If a language has Postp word order, then if the demonstrative precedes the noun, the genitive precedes the noun (Postp ⊃ (DemN ⊃ GN))".
- "If a language has Postp word order, then if the numeral precedes the noun, the genitive precedes the noun (Postp ⊃ (NumN ⊃ GN))".

Hawkins (1983) defined other implicational universals observing that the relative instability of demonstratives and numerals compared to adjectives in prepositional languages cannot be extended to postpositional ones: in the case of demonstratives, postpositional languages obey the same implicational regularity as prepositional languages (Postp ⊃ (DemN ⊃ NA)), thus, the adjective is more unstable than the demonstratives for Postp languages. From this analysis, Hawkins (1983) proposed the following generalization:

- "If a language has noun before demonstrative, then it has noun before adjective; i.e., NDem ⊃ NA (equivalent to AN ⊃ DemN)"

In a similar way, Hawkins revised other Greenberg's universals concerning numerals and adjectives; relative clauses and genitives; demonstratives and relative clauses; as well as numerals and relative clauses.

As it was done for prepositional languages, Hawkins (1983) also defined a Postpositional Noun Modifier Hierarchy (PoNMH):

"PostP ⊃ ((AN ∨ RelN ⊃ DemN & NumN) & (DemN ∨ NumN ⊃ GN))"

This hierarchy implies that just 8 out of the 32 possible co-occurrences are possible for postpositional languages. Of the allowed co-occurrences, 3 overlap with the ones permitted in prepositional languages.

After proposing his set of revised implicational universals (most of them exceptionless), Hawkins (1983) attempted to explain why the implications allow or disallow certain co-occurrences and why prepositional and postpositional languages differ in some features while agreeing in others.

The first explanation proposed by Hawkins (1983) concerned the concept of heaviness which is the base of his "Heaviness Serialization Principle" (HSP). Considering the PrNMH, the author suggested that prepositional languages place "lighter" constituents before the head and "heavier" ones after it: usually, demonstrative and numeral determiners are lighter than descriptive adjectives which are, generally, shorter than genitives, etc. Therefore, the heaviness hierarchy was defined as:

$$Rel \ \geq Gen \geq Adj \geq \begin{Bmatrix} Dem \\ Num \end{Bmatrix}$$

Where "≥" means "greater than, or equal to" in terms of heaviness. This concept follows four aspects: length and number of morphemes, number of words, syntactic depth of branching nodes, and presence of dominated constituents (Hawkins, 1983). The ranking on the HSP is explicated by the processing preference for the head (noun) to happen as early as possible in noun phrases (heavy modifiers delay the recognition of the head, thus, retarding its processing which also slows the identification of arguments and predicates).

Nevertheless, the HSP cannot explain on its own the differences encountered between Prep and Postp languages due to the fact that some Postp co-occurrences show some variance with the HSP. Thus, Hawkins (1983) proposed an additional concept, the Mobility Principle (MP):

$$\begin{Bmatrix} Adj \\ Dem \\ Num \end{Bmatrix} \geq \begin{Bmatrix} Rel \\ Gen \end{Bmatrix}$$

In this case, "≥" means "exhibits greater or equal mobility away from the adposition + NP serialization". In other words, Adj, Dem and Num are more mobile than Rel and Gen and, thus, move around the heads more easily. It produces a serialization which is the opposite to the one regarding the adposition in relation to modifiers. Hawkins (1983) argued that the MP is based on casual factors related with syntax (more moveable constituents being nonbranching and non-phrasal), and history (classically, if one language reorganizes any of its noun modifiers, it involves single-constituent and non-phrasal nodes first).

Both HSP and MP interacts according to the "Mobility and Heaviness Interaction Principle" (MHIP): for some word order co-occurrences the two principles make the same predictions; for others the HSP makes prediction with respect to which the MP is neutral and, sometimes, the two principles are in conflict (which happens only in Postp languages) (Hawkins, 1983). Consequently, where conflicting predictions are made for 2 noun modifiers in terms of word order, the larger is the heaviness difference of the two modifiers, the greater or equal is the ability of HSP to prevail over the MP's contrary predictions; however, HSP will surpass MP's predictions only when the heavier modifier is a relative clause.

To sum up, prepositional languages consistently place lighter modifiers to the left and heavier ones to the right, while postpositional languages have some heavier constituents to the right with lighter ones to the left.

Following the revision of Greenberg's universals, Hawkins (1983) proposed a new "word order type" notion, stating that it does not mean uniform ordering for all possible operator-operand pairs. Instead, a language included in a type respects a specific set of co-occurrence possibilities which are determined by the implicational universals.

Therefore, a type does not define "exactly a set of co-occurring properties which are present in the languages of the type. It actually determines a limited number of families composed of co-occurring properties and establishes exceptionless combinatorial orderings that are correlated with the typological indicator. Furthermore, all the correlating word order combinations are not unique to each type, some overlap are possible and have been attested" (Hawkins, 1983).

The most important typological indicators of Hawkins' approach (1983) are Prep, Postp, V-1 and SOV (or only OV). All of them are principally "operand-peripheral" (the operand occurs in the extremes of the phrases). The attested ambivalent behaviour of SVO languages does not allow a verb-based typology, hence, Hawkins (1983) suggested that Prep and Postp are much better typological indicators. This way, two major language types concerning word order can be defined: prepositional and postpositional.

After these main considerations, Hawkins (1983) briefly analysed more complex cases (e.g.: noun phrases containing more than one modifier at the same time). His objective was to analyse what relative orderings are observed and what are the possible cross-categorial generalizations for longer sequences. Again, Hawkins' strategy was to revise one specific sequencing universal concerning noun modifiers from Greenberg's universals set: "When any or all the items (demonstrative, numeral and descriptive adjective) precede the noun, they are always found in that order. If they follow, the order is either same or its exact opposite" (Greenberg, 1963).

By scrutinizing his language set, Hawkins (1983) proposed the following revision to the above-mentioned universal: "When any or all of the modifiers (demonstrative, numeral, and descriptive adjective) precede the noun, they (i.e.: those that precede) are always found in that order. For those that follow, no predictions are made, though the most frequent order is the mirror-image of the order for preceding modifiers. In no case does the adjective precede the head when the demonstrative or numeral follows".

Hawkins (1983) also showed that the attempt of using constituency and adjacency to explain this specific universal fails as there are three major sources of variation:

1) "Languages may vary within the constraints permitted by constituency and adjacency".
2) "Languages may vary by having different constituent structures".
3) "Languages may vary in the extent to which adjacency holds".

Another important contribution from Hawkins' work (1983) concerns the "Cross-Category Harmony" (CCH) principle which explains the statistical co-occurrences distributions throughout the attested types defined by the universals. The CCH predicts the relative quantities of languages that have the implicationally allowed word order co-occurrences sets.

The CCH states that there is a calculable preference for the proportion of preposed and postposed operators within one phrasal category (i.e.: NP, VP, AdjP, etc.) that can be extended to the others. The operand position of one phrasal category has an influence on the operand position of the other ones. Accordingly, the more inconsistencies there are concerning word order co-occurrences compared to the ideal harmonic ordering, fewer languages are attested. For example, "if the operators on the verb and on the adposition are all preposed (SOV & Po), then, the most favoured languages are those whose operators on the noun are also all preposed (thus, more languages being attested in the language sample), the next most favoured languages possess only one noun operator postposed, and so on" (Hawkins, 1983).

The language set analysed by Hawkins (1983) also allowed the addition of subject and verb in the operator-operand relation. When subjects are preposed, the operator-before-operand ordering is harmonic (preposed noun modifiers and postpostions). Thus, "SVO languages have one solid operator-before-verb ordering, even though the other operators on the verb are postposed" (Hawkins, 1983). However, the observed predominance of prepositions and postposed noun modifiers in SVO languages reflects the common serialization relative to the verb: operand before operator. It leads to the definition of the following types: VSO, SVO, SOV non-rigid and SOV rigid (verb final) (Hawkins, 1983).

Moreover, even if the CCH can base its predictions on the concept of cross-categorial balance, it is conceivable to establish a valid model for cross-categorial generalization even with only 2 operators. Hawkins (1983) investigated the subject and object order as a characteristic of all operator orders with respect to the verb:

- SVO languages: all operators after the verb, except for the subject.
- VSO languages: all operators after the verb.
- SOV languages extremely non-rigid: only subject and direct object operators before verb.
- Other SOV languages: other operators than subject and verb can also be positioned before the verb.

Thus, “SVO is more similar to VSO than to SOV and the cross-categorial generalizations mirror this similarity” (Hawkins, 1983).

The CCH predictions for shared category word order pairs can be defined as (Hawkins, 1983):

- “Given two word order co-occurrence pairs, W and W’, which satisfy the following conditions:
    1) W consists of word order co-occurrence pair A & B, and W’ of A’ and B’.
    2) A, A’, B, B’ are all ordered sets of grammatical categories.
    3) Sets A and A’ have the same categories as members: one operand a and at least one operator upon a. Sets B and B’ have the same categories as members: one operand b (where b ≠ a), and at least one operator upon b.
    4) The relative orderings of operand to operator(s) differ either between A and A’, or between B and B’, or between both, and are subject to the co-occurrence predictions of implicational universals.
- Then, the relative cross-category harmony of W and W’ is determined as follows:
    1) Calculate the number of operator-operand deviations from the nearest operand ordering(s) with no deviations for each pair.
    2) The fewer the number of deviations, the greater is the cross-category harmony of co-occurrence pair W(A&B) or W’(A’&B’).”

The prediction that can be made is that whichever co-occurrence pair W and W’ respects more the CCH, the higher will be the number of the attested languages. The above definition can also be made considering more categories within the co-occurrence pairs (i.e.: W consisting of A&B, or B&C, or A&C), or considering whole language types (i.e.: language type T consisting of co-occurring word orders A & B & C) (Hawkins, 1983). This principle was verified by Hawkins (1983) in his language sample, however, some exceptions were encountered, mainly due to general counter-principles, or because of the unrepresentative condition of some types concerning specific word order patterns in his set. For his test concerning the CCH applied on the whole language sample, the total success rate of the predictions was 95.7%.

Hawkins (1983) also compared his approach with a statistical view of Vennemann's NSP and with Greenberg's statistical implications. The conclusion is that the CCH is more successful in predicting the number of languages for each co-occurrence word order types. With the CCH, he was able to predict the regular frequency differences among the observed language-types: one type may be harmonic, however, if a specific implicational universal is disrespected, it will not be attested. Consequently, while implicational universals do not allow frequency predictions, the CCH is a reliable method which gives relative language frequencies and does not refer to individual categories.

To go beyond the descriptive character of the CCH, Hawkins (1983) suggested that a large variety of factors (syntactic, semantic, psycholinguistic, historical, etc.) could be used as explanation. More precisely (Hawkins, 1983):

1) The CCH suggests "the validity of the syntactic-semantic parallelism between the verb and its modifiers, the noun and its modifiers, and so on (reflecting the reality of modifier-head theory)".
2) Some kind of analogy also arises from the CCH: "the operator preposing and postposing balance within one category generalizes to others due to the generalization linking concerning operators and operands". Also, languages have natural tendency for similar elements to be treated the in the same way.
3) The syntactic complexity also contributes in explaining the CCH. According to Jackendoff's (1977) X-bar theory, "languages with a more harmonic balance of operators and operands regarding different categories are preferred over disharmonic ones: harmonic orderings permit the formulation of more cross-categorial syntactic rules". In other words, a decrease in the CCH means an increase of the grammatical complexity.

The other themes developed by Hawkins in the book "Word Order Universals" (1983) concern mostly how the proposed universals and principles can predict language change in a diachronic perspective, thus, less relevant to this thesis in which languages are compared synchronically.

2.2.3. Dryer's contribution to syntactic typology

In 1992, Matthew S. Dryer published the article "The Greenbergian Word Order Correlations" in which he challenged Hawkins analysis focused in head and dependents, thus, being called as "Head-Dependent Theory" (HDT), and proposed a new theory called "Branching-Direction Theory" (BDT).

Dryer (1992) presented an empirical study based on a sample of 625 languages to determine exactly which pairs of elements correlate with verb and object in terms of word order. He affirms that the previous proposed theory (based on head an dependents) is not valid. The HDT affirms that correlations reflect a tendency towards constant serialization of heads and dependents. In opposition to that, he proposed a new theory ("Branching-Direction", BDT) for which "word order correlations are related to the tendency for languages to be consistently right-branching or consistently left-branching" (Dryer, 1992).

Dryer's work follows Greenberg's paper (1963) which showed that "the order of certain pairs of grammatical elements correlates with the order of verb and object" (Greenberg, 1963). However, Greenberg worked with a much smaller language dataset, therefore, generating questions about areal and genetic bias. Other previous works (Lehmann, 1973 and 1978; Vennemann 1973, 1974 and 1976; and Hawkins, 1983) do not present enough empirical support concerning their assumptions.

First, Dryer (1992) defined precisely what is a correlation in his own terms: "If the order of elements X and Y exhibits a correlation with the order of verb (V) and object (O), then, <X,Y> is a correlation pair. X being a verb patterner and Y being an object patterner".

Then, he defines the two main questions he intended to answer with his work:

1) "What are the correlation pairs?"
2) "What general property characterizes the relationship between verb patterners and object patterners".

To avoid problems that may arise from a sample where elements (languages) are not independent (due to genealogical or diffusion effects), Dryer (1992) proposed a methodology to guarantee that results are statistically valid: grouping languages in genetic groups (genera).

Therefore, for each correlation, it is not the number of languages but the number of genera that is considered. Also, he divided the sample into 6 large geographical areas which are known for sharing some macro-areal features: Africa, Eurasia, Southeast Asia & Oceania, Australia- New Guinea, North America and South America. In this way, if the word-order phenomenon is more recurrent (in terms of number of genera) than its opposite in all the 6 areas, it is consistently dominant, and, therefore, can be considered as a universal property of language, not only an areal phenomenon.

From these assumptions, Dryer (1992) refined his definition of correlation based on verb and object relative order: “If a pair of elements X and Y is such that X tends to precede Y significantly more often in VO languages than in OV languages, then <X,Y> is a correlation pair and X is a verb patterner and Y an object patterner with respect to this pair”.

Most of previous work assumed that HDT is correct (sometimes with different terms to define head and dependents), thus, basically considering that languages follow two ideals: head-initial or head-final. However, for many observed phenomena, the predictions depend on one’s assumption about which element is the head, and this can be quite controversial. Therefore, when HDT is evaluated, the implications of these assumptions must be considered.

Dryer’s theory (1992) avoids terms that may be controversial. The BDT can be defined as such: “verb patterners are non-phrasal (nonbranching, lexical) categories and object patterners are phrasal (branching categories)”. Therefore, languages can be: “right-branching (phrasal categories follow non-phrasal ones) or left-branching (phrasal categories precede non-phrasal ones)” (Dryer, 1992).

Furthermore, Dryer (1992) analysed within his language sample the pairs of elements that can be explained using HDT. Six pairs were examined (the first element of the pair being the verb patterner and the latter, the object patterner): noun/genitive, adjective/standard of comparison, verb/PP, verb/manner adverb, copula verb/predicate, and “want”/VP. For each one of the listed pairs, the author shows evidence that they correspond to a correlation pair, respecting the statistical criteria explained above.

Additionally, Dryer (1992) scrutinized five other pairs which do not correspond a correlation one when tested within the language sample: adjective/noun, demonstrative/noun, intensifier/adjective, negative particle/verb, and tense or aspect particle/verb. According to him, if HDT was correct, these pairs should be correlations, however, this is not what was observed, thus, building strong evidence against HDT.

Next, Dryer (1992) proposed an analysis of eight controversial pairs for which there is no theoretical agreement regarding the head and dependent definition. In these cases, the HDT can be correct only if a precise assumption is made. The analysed pairs were: "tense and aspect auxiliary verb/VP, negative auxiliary/VP, complementizer/S, question particle/S, adverbial subordinator/S, article/N, plural word/N, and verb/subject" (Dryer, 1992).

Thus, he attested that a new theory should be considered, the "Branching-Direction Theory" which was described as: verb patterners are non-phrasal (nonbranching, lexical) categories and object patterners are phrasal (branching) categories. Thus, for "a pair of elements X and Y, XY is more often present in VO languages if X is non-phrasal and Y is phrasal" (Dryer, 1992). However, BDT depends on one's assumptions about the constituent structures.

Dryer (1992), then, presented one point on which his first definition of BDT does not account and proposes a new precise definition: "Verb patterners are non-phrasal categories or phrasal categories that are not fully recursive and object patterners are fully recursive phrasal categories in the major constituent tree". Consequently, only major constituents can be considered by the BDT.

However, if some assumptions can be done in terms of heads, the BDT can be more elegantly defined (Dryer, 1992): "Verb patterners are heads and object patterners are fully recursive phrasal dependents". The author says that both versions can be considered and states that further studies should be done to see which one is the best. The BDT requires a "high degree of hierarchical constituent structure" (Dryer, 1992), therefore could be problematic for non-configurational languages (Japanese for example).

Some cases that present complications which may impact the BDT were also listed (Dryer, 1992): numerals, demonstratives, manner adverbs, subjects, and affix position. For each one, details regarding the possible problems were described and some explanation or assumptions to validate BDT were provided.

Dryer is also responsible, together with Martin Haspelmath, David Gil, and Bernard Comrie, of the edition of "The World Atlas of Language Structures" (WALS) published in 2005 and of its the online database[5]. They provide information of structural properties of languages gathered from a large variety of descriptive materials. The book consists of 142 maps mostly enriched with texts concerning a large variety of language features (produced by a team of more than 40 authors). A total of 2,650 languages is represented in the atlas, however, not all information is available for all languages. It is considered as a valuable resource for cross-lingual comparison and for typological analysis of specific linguistic phenomena and has become a reference for typologists all over the world, as it will be described in later sections. However, this database is mainly descriptive, the properties are described individually, thus, it does not include word order co-occurrences (Hawkins, 1982) or correlations (Dryer, 1993).

---

[5] https://wals.info/

The information provided by WALS allows languages to be compared in terms of specific properties. It has been used as one of the sources of the lang2vec tool (Littell et. Al, 2017) which provides language representations in the format of vectors (composed of their linguistic features). These vectors will be used to define the base-line in terms of language classification in terms of syntactic features in this thesis.

If universals are not a part of WALS, it is the main information provided by the RaRa and the Universals Archive[6] which have been created as the outcome of the project "Sprachbaupläne" of the Universität Konstanz (Plank and Filimonova, 2000). The objective of this group was to collect and document linguistic universals that have been proposed in relevant literature, especially the implicational ones ("if a language has property X, then it will also have Y"). Currently, it includes over 2,000 entries freely available, and with this database, users can search universals in terms of any of the individual words or combinations of words that occur in their formulation. The domain can also be specified: syntax, morphology, phonology, phonetics, semantics, and lexicon.

The linguistic universals inside the Universals Archive are substantive, thus, design features defining natural languages (such as vocal-auditory channel) are not included in this database. Moreover, the substantive universals are statements preceded by a universal quantifier ranging over all natural languages: they are either unconditional ("for all languages, there is or is not X"), implicational, or describe correlations. Consequently, a property to be considered universal must be universally shared by all languages, or, if not shared, must not vary independently across them. This universals dataset defines a repository of anything that any language is free to select. The archive also contains statistical universals (tendencies). Furthermore, the collected universals are timeless and its properties includes all kind of units, categories, constructions, rules, constraints, principles, and relationships. A work of standardization (and translation to English) was conducted in order to guarantee the homogeneity of the provided information (Plank and Filimonova, 2000).

---

[6] https://typo.uni-konstanz.de/rara/

Each universal in the Universals Archive is presented in its original formulation (from the literature) and in its standardized form. Information about its domain (e.g.: syntax, morphology), type (e.g.: implication, mutual implication, rarum, unconditional, etc.), quality (e.g.: statistical, absolute, etc.), source, and possible counterexamples is also provided. Although providing valuable information concerning language universals, as the aim of this study is to classify languages concerning their syntactic differences and similarities and not to find cross-lingual regularities, this database will not be used in our experiments.

While the aim of the works presented in this section was basically the identification and explanation of universals, implications, and correlations, what is proposed in this thesis is an applied investigation with the aim to find which typological aspects (concerning word order) are involved when different languages are combined to train dependency parsing tools, greeing, in a certain way, with the orientative usage of typology proposed by Vennemann (1981). In this work, the weight of verb and object relative position, as well as the head and dependency ordering will be tested using quantitative methods (along with more complex measures concerning syntactic structures) and will be compared to qualitative typological language descriptions provided by typological databases.

As it has been demonstrated by Dryer (1993), the head and dependent theory does not account for some specific correlations attested in some genera. Moreover, in some cases heads and dependents depend on specific assumptions, varying according to different authors. Nevertheless, in the specific case of the usage of head and dependents in typological studies concerning dependency parsing improvement, this fact is not a weakness as the experiments are conducted using language corpora annotated following a specific framework, with a unified view concerning the definition of these elements.

2.3. Corpora-based Typology

In 2022, Levshina presented an overview of corpora-based typological studies. The author claims that even though no corpus can replace the traditional typological data from reference grammars, the usage of corpora can increase the diversity in typological research specially by providing means for the investigation of probabilistic and gradient properties of languages. The importance of the corpora-based approach is attested by the increasing number of available cross-lingual corpora. Also, this approach is a useful way to identify and interpret cross-linguistic generalizations. Traditional grammars are based on someone’s judgements and intuitions based on restrict corpora. On the other hand, corpora-based studies permit the language to be investigated in a more direct and detailed manner, although also presenting some bias concerning the choices during the sampling phase (e.g.: word segmentation, annotation framework, etc.).

An overview of existing corpora, which can possibly interest typologists, was also presented by Levshina (2022). They are divided in three main types:

1) Parallel corpora: composed by “aligned sentences or other chunks of text in two or more languages” (Levshina, 2022). This type presents the highest semantic and pragmatic similarity between the contents.
2) Comparable corpora: in this case, texts are not parallel but have similarities regarding text types and topics.
3) Unified annotation: texts from this type can differ in terms of topics but are annotated uniformly following a precise framework.

Dahl (2004) showed, by using parallel corpora composed by Bible translations, that techniques which were developed for word alignment of parallel corpora are also efficient regarding the comparison of the distribution of grammatical phenomena across languages. The Parallel Universal Dependencies (PUD) corpora (Zeman et al., 2017) have been selected for this thesis as this collection follows a unified annotation and is parallel, thus presenting semantic and pragmatic similarities allowing the focus to be on the syntactic cross-lingual differences.

In terms of word-order analysis, reference grammars usually present the main observed strategy of each language regarding each ordering phenomenon. Nevertheless, less frequent or context-dependent occurrences may be left apart in this type of references, thus, not reflecting the gradient of the language in use (Levshina, 2019). For example, Östling (2015) presented a more detailed analysis of verb, object, and subject order with precise numeric frequency estimations instead of just proposing categorical labels such as SOV, SVO, etc. Moreover, Wälchi (2009) identified some specific context of usage of constructions formed by verb and locative phrases depending on the relative position of the components.

Corpora can be a useful resource concerning the discovery of new descriptive typological measures. Greenberg (1960) was a pioneer in this field by manually analysing 100-word samples of text to determine indices concerning morphological typology (e.g.: index of agglutination, index of synthesis, suffixal index, etc.) based on Sapir's previous typological work. Greenberg's work has been further developed by Szmrecsanyi (2009) who studied not only cross-lingual differences but also geographical and diachronic varieties of English, showing that this language is not homogeneous concerning analyticity and syntheticity. One identified problem faced by researchers when investigating morphology using corpora is the lack of abundant annotated data with morpheme segmentation or reliable tools for automatic analysis for low-resourced languages.

Another topic which has been explored by many authors using a corpora-based approach is the comparison of linguistic complexity (e.g.: Hawkins, 2004). Sinnemäki (2014) provided an overview of such studies, basically dividing them in two different types: complexity of grammar in general and local complexity (e.g.: of the tense and aspect system). Usually, what is measure in these works is the morphological or word order complexity as a global property of the language which can be estimated in several ways (e.g.: Juola, 1998 and 2008; Li et al., 2004; Cilibrasi and Vitányi, 2005; and Benedetto et al., 2002).

An interesting study about measuring language complexity was published by Bentz et al. (2016). The authors have analyzed a total of 500 languages from many different linguistic families and genera and compared three different measures (distributional-based approach) obtained from language corpora (word entropy, relative entropy of word structure, and type and token ratio) to the typological information in terms of morphological complexity provided by WALS database (paradigm-based approach) (Dryer and Haspelmath, 2013). What was observed is that although language complexity is conceptualized differently for each method, they are all strongly correlated, thus, each one reflecting different nuances of the fact that linguistic complexity is related to fundamental information-theoretic concepts of uncertainty or choice when encoding and decoding a message.

Still concerning linguistic complexity, Ehret and Szmrecsanyi (2016) analysed a sample of 1,200 languages (with different typological characteristics) from all over the world. Their strategy was based on reshuffling characters or words and on the usage of entropy scores calculated over minimum substring lengths. The obtained results confirmed the similar trend obtained in previous studies: the more information is carried by word order, the less information is conveyed by the morphology, and vice-versa.

Entropy measure using corpora can be used to investigate word order flexibility, for example, for two components, entropy value is 1 if two possible orders co-exist and have equal frequency

(50/50) and 0 if only one order is possible. Levshina (2019) used this approach to analyse several syntactic dependencies. Furthermore, Allassonnière-Tang (2020) showed that around 500 sentences is enough data to obtain simple coarse-grained measures of entropy for basic constituents. Moreover, Futrell et al. (2015) concluded that entropy measures allows to estimate the variability regarding head direction and ordering relations for many restricted relation types. They also showed that the usage of corpora from different sizes can interfere in the obtained results as entropy is highly dependent on that.

Another common usage of corpora in typology concerns the identification and tests of typological correlations such as the ones proposed by Greenberg (1963) and Dryer (1992). Naranjo and Becker (2018) published a study attesting several correlations between verb-headed and noun-headed dependencies using Universal Dependencies corpora. Another example is the work presented by Bentz and Ferrer-i-Cancho (2016) who investigated a sample of around 1,000 languages (from 80 different linguistic families) and observed a negative correlation between word length and its frequency, which enable them to conclude that the Zipf's law of Abbreviation (Zipf, 1965 [1935]) is an absolute universal (synchronic).

Corpora-based typology has also been implemented in the verification of typological implications. An example of this is the work developed by Gerdes et al. (2019b) concerning Greenberg's Universal 25: "If the pronominal object follows the verb, so does the nominal object" (Greenberg, 1963). Additionally, some works have been carried on with regard to head and dependent distances (Ferrer-i-Cancho, 2006; Futrell et al., 2015 and Liu, 2008, and Liu, 2020).

A different and more specific typological study, with a diachronic perspective, was presented by Sergey Say (2014) who quantitively examined 29 languages to investigate contextualized uses of bivalent predicates. The approach was based on the measure of a distance metric regarding entropy and pairwise mutual information between distributions. The results showed that distributions of verbs into valency classes develop quickly and are transferable in contact situations even when there are differences in terms of argument-coding devices.

Concerning low-resourced languages, Haig et al. (2011) developed a specific annotation system (GRAID) based on morpho-syntactic annotation of connected discourse (Haig and Schnell, 2011) more adapted to the documents regarding their language set (speech samples from 4 indigenous languages). After presenting the annotation system in details, they applied it to examine the different languages in terms of certain domains of discourse organization and showed that the strategy applied allowed to identify distribution of the linguistic phenomena even in the small available language samples. This fine-grained annotation system was also successful to provide enough information regarding pronoun deployment, phenomenon overlooked in previous works. This study shows how methodologies can be fine-tuned to adapt to low-resource scenarios.

With a perspective of studying diachronic syntactic changes which characterize the evolution from Latin language to Romance ones, Liu and Xu (2012) developed a method to analyse 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) composed their dataset. The selected treebanks came from the CoNLL-X Shared task (Buchholz and Marsi, 2006). The dependency syntactic networks for each language was characterized with the calculation of the following syntactic parameters extracted from each corpus:

- The mean sentential length;
- The percentage of the head-final dependencies;
- The percentage of the head-initial dependencies;
- The percentage of dependencies between adjacent words;
- The percentage of dependencies between non-adjacent words;
- The mean dependency distance of all head-final dependencies;
- The mean dependency distance of all head-initial dependencies;

They showed that the dependency syntactic networks coming from dependency treebanks 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. Nevertheless, CoNLL-X is not a unified annotation framework, the only universal principle being that syntax is represented by some kind of direct tree with labels for each dependency relation. Moreover, the selected corpora have difference sizes, although this fact does not seem to have affected the analysis and the final conclusions.

Finally, another approach regarding the extraction and comparison of syntactic information from treebanks is proposed by Blache, P. et al. (2016). Typological syntactic information is obtained by “inferring context-free grammars (together with statistics) from syntactic structures inside annotated corpora” (Blache et al., 2016). The cluster analysis comparing 10 different languages showed the potential of the proposed tool (MarsaGram) in terms of typological classification. This tool will be used in this thesis; hence, it will be later described in details.

Corpus-based typology has proven to be an efficient way to investigate syntactic phenomena, either to give a quantitative perspective to concepts described in grammar or to investigate and compare languages with new approaches. The set of parameters identified by Liu and Xu (2012) and the patterns which can be extracted using MarsaGram are part of this thesis methodology. The innovation proposed here is that the possible language classifications will be compared (in terms of correlation metrics) to the improvement (or decrease) in terms of dependency parsing metrics obtained when different languages corpora are combined, thus, proposing an extrinsic evaluation of the ways of classifying languages quantitatively and syntactically.

2.4. Typology and Natural Languages Processing

Typological information has been used in different ways in many studies aiming to improve dependency parsing results. It has been proved that “typological comparison of languages is a powerful way of increase overall metrics concerning dependency parsing automatic annotation, especially regarding low-resource languages and unannotated ones (which do not have any corpora annotated in terms of syntactic relations)” (Alves et al., 2022).

O’Horan et al. (2016) published a survey about the usage of structural typological information (concerning phonological and morphosyntactic features) in natural language processing. This prior study was, then, completed in 2019 by Ponti et al. who described the state-of-the-art concerning this topic in a much vaster survey that included semantic features and some aspects of typological strategies regarding machine and deep learning methods. For these authors, the importance of understanding linguistic variation at the surface level is decisive for the development of effective multilingual NLP tools, allowing NLP technology to become more globally accessible.

According to Ponti et al. (2019) language variation at the surface level has undesired consequences for NLP as most of algorithms are developed and tested (in terms of architecture and hyper-parameters) on a limited set of languages which can generate language-specific bias as described by Bender (2009 and 2011). Also, due to the machine learning and deep learning dependency on supervised and labelled data, low-resourced languages (and languages with no annotated data) cannot be effectively used to train these models.

Typology can contribute to overcome some of these limitations as it has been shown by some experiments where multilingual models performed better than monolingual ones (Pappas and Popescu-Belis, 2017). Moreover, typology can lessen several of these restrictions by helping the development of unsupervised models which do not rely on the availability of manually-annotated resources which are expensive, time-consuming, and require skilled labour. It can be achieved via three main categories of methods: by guiding the transfer of models or data from well-resourced languages to low-resourced ones, by proposing multilingual joint learning strategies, and by creating multilingual distributed word representations.

"Typology studies the variation across languages through their systematic comparison" (Comrie, 1989). It is a challenging task as linguistic categories cannot be universally predefined: there is "a lot of cross-linguistic variation in lexicons and grammars and newly discovered languages often exhibit unusual properties" (Ponti et al., 2019). Thus, language comparison should be functional and not based on formal criteria. The definition of benchmarks for cross-lingual comparison is usually based on solid documentation (Bickel 2007a) coming from the gathering and analysis of linguistic data. Typological information is usually stored in large databases of attribute-value pairs (also called "typological feature"): each attribute corresponds to an attested structure and each value to the most common observed strategy.

Ponti et al. (2019) also pointed out that "any cross-lingual generalization must be demonstrated through a representative sample of languages: the sample should be large enough to include even rarer features". However, some bias can appear due to the fact that many languages do not possess information concerning all features due to insufficient documentation, and because some similarities between them may not always be caused by language-internal dynamics but from external factors: it can be inherited from a common ancestor (genealogical bias) or borrowed by contact with a neighbour (areal bias). Thus, some features may be widespread inside a genealogical family or geographical region, but extremely rare elsewhere.

As previously mentioned, typological features concerning a huge number of languages and regarding multiple distinct levels of linguistic description have been gathered by typologists in open-source databases. These catalogues, presented in table 2.2, organize the obtained information in terms of universal attributes and language-specific values (e.g.: WALS, which has been briefly described in the previous section[7]).

Database Name	Levels	Coverage
World Loanword Database (WOLD)	Loanwords (lexicon)	41 languages (24 attributes)
Syntactic Structures of the World's Languages (SSWL)	Morphosyntax	262 languages (148 attributes)
World Atlas of Language Structures (WALS)	Phonology, Morphosyntax, Lexical semantics	2,676 languages (192 attributes)
Atlas of Pidgin and Creole Language Structures (APiCS)	Phonology, Morphosyntax	76 languages (335 attributes)
Valency Patterns Leipzig	Predicate-argument structures	36 languages (80 attributes)
Lyon-Albuquerque Phonological Systems Database (LAPSyD)	Phonology	422 languages (70 attributes)
PHOIBLE Online	Phonology	2,155 languages (2,160 attributes)
StressTyp2	Phonology	699 languages (927 attributes)
Intercontinental Dictionary Series (IDS)	Lexical Semantics	329 languages (1,310 attributes)
URIEL Typological Compedium	Phonology, Morphosyntax, Lexical semantics	8070 languages (284 attributes)
Automated Similarity Judgment Program (ASJP)	Lexical Semantics	7,221 languages (40 attributes)
AUTOTYP	Morphosyntax	825 languages (1,000 attributes)

Table 2.2. List of databases concerning typological information (Ponti et al., 2019).

Among these databases, WALS has been the most extensively used in NLP systems as it provides phonological, morphosyntactic and lexical information for a high amount of languages (Ponti et al., 2018). The URIEL Typological Compendium is a meta-repository which wraps several databases together (Littell et al., 2017), being the base of the lang2vec tool previously mentioned in the last section (used for our baseline in terms of syntactic typological classification).

---

[7] The Universals Archive is not considered here as the type of information it contains does not correspond to the type attribute-value (typological feature), instead, it corresponds to cross-lingual generalities (universals).

One problem usually present in these databases is the fact that they suffer from discrepancies which are caused by their variety of sources. Furthermore, there are many gaps as not all languages have the same amount of descriptive literature. Moreover, most databases fail to illustrate the variations that can occur within a single language (as only the most frequent phenomena are reported, not all possible ones), and, finally, some redundancy can be found (e.g.: WALS feature 81A "Order of Subject, Object and Verb" which is the sum of WALS 82A "Order of Subject and Verb" and WALS 83A "Order of Object and Verb").

As previously mentioned, typological information can help improving linguistic information transfer from well-resourced languages to low-resourced ones, also called in this scenario as source and target languages respectively (Ponti et al., 2019).

There are three mainstream strategies for language transfer:

1) Annotation projection: a source-labeled text is aligned at the word level with a target raw text and the annotations are projected (Yarowsky et al., 2001 and Hwa et al., 2005). Or, "in a more optimized scenario, the propagation of labels can be conducted over multiple steps based on bilingual graphs built with distributional similarity functions (Das and Petrov, 2011) or constituents (Padó and Lapata, 2009)" (Ponti et al., 2018). Moreover, when propagation is done via model expectations on labels or sets of most likely annotations, it is called soft projection (Wang and Manning, 2014, Khapra et al., 2011, and Wisniewski et al., 2014). Dependency parsing projections are more complex as it involves sets of vertices (words) and edges (dependencies). Yet, the transfer can be improved with auxiliary linguistic resources: token-level constraints on labels (Li et al., 2012 and Täckström et al., 2013) and type-level constraints extracted from dictionaries during projection (Ganchev and Das, 2013).
2) Model transfer: a model is trained on a source language and tested on a target one (Zeman and Resnik, 2008) where, usually, models are delexicalized before being transferred due to the vocabulary incompatibility. In this approach, the models are either fed with language-independent features or with harmonized ones (Zhang et al., 2012). Softening the source constraints and enforcing the linguistically motivated ones is a way of reducing cross-lingual differences in linear order structures and lexicalization and translation can help by corelating non-overlapping vocabularies

(Agić et al., 2014). However, when languages are not close to each other, the quality of alignment is worsened (Agić et al., 2016). One limitation of annotation projection is the necessity of having parallel data (Agić et al., 2015). Typological strategies can be used with the aim of simplifying the annotation projection by tying universal features together, especially in multi-source transfer (Agić et al., 2016 and McDonald et al., 2011).

3) Multilingual Joint Supervised Learning: Multiple languages are used to train jointly probabilistic models. Ammar et al. (2016) and Khapra et al. (2017) showed that this type of model often surpasses monolingual ones, especially where all languages are low-resourced. Moreover, in the low-resourced scenario, results are improved via the optimization of the learning phase by discovering the most relevant examples to annotate (Fang and Chon, 2017). The main challenge regarding this strategy is to tailor the joint model to be optimized for a target language with a balance between private (monolingual) and shared network components (multilingual). The latter can be enhanced with typological approaches, for example, by decoding specific typological properties from monolingual representations (Malaviya et al., 2017), or from the multilingual components (Johnson et al., 2017).

In their survey, Ponti et al. (2018) focus on typological features extracted from the crafted databases presented in table 2.2. Typological strategies in NLP studies concerning morphosyntactic annotation most often involves cross-lingual comparisons regarding a selected subgroup of word order features from WALS database. Several studies consider only nouns, verbs, and modifiers (following the work of Naseem et al., 2012), it is the case of the work developed by Ammar et al. (2016), Daiber et al. (2016), Täckström et al. (2013), Zhang et al. (2012), and Barzilay and Zhang (2015). These studies differ mostly in terms of the language-set and total number of selected word-order features for cross-lingual comparison. In 2016, Berzak et al. presented a study concerning all non-redudant morphosyntactic features in WALS (119 in total), while Agić et al. (2017) and Ammar et al. (2016) used all WALS available features, and Deri and Knight (2016) preferred to compare languages regarding all URIEL listed properties.

Ponti et al. (2018) also warned about the unrestricted usage of these typological databases as, even though they are rich in terms of linguistic information, their feature sets are often incomplete for many languages, especially low-resourced ones. Moreover, when typological information is considered, it is often encoded as vectors for which each dimension corresponds to a feature associated to its language-specific value (most commonly in a binarized form as presented by Georgi et al., 2010). This fact can be problematic as not all features are compatible with this type of representation. Some work has been done to provide automatic filling of missing values (using genealogical information to predict them). It can be useful in the case of stable features but not for the ones which are more sensible regarding time changes. Absent feature values can be predicted based: on morphosyntactic annotated text (e.g.: treebanks) as it has been demonstrated by Liu (2010), on the propagation from other values in a database via language clustering regarding linguistic similarities (Teh et al., 2007 and Littell et al., 2016) which can be done using supervised learning from Bayesian models or neural networks (Takamura et al., 2016), and on heuristics methods concerning co-occurrences metrics (using multi-parallel texts as showed by Wälchli and Cysouw, 2012).

Generally, typological features are joined in NLP algorithms in three different ways: by assisting the development of such models when features are converted into rules or prior assumptions (e.g.: in Bayesian graphic models), by their usage to expand the input representations or to tie together specific parameters across languages, or by guiding data selection and synthesis.

Regarding rules and prior assumptions, Bender (2016) developed the "Grammar Matrix kit" which consists of "one universal core grammar and language-specific libraries for phenomena where typological variation is attested". Also, typological features can determine the design of graphical models of Bayesian networks (Schone and Jurafsky, 2001) by assigning part-of-speech tags to word clusters learned in an unsupervised way.

The most common usage of typological features is to tie specific parameters together and provide input representations of language properties in language transfer of multi-lingual joint learning. One well-known approach is the one developed by Naseem et al. (2012) and adopted by Täckström et al. (2013) and by Barzilay and Zhang (2015). It is called "selective sharing" and is a way to parse sentences in a language transfer situation where many source languages are used to develop a model to parse a target language without any annotated available dataset. The assumption is that parts of speech of pairs (composed by a head and a dependent) are universal, but their ordering is specific.

The selective sharing strategy factorizes "the recursive generation of dependency tree fragments into two steps" (Naseem et al., 2012). The first one is universal; the algorithm selects an unordered set of dependents, and the second step is a language-specific phase where each dependent is assigned with a direction (left or right) with relation to the head based on the language and following a specific probability. Dependents in the same direction are eventually ordered with a probability drawn from a uniform distribution of their possible permutations. Typology information (represented in vectors) is used in the second step, guiding the calculation to optimize the likelihood of the observations.

The previous method has been improved by Täckström et al. (2013) who proposed a discriminative model (delexicalized first-order graph) allowing to dispose strong independence assumptions (e.g.: between choice and ordering of dependents) and display invalid combinations. They also proposed the usage of language-specific features for the directionality of dependents (combination of the part-of-speech tags of head and dependents with WALS values).

Barzilay and Zhang (2015) modified Täckström approach by proposing tensor-based models that avoid problems linked with the manual feature selection. Their idea was to "induce a compact hidden representation of individual features and languages by factorizing a tensor built with their combination, thus, generating intermediate feature embeddings in a hierarchical structure".

Beside the selective sharing approach, other methods have been developed using the multi-lingual biasing strategy. It is the case of the multilingual parser proposed by Ammar et al. (2016) which interlaces both language-specific and language-invariant features in the feature set: universal coarse part-of-speech tags, multi-lingual word embeddings and multilingual word clusters. Typological information is used to condition the hidden states of language models.

The other typical usage of typological features concerns data selection with the intention to: choose the most appropriate source language and/or to weight the influence of each language in multilingual combined models. This selection is normally defined by using general language similarity metrics or by measuring the overlap concerning language-independent properties (e.g.: part-of-speech sequences) (Ponti et al., 2018).

2.5. Typology and Dependency Parsing

As presented in the previous section, many studies regarding unannotated languages are based on a typological characterization (using typological features and/or part-of-speech combination patterns) that allows the determination of the most similar language whose annotated corpus is, then, used to train the model that will serve to annotate the target one. This method is called "Single Source" which is different from the "Multiple Source" one where all possible training corpora are concatenated and used to train only one universal model. In most studies, the training corpora are delexicalized which avoid lexical interference when processing the target language.

Methods using part-of-speech patterns as a comparison feature require that the target language must have at least some sentences annotated in terms of part-of-speech. When other typological features coming from typological databases (such as WALS) are considered for language comparison, it is necessary that the target language is sufficiently typologically described in these catalogues.

Lynn et al. (2014) focused on the implementation of a cross-lingual parsing strategy for the Irish language. McDonald et al. (2011) described two different methods concerning this approach:

1) Direct transfer: a delexicalized version of the source language treebank is used to train a parsing model which is then used to parse the target language
2) Projected transfer: the direct transfer approach is used to seed a parsing model which is then trained to obey the constraints of the source language which are learned from a parallel corpus

McDonald et al. (2011) also showed that genealogically related languages were not always the best source-target pairs. On the other hand, Petrov et al. (2011) obtained interesting results for languages from the same linguistic genera (such as Romance and Germanic) but also in experiments using training data from more heterogeneous combinations. In their study, Lynn et al. (2014) used delexicalized corpora from 10 different languages to train parsing models (using MaltParser, Nivre et al. 2006) which were tested with a delexicalized Irish test set. The best results (in terms of UAS and LAS) were obtained with the model trained with Indonesian language (Austronesian language), confirming that the in some cases, better improvements are obtained with non-related languages.

In 2017, De Lhoneux et al. also used the combination of corpora as a strategy to train dependency parsing models for languages without any annotated data (Buryat, Kurmanji, North Sami, Upper Sorbian, Kazakh, and Uyghur).

The adopted method was to develop training corpora based on support languages which were defined using four criteria:

1) Language relatedness: languages from the same genealogical family;
2) Script: languages which have the same type of script;
3) Geographical: closeness in terms of geographical distance;
4) Performance: all possible support languages were used to train parsing models and were evaluated with the test sets of the challenging languages, the languages with the best results were selected to be part of the final training corpora.

The obtained results were encouraging concerning the languages with no training resources, again, confirming the benefits of corpora combination.

Another example is the method proposed by Agić (2017) where three language are combined via comparative techniques that choose the best single source for an unannotated language (containing only part-of-speech information) using: part-of-speech trigrams, a language identification software (langid.py tool, developed by Lui and Baldwin, 2012), and WALS features. The strategy considers the available data of the target language to determine the best training corpus (source language) by calculating which is most similar language in terms of the described comparative features. Later, it has been showed by Litschko et al. (2020) "that better outcomes are obtained when the same typological features are used to analyse separately each sentence of the target corpus, defining, for each instance the best source model, thus, not using the only one source language to parse the whole target text. In both studies, only qualitative typological features and surface level word order (part-of-speech trigrams) are analysed" (Alves et al., 2022).

While "the studies described in the previous paragraph are based on the analysis of part-of-speech trigrams for cross-language comparison, Wang and Eisner (2018) proposed a method to compare word order (again using part-of-speech possible combinations) which is based on a deep-learning algorithm (multilayer perceptron architecture) that classifies languages in an unsupervised way with the information extracted from delexicalized corpora" (Alves et al., 2022). This model is, then, processed to allow the identification of the most appropriate source language. Their main goal was to prove that part-of-speech (POS) sequences convey valuable information about syntax. The authors used, as part of their dataset, the Galactic Dependencies treebank (Wang and Eisner, 2016) which is composed by around fifty thousand artificial languages. The new synthetic corpora were generated by selecting a substrate language (represented in the Universal treebanks), and systematically reordering of the dependents of some nodes using the order features of other UD languages. They have shown that even though the fact that the synthetic languages violate some typological universals or typological tendencies, and that the parsability and the perplexity of a real training language usually get worse when nodes are permuted, the new languages can improve dependency parsing results for unannotated languages when the best single model is selected using trigrams part-of-speech comparison.

Another strategy concerns uniquely the usage of typological information from URIEL database (lang2vec tool, Littel et al., 2017), as presented by Glavaš and Vulić (2021). The technique consists of comparing the vector composed by the values of the linguistic features of the target language with the ones containing the typological characteristics of well-resourced languages. The idea was not to determine the best corpus, but to associate the most similar data from different languages as long as the similarity metric respects a specific threshold.

Fisch et al. (2019) analysed the challenges of integrating typology into neural dependency parsers concerning two different approaches for delexicalized dependency parsing (using Biaffine Parser, Dozat et Manning, 2016). The first approach is the selective sharing proposed by Naseem et al. (2012), while the second concerns the addition of typological information as a complementary feature of the input sentence. The typological analysis for language comparison was based on a specific subset of features regarding word order (WALS). The typological approach was compared to the average directionality of each corpus (as proposed by Liu, 2010) and to the surface statistics of part-of-speech tags. Fisch et al. (2019) observed that typological information is an effective way of improving parsing metrics (statistically similar to the language comparison using the corpus directionality but superior to the surface statistics).

Another approach was developed by Scholivet et al. (2019) where a parser model was trained with a multilingual delexicalized corpora where each token was also associated to a vector derived from WALS database. The obtained results showed that this approach consistently improved LAS results when compared to the baselines (multilingual training corpus without typological information and multilingual training corpus with language identification label for each token).

Beside the usage of typology to determine the best source language to train a model to parse a target language without any training data, another possibility is to define typological strategies to guide corpora combination where datasets are combined to improve low-resource languages parsing results.

Stymne et al. (2018) presented a study regarding the combination of heterogeneous treebanks to train dependency parsers. The strategy of simply combining training data has the advantage of not requiring any modifications to the parser itself. This method was previously briefly tested by Björkelund et al. (2017) but with inconclusive results. Beside the simple approach of purely combining corpora, it is possible to improve the final results by either concatenating data from multiple languages for the first phase of the training step, then perform a fine-tuning step based on the target language dataset, or by providing specific language embeddings. Stymne et al. (2018) showed that all the described strategies present an improvement in terms of LAS metric when compared to models trained with only the target language dataset. Hitherto, the usage of treebanks embeddings proved to be the best method in terms of overall results. This study is interesting as it evidences the advantage of using combined corpora, however languages were combined without using any typological principle.

Another interesting study concerning corpora combination of related languages was presented by Smith et al. (2018) for the CoNLL 2018 Shared Task. Their system (Uppsala) was based on three components: the first one performs joint word and sentence segmentation, the second predicts part-of-speech tags and morphological features, and the third one determines the dependency trees from the words and tags. The parsing training step (greedy transition-based parser) was not conducted using single parsing model for each treebank but with multiple treebanks composed of closely related languages. Nevertheless, in the article, the way the authors used to define the relation between languages is not described in details. They used genealogical information but the source is not informed. Their objective was to optimize the performance of the system with a reduced number of o parsing models. The language balance in the multilingual training sets was guaranteed with the selection of 15,000 sentences of each language. The obtained results showed that, in general, the aggregated sets of treebanks (multilingual) presented better LAS scores when compared to single language ones, especially for low-resourced languages.

Deri and Knight (2016) proposed a method which extracts information from URIEL (genealogical, geographic, syntactic, and phonetic features) to compare languages and select the closest to the target one. Rosa and Žabokrtský (2017), on the other hand, established a comparitive method based on the divergence between part-of-speech trigram distributions.

This method was further optimized by Ponti et al. (2018) who used the Jaccard distance on morphological feature sets and the tree edit distance of lexicalized dependency parses regarding translationally equivalent sentences.

Besides, Ponti et al. (2018) also presented a strategy consisting of the usage of typological features to pre-process treebanks to reduce the variation in language transfer tasks. The source trees are adapted to the typology of the target language with respect to several constructions. This pre-processing method is rule-based: when a source subtree matches a construction documented in a typological database, it converts it to the target strategy. The conversion hinges upon a sequence of node addition, node deletion, and label change.

In a different approach, De Lhoneux et al. (2018) examined how several typological features are related to the dependency parsing scores when a set of 27 different deep-learning parameters are used for cross-lingual parameters sharing. These parameters correspond to three sets: character based one-layer (bidirectional LSTM), word based two-layer (bidirectional LSTM), and multi-layered perceptron (MLP) with a single layer. De Lhoneux et al. (2018) showed that the linguistic intuition concerning character and word-level LSTMs are very sensitive to both phonological and morphosyntactic variances, whereas the MLP parameters learn to predict less idiosyncratic (hierarchical relations from relatively abstract representations of parser configurations). The selected languages were classified considering genealogical family, and concerning the subject, verb and object order (in a qualitative manner).

Schuler and Agić (2017) developed an innovative method to compare languages in order to empirically sample the best source language to be used to train a parsing model for many target languages. They did not use typological databases to determine which language is the best candidate to become the source language for each specific target, instead, they used the parser performance (UAS) as a measure of similarity (parser generalization capacity from one language to another), generating language groups by using a standard network clustering algorithm (Infomap[8]). The obtained language clusters were diverse between them but coherent within. Thus, they were able to identify 9 representative languages, generating the same number of models which were able to parse, in an optimized way, a total of 47 languages. The obtained language clusters present many similarities with genealogical families but also some differences (e.g.: Hungarian and Chinese in the same cluster and Dutch in an isolated cluster, not together with English and Swedish).

---

[8] http://www.mapequation.org/code.html

Moreover, the UDapter tool (Üstün et al., 2020) was developed with the integration of linguistic typological features into the parsing network (103 syntactic, 28 phonological and 158 phonetic features from URIEL database). When the values regarding the features were not available, they were estimated by a k-nearest neighbours approach based on genetic, geographical and other features distances between languages. Considerable improvement was obtained for low-resources languages (zero-shot scenario), while for rich languages, results were similar. In terms of typological gain, the authors observed that all provided typological features contributed to the improvement of the results (not only the syntactic one as one could expect for the dependency parsing task).

It is possible to conclude from the overview of the related work presented in this section that typology is an effective way of improving dependency parsing results. A large number of the studies concerns cross-lingual parsing, meaning that one source language is selected and its corpus is used to train a parsing model to annotate a target language with no training corpus available. This is not the objective of this study where the idea is to combine corpora of similar languages (therefore using both source and target language training sets) to improve final results. Additionally, while in this thesis the selected parsing tool is based on multilingual BERT and training corpora are lexicalized, many related works are based on the usage of delexicalized training sets and parsing algorithms that do not include language models in their structure, even though these models have been proved to be very efficient in terms of dependency parsing results.

Two main tendencies can be observed when typological information is implemented in parsing systems. First, many authors choose to use the provided typological information from typological databases (mostly using them as typological vectors) with a variation concerning the number and type of features selected (e.g.: only syntactical features, syntactical and phonological features, etc.). Secondly, some researchers prefer to use some statistical data concerning word order at the surface level (e.g.: part-of-speech trigrams). Both strategies seem to guarantee an improvement in terms of dependency parsing metrics, however, no study has been conducted in terms of more elaborated quantitative methods regarding word order patterns which is the object of this dissertation.

Finally, it is important to notice that most studies are based on experiments with the objective to improve overall metrics without further analysis in relation with theoretical aspects of syntactic typology. One exception is the work developed by Lynn et al. (2014) in which the authors provided some possible explanations concerning specific dependency relations linking the best candidate for source language (Indonesian) and the target one (Irish).

Hence, this thesis aims to propose a comparative study concerning the usage of language combination (association of lexicalized corpora) with respect to different syntactic typological features and word order patterns extracted quantitatively from annotated corpora in a low-resource scenario. By using correlation measures, the idea is to verify which exact typological method (qualitative or quantitative) best represents what phenomena are more relevant for deep learning algorithms using language models when languages are combined, thus, proposing a more accurate typological language classification.

2.6. Dependency Syntax

Mel’čuk and Polguère in the book “Dependency in Natural Language” (2009) defined in its foreword section the two main assumptions shared by many linguists concerning the dependency approach to syntax:

1. “A sentence is associated with a formal object representing its inner organization which is called syntactic structure”.
2. “The syntactic structure of a sentence is a set of its lexical units linked together by syntactic relations”.

Moreover, four defining properties of the dependency syntax are derived from the assumptions above (Polguère and Mel’čuk, 2009):

1. “Connectedness of the syntactic structure: the syntactic structure forms a unified whole (continuous system of syntactic relations, thus, a connected graph) and no lexical unit is left out of the structure”. The minimal phrase is formed by two elements which are syntactically connected”.
2. “Directedness of syntactic relations: syntactic relations are directed; consequently, phrases have an asymmetric nature, one component dominates the other. Therefore, the phrase behaves rather like its dominant component (head or governor). The representation “$L_1 \rightarrow L_2$” indicates that L_1 is the syntactic governor of the dependent L_2”.

3. "Strict hierarchical organization of the syntactic structure: each lexical unit has one and only one syntactic governor, the exception being the one unit which has no governor at all (i.e.: the top node of the syntactic structure; head of the sentence). Also, the governor controls the linear position of the dependent. A formal consequence of this fact is that the syntactic structure is an acyclic directed connected graph".
4. "Meaningfulness of syntactic relations: to completely specify one sentence, it is not enough to indicate the oriented syntactic relation between two lexical units. The existing syntactic relations must be described by a set of determined labels concerning the dependents. The syntactic relations carry more information than simply the hierarchical organization. They are a bridge between the meaning of the phrase and its surface form. However, syntactic relations do not correspond, in general, to a specific meaning, they correspond to semantic roles (and vice-verse) but these correspondences are not direct nor systematic".

Consequently, combining these assumptions, it is possible to say that "the syntactic structure of a sentence corresponds to a tree whose nodes are labelled with lexical units and whose arcs are labelled with names of the syntactic relations" (Polguère and Mel'čuk, 2009).

Mel'čuk and Polguère (2009) proposed a "Meaning-Text" approach to syntactic dependency. In other words, for them, "syntactic structures are considered within a "Meaning-to-Text" perspective, thus, being perceived as an intermediate structure between the source (semantic non-hierarchized network) and the target (linearly ordered morphological string)". This approach allows the description of language rules which links the semantic, the syntactic and the morphological structures. Furthermore, the formal proximity of the syntactic dependency structure (in the form of graphs consisting of connected lexical units) with the semantic network (graph of connected lexical meanings) facilitate their analysis.

Their theory focusses on the precise "description of syntactic dependency relations (not on the sentence elements connected by them): each syntactic relation is considered as a linguistic unit in its own right. Besides, two levels of syntactic dependency are distinguished: the deep-syntactic structure (closer to meaning) and the surface-syntactic one (closer to the linear sequence of lexical units)" (Polguère and Mel'čuk, 2009). The deep-structure presents only the hierarchization of the full lexical units in terms of meaning and does not reflect directly neither the word order, nor the morphological aspects. "The dependency syntactic structure of a sentence must contain all the information needed to calculate all possible word orders in the sentence. Each individual relation indicates (via syntactic rules) the ordering of its dependent element with respect to the governor" (Polguère and Mel'čuk, 2009).

The dependency approach rejects phrase-structures as a meaning of representing the syntactic organization of a sentence. However, Mel'čuk and Polguère (2009) admit that phrases are necessary at some specific levels, for example, at the deep-morphological level of the sentence representation.

In the article "Dependency in Natural Language" (2009), Mel'čuk claimed that dependencies appear in linguistics via wordforms in an utterance and are linked by them: "one wordform depends on another concerning its linear position and its morphological form" (Mel'čuk, 2009). Thus, dependency is one of the most basic concepts of linguistics due to the fact that the speaker, in order to communicate, must first select the necessary signs (paradigmatic axis), and then arrange them into a linear sequence (syntagmatic axis). The specific arrangement allowing communication to happen is controlled by the dependencies between the signs.

Mel'čuk (2009) defines three major types of dependencies: semantic, syntactic, and morphological. Our focus in this study concerns dependency parsing, therefore, only the author's approach concerning syntactic dependencies will be described. His description is based on nine required notions (Mel'čuk, 2009):

1. "Utterance: an autonomous speech segment. It can appear between two major pauses, constitutes a prosodic unit and is understandable by speakers of the language";
2. "Wordform: it is a minimal utterance, it corresponds to a disambiguated word (or lexeme) in a specific inflectional form. It concerns the unit dealing with dependency syntax";
3. "Phrase: an utterance consisting of one or several wordforms";
4. "Clause: a syntactically organized phrase. It can constitute a simple sentence by itself or can be a constituent of a sentence";
5. "Sentence: a maximal utterance composing a complete communication unit and the upper limit of the dependency analysis";
6. "Semantic predicate, semantic name and argument of a predicate: a semantic predicate is a required meaning which is incomplete without other meanings. A meaning that is not a predicate (actions) is a semantic name (i.e.: objects, beings, substances, and points in space and time). The argument is a meaning that is inserted into an open slot of a predicate";

7. "Inflectional category: it corresponds to an ensemble of opposed inflectional values (also called grammemes). The selection of one of them is obligatory for lexemes of particular classes in some languages";
8. "Syntactics: it specifies the cooccurrence of the sign which is not determined by its signified nor by its signifier. The syntactics of a sign is represented by a set of syntactic features (each one with mutually exclusive values)";
9. "Passive syntactic valence of a lexeme and of a phrase: it is a set of syntactic roles which the lexeme or the phrase can receive in larger constructions (syntactic distribution). Generally, it is defined for major classes of lexemes (parts of speech)".

Beside the above described notions, dependency syntax requires three other assumptions. The first is that a sentence has different representations on four levels: semantic, syntactic (deep and surface), morphological and phonological. Moreover, each representation displays a set of properties of the sentence. Thus, "a sentence representation is a set of formal objects called structures which are responsible for particular aspects of the sentence organization. The second assumption is that a sentence representation appears formally as a labelled graph whose vertices (nodes) represent linguistic units of the corresponding level, while the arcs correspond to the relations between the nodes. The major type of relation between linguistics units in a sentence is dependency" (Mel'čuk, 2009). Finally, the last supposition is that for both syntactic and morphological level, it is possible to distinguish the deep and the surface sublevels of the sentence structure. The deep one is related to the meaning, thus, expressing relevant semantic contrasts, whilst the surface level expresses relevant formal contrasts (Mel'čuk, 2009).

The Semantic Structure (SemS) of a sentence can be defined as "a network whose nodes represent meanings and are labelled with semantemes (lexical meanings). The arcs represent predicate-to-argument relations and are labelled with numbers which identify arguments of the predicate" (Mel'čuk, 2009) (as showed in the Figure 2.1).

Concerning the syntactic structure (SyntS), as explained previously, it holds two subtypes: the deep-syntactic structure (DSyntS) and the surface-syntactic one (SSyntS).

The DSynt is a tree and its nodes are labelled with full lexemes of the sentence and whose arcs (branches of the tree) receive labels corresponding to the deep-syntactic relations (DSyntRels), as exemplified in the Figure 2.2. Mel'čuk (2009) defines 12 different relations across languages: seven actantial, two attributive, two coordinative and one appendititive (or extra-structural).

Figure 2.1. The semantic structure (SemS) correspondent to the English sentence: “For decades, cocoa farming has escaped such problems by moving to new areas in the tropics” (Mel’čuk, 2009).

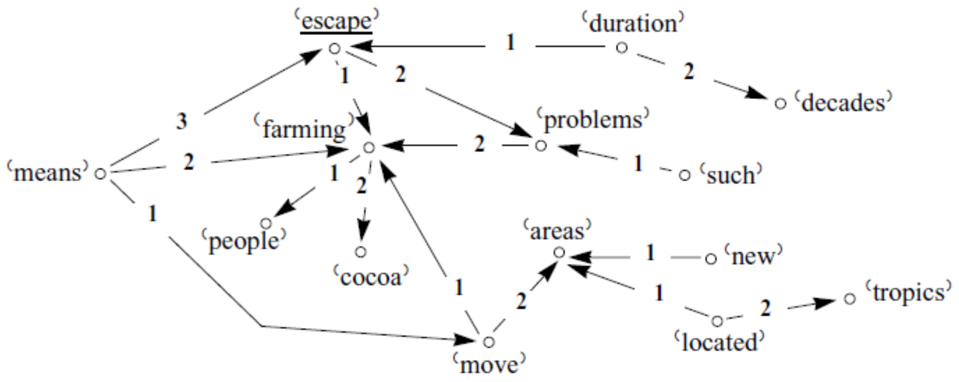
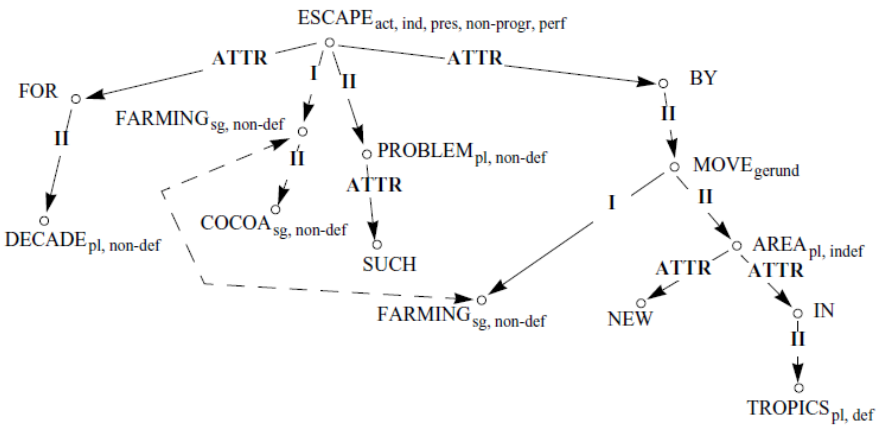

‘escape’
‘duration’
‘decades’
‘problems’
‘farming’
‘means’
‘such’
‘people’
‘cocoa’
‘areas’
‘new’
‘tropics’
‘located’
‘move’
1
2
3

Figure 2.2. The Deep-Syntactic Structure (DSyntS) of the English sentence: “For decades, cocoa farming has escaped such problems by moving to new areas in the tropics” (Mel’čuk, 2009).


ESCAPE act, ind, pres, non-progr, perf
ATTR
I
II
FOR
DECADE pl, non-def
FARMING sg, non-def
COCOA sg, non-def
PROBLEM pl, non-def
SUCH
BY
MOVE gerund
FARMING sg, non-def
AREA pl, indef
NEW
IN
TROPICS pl, def

On the other hand, the SSyntS of a sentence is also a tree and its nodes are also labelled with the lexemes, however, its arcs (branches) are characterised by language-specific surface-syntactic relations (SSyntRels), each one representing a specific construction of the language (as seen in the Figure 2.3).

The dependency parsing task in natural language processing, which is our focus in this thesis, corresponds to the automatic identification of the surface-syntactic structure. While Mel’čuk suggests language-specific surface-syntactic relations, the Universal Dependencies framework propose a set of dependency relations (called “deprel”) for all human languages, allowing cross-lingual comparative studies (de Marneffe et al., 2021).

While the semantic and the syntactic structures are represented by trees, the deep-morphological structure (DMorphS) of a sentence corresponds to a string of lexico-morphological representations of the wordforms respecting the strict linear order. Morphological dependencies do not have a representation as they are not universal, thus, they are computed according to syntactic dependencies. The Figure 2.4 presents an example of a DMorphS.

Figure 2.3. The Surface-Syntactic Structure (SSyntS) of the English sentence: “For decades, cocoa farming has escaped such problems by moving to new areas in the tropics” (Mel’čuk, 2009).

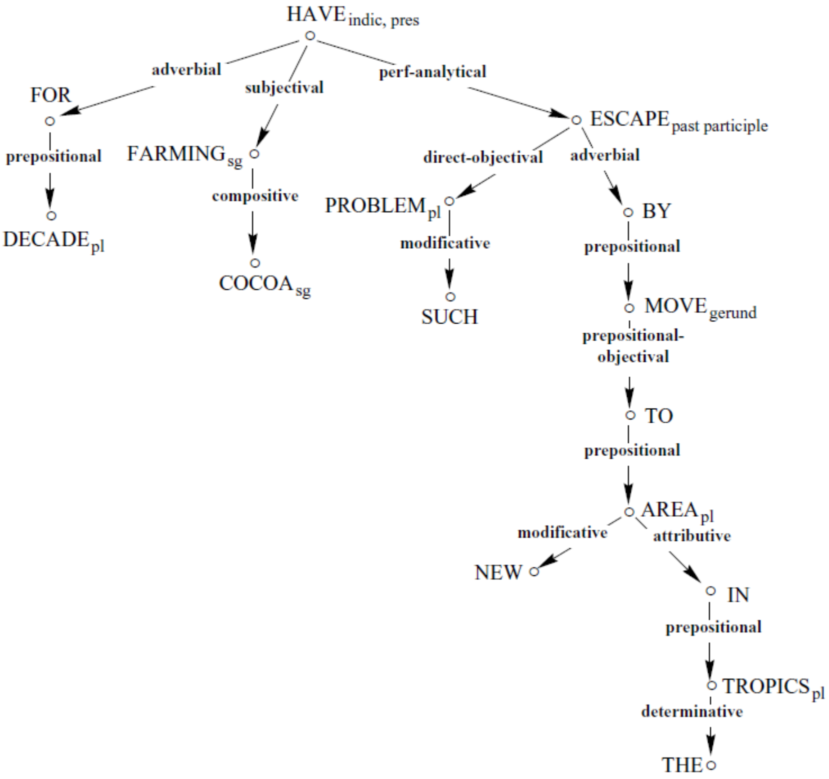

HAVE indic, pres
adverbial
subjectival
perf-analytical
FOR
prepositional
DECADE pl
FARMING sg
compositive
COCOA sg
ESCAPE past participle
direct-objectival
adverbial
PROBLEM pl
modificative
SUCH
BY
prepositional
MOVE gerund
prepositional-objectival
TO
prepositional
AREA pl
modificative
attributive
NEW
IN
prepositional
TROPICS pl
determinative
THE

Figure 2.4. The Deep-Morphological Structure (DMorphS) of the English sentence: “For decades, cocoa farming has escaped such problems by moving to new areas in the tropics” (Mel’čuk, 2009).

FOR DECADE$_{\text{pl}}$ COCOA$_{\text{sg}}$ FARMING$_{\text{sg}}$
HAVE$_{\text{ind, pres, sg, 3}}$ ESCAPE$_{\text{ppart}}$ SUCH PROBLEM$_{\text{pl}}$
BY MOVE$_{\text{ger}}$ TO NEW AREA$_{\text{pl}}$ IN THE TROPICS$_{\text{pl}}$

Consequently, Mel’čuk (2009) defined the three main types of dependency relations between wordforms of a sentence playing a role on the syntagmatic level: semantic dependency [= Sem-D], syntactic dependency [= Synt-D], and morphological dependency [= Morph-D]. In this specific approach, paradigmatic relations such as synonymy, antonymy and derivation are excluded. Moreover, only direct dependencies are analysed, thus, anaphoric relations, inclusion and ordering relation (between wordforms, phrases and clauses), and the communicative dominance relation (between semantic units) are also omitted.

Dependency is defined (Mel’čuk, 2009) as being a non-symmetric relation similar to a logical implication: one element implies the other one, but not inversely. It is represented by an arrow: in “w1 → w2”, w2 (the dependent) depends on w1 (the governor, head, regent or ruler). Mel’čuk (2009) preferred the term “governor” instead of “head” as the latter is inherited from phrase-structure syntax and, therefore, was considered to carry connotations such as constituency which are not needed in dependency syntax. Also, Mel’čuk (2009) distinguished “head” and “governor” designations: for a phrase, its governor is outside of it, while its head in inside (being governor of the other wordforms within the phrase).

Following this presentation of the grounding notions concerning dependencies in natural languages and the introduction about the different types of dependency structures of a sentence, the next paragraphs describe the syntactic level following Mel’čuk (2009) approach which is, as defined by Jurafsky and Martin (2021), the theoretical ground that allowed the development of dependency parsing methods.

The syntactic dependency approach was first formally described by Tesnière (1959), however, dependencies have been used to describe sentences structures since the Antiquity (Mel’čuk, 2009). Arab grammarians in the eighth century, such as Sībawaih, already used the terms governor and dependent to talk about syntax. On the other hand, phrase-structure was first introduced in the early twentieth century and has become dominant due to the Chomskian Transformational-Generative Grammar.

Since the sixties, dependency syntax has become the base of the first computational applications of linguistics such as Hays (1960 and 1964); Lecerf (1960); Fitialov (1962); Iordanskaja (1963); Padučeva (1964); Gaifman (1965); Baumgärtner (1965 and 1970); Marcus (1965); Robinson (1970); and Heringer (1970). And later on, many linguistics theories were built based on the same approach, such as: Case Grammar (Fillmore, 1967 and Anderson, 1977), Meaning-Text Theory (Mel'čuk 1974, 1979, 1988, 1997b), Lexical-Functional Grammar (Bresnan & Bresnan 1982), Relational Grammar (Perlmutter 1983), Word Grammar (Hudson 1984, 1990), Functional Generative Description (Sgall et al. 1986, Petkevič 1995), Lexicase Theory (Starosta 1988), etc.

Mel'čuk approach (2009) towards dependency syntax concerns specifically the representation of the structure concerning sentences, and not a type of dependency grammar (i.e.: rules ensuring the generation and parsing of sentences).

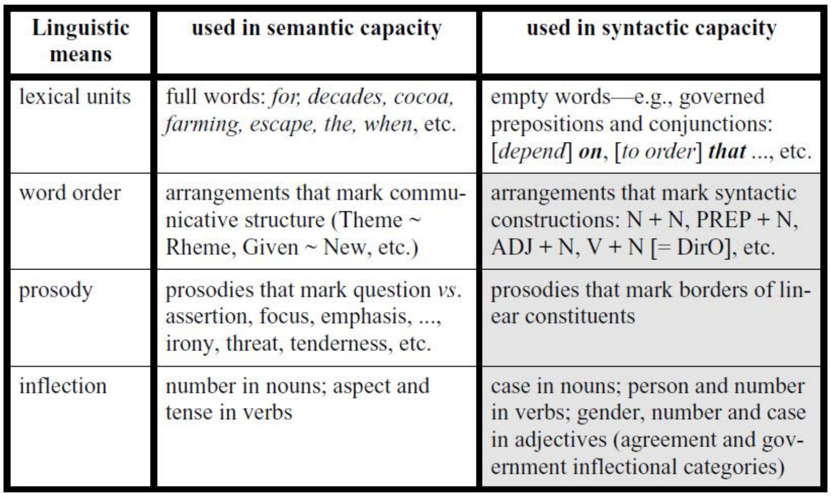
The SyntS of a sentence is defined as being the mediator between its SemS (n-dimensional graph) and its MorphS (1-dimensional graph), meaning that the SynS must be straightforwardly obtained from the semantic network and effortlessly converted to the morphological chain (Mel'čuk, 2009). Also, it must permit inverse processes to be conducted in order to go from text to meaning. Thus, the simplest formal object which satisfies these necessities is a 2-dimensional graph (tree).

Defining the SyntS as a tree means that each arc composing the graph represents an anti-reflexive, anti-symmetrical and anti-transitive binary relations between lexemes (i.e.: Synt-D relations). Mel'čuk (2009) claims that this representation is not only a linguistic tool but that it also, somehow, represents a psychological reality (how sentences are organized in the brain of speakers).

The meaning of a sentence is, therefore, expressed in four types of linguistic means (which have distinguished semantic and syntactic capacities as presented in Figure 2.5).

The SyntS is composed only by lexical means from the syntactic capacity and it defines an order relation: first, where to position a specific wordform in relation to another one (before or after) and, secondly, the details of the positioning of mutual orderings of different wordforms linked to the same governor.

Figure 2.5. Linguistic expressive means and their possible uses (Mel'čuk, 2009).



Linguistic means	used in semantic capacity	used in syntactic capacity
lexical units	full words: *for, decades, cocoa, farming, escape, the, when*, etc.	empty words—e.g., governed prepositions and conjunctions: [*depend*] ***on***, [*to order*] ***that*** ..., etc.
word order	arrangements that mark communicative structure (Theme ~ Rheme, Given ~ New, etc.)	arrangements that mark syntactic constructions: N + N, PREP + N, ADJ + N, V + N [= DirO], etc.
prosody	prosodies that mark question *vs.* assertion, focus, emphasis, ..., irony, threat, tenderness, etc.	prosodies that mark borders of linear constituents
inflection	number in nouns; aspect and tense in verbs	case in nouns; person and number in verbs; gender, number and case in adjectives (agreement and government inflectional categories)

The object of dependency parsing being the surface-syntactic dependencies, the following definitions are focused on the surface-syntactic structure (not on the deep one). The establishment of a SSynt-D relation between two wordforms is based in three main criteria: one criterion for SSynt-connectedness, one concerning the SSynt-dominance, and one determining the specific type of SSynt-D.

Concerning the SSynt-connectedness criterion, it defines "whether two particular wordforms (w_1 and w_2) in an utterance are syntactically directed linked, thus, it is respected when the position of w_1 or w_2 must be defined in relation to the other (w_2 either precedes or follows w_1, and the order is either compulsory or optional in some conditions). The wordform which determines the linear position is not necessarily the governor" (Mel'čuk, 2009).

Furthermore, w_1 and w_2 are considered to have a direct SSynt-D link between them, only if one of the following conditions are satisfied (Mel'čuk, 2009):

1. "If w_1 and w_2 form a prosodic unit (phrase)";
2. "Or, if w_1 and w_2 do not form a phrase, but w_1 is the Synt-head of the phrase and w_2 is also a Synt-head of another phrase concerning other wordforms".

The second criterion concerns the SSynt-dominance, which is defined as: "if w_1 and w_2 are syntactically linked in the utterance, one of them dominates the other" (Mel'čuk, 2009). Thus, the notation "w_1 –synt→ w_2" means that the w_1 is the governor and dominates syntactically w_2.

It also indicates that the governor is responsible of the determination of the external links of the phrase (i.e.: its distribution in the sentence). Moreover, this criterion involves the morphological links between the elements of a phrase and its outside context. Thus, in a phrase (composed by w_1 and w_2), if the governor cannot be recognized concerning the rules related to its distribution in the sentence, a morphological rule can be applied: “if w_1 controls the inflection of wordforms outside the phrase, or if its inflection is controlled by external ones, then, w_1 is the governor and is considered as the morphological contact point” (Mel’čuk, 2009).

If “neither the distribution nor the morphological contact point can define the head, the semantic content of a phrase can be checked: the governor is the wordform with the more defined semantic content. Thus, the Synt-governor is more prominent than its Synt-dependent (syntactically, morphologically, or at least semantically)” (Mel’čuk, 2009). The definition of the head follows the hierarchy: syntactic prominence > morphological prominence > semantic prominence.

It is important to notice that this criterion is language dependent, if X–synt→ Y is attested for a precise language, it does not mean that a comparable construction in terms of part-of-speech will be the same (in terms of governor and dependent) for a different language.

Finally, the third criterion deals with the “definition of the types of syntactic relations (labelled SSynt-dependencies). When two wordforms are directly linked by a Synt-D, the specific type of relation must be described (in terms of surface-syntactic relations or SSyntRels). Each type of SSyntRel is associated with a label which has to be meaningful and has to refer to a family of well-defined syntactic constructions that have an impact on the morphological structure of the sentence. Therefore, SSyntRel can be described as a linguistic sign whose signifier is an ordered pair of lexemes of particular syntactic classes with specific morphological characteristics” (Mel’čuk, 2009).

Furthermore, concerning syntactic substitutability, a SSyntRel must respect what Mel’čuk (2009) calls the “quasi-Kunze property”, an adaptation of “Kunze property” (Kunze, 1972) but in a less strict manner: considering lexemes L(X), L(Y), …, of part-of-speech X, Y, …, which form complete SSynt-configurations[9] Δ(X) and Δ(Y). “A SSyntRel has the quasi-Kunze property if and only if there exists a part-of-speech X for which any SSynt-configuration

---

[9] Subtrees having L(X) and L(Y) as their top nodes and a SSyntRel.

L(X)—SSsyntRel→Δ(Y), replacing Δ(Y) by Δ(X), but not necessarily vice-versa, in any SSyntS does not affect its syntactic well-formedness” (Mel’čuk, 2009).

Additionally, “a given SSyntRel with respect to the same governor can be either non-repeatable or unlimitedly repeatable. It is non-repeatable if, and only if, no more than one branch labelled with a specific SSyntRel can derive from any governor. It is the case for actantial SSyntRels whose dependents are marked only by syntactic means such as subject and direct object. On the other hand, the SSyntRel is unlimitedly repeatable if and only if several branches labelled as such can start from a governor. This corresponds to the cooccurrence (or iteration) test used in linguistic analysis” (Mel’čuk, 2009).

In summary:

- “Synt-D is defined as: the w_2 is syntactically dependent of w_1 via a specific SSyntRel in a particular utterance if the three criteria described above are satisfied”.
- “Synt-D is anti-symmetrical (w_1–synt→w_2 entails ¬(w_1←synt–w_2)), anti-reflexive (a wordform cannot be linearly positioned with respect to itself), anti-transitive (w_1–synt→w_2 and w_2–synt→w_3 entails ¬(w_1←synt–w_3)), and Synt-D must be distinctively labelled and presupposes the uniqueness of the governor”.
- “Synt-D is universal (present in all languages), it appears in all sentences and concerns all wordforms, forming a dependency tree which is a connected graph in which each node depends only on one other node. Only one node does not depend on anything (top node or root)”.
- “The top node does not depend syntactically on anything else and all other wordforms in the sentence depend somehow (direct or indirectly) on it. In most versions of dependency approaches, when a complete clause (or sentence) is involved, the top node is filled by a finite or tensed verb. This fact is not an arbitrary choice, it reflects the verb properties in the sentence which agree with the dependency established criteria”.

Thus, “the linear order of the nodes is not explicitly specified due to the fact that the syntactic dependency description separates the SSynt-links and the ordering of the wordforms. The linear position of wordforms is determined by the SSyntS via a set of the language-specific syntactic rules” (Mel’čuk, 2009).

Beside the properties concerning Synt-governors and Synt-depentents previously presented, these elements also possess three other characteristics which are not compulsory (being absent for some governors and dependents in particular languages): “omissibility, cooccurrence control, and incorporability” (Mel’čuk, 2009).

These characteristics can be described as:

- Omissibility property differentiates governors and dependents: considering the configuration w_1—synt→w_2, the synt-dependent (w_2) can be omitted (without provoking an ellipsis) without posing a problem to the correctness of the SSyntS. In some cases, the Synt-dependent may be compulsory (thus, non-omissible), for example in exocentric constructions such as PREP → N. Also, the synt-governor can sometimes be omissible (e.g.: the English subordinate conjunction “that” in the sentence “He knows that she is in town” which is syntacticaly equivalent to “He knows she is in town”).
- Cooccurrence (or subcategorization) control is the property which states that the governor w_1 is subcategorized by the governor w of the whole sentence, thus w must consider some properties of w_1 and not of its dependent. This property can also be described as: the governor w_1 tends to subcategorize for its dependent w_2 (w_1 tends to determine the choice of w_2).
- Incorporability corresponds to two possible phenomena regarding the orientation of the dependency in a configuration (w_1—synt—w_2): internal and external. Concerning the internal incorporability, if “w_2 can be incorporated into w_1 (and not vice-versa), then w_1 is the governor of w_2. On the other hand, the external incorporability property claims that if w_1 (or both w_1 and w_2, but not w_2 alone) can be incorporated into the governor of the whole phrase, then w_1 is the governor of w_2” (Mel’čuk, 2009).

Other non-obligatory properties of governors include class size (governors belong to larger word-classes than its dependents), versatility (a governor occurs in a larger diversity of syntactic environments), frequency (a particular governor is less recurrent than a specific dependent), etc. However, many languages present exceptions concerning these non-compulsory properties, they can only be considered for heuristic reflexions.

Mel’čuk (2009) also defined “three subtypes of syntactic dependency: complementation, modification and coordination”. The first two being particular cases of subordination. The figure 2.6 presents the structure of these major subtypes.

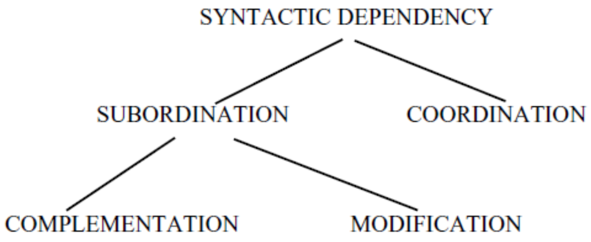
Some dependencies do not belong to any of these subclasses at the SSynt-level, thus, a fourth major subtype (ancillary) is needed to link specific syntactically-induced wordforms (structural words, chunks of idioms, etc).

Thus, for w_2 –synt→ w_1, these subtypes can be defined as:

- Complementation: w_2 is a complement of w_1 (or a Synt-actant) if w_2 is also depends on w_1 semantically.
- Modification: w_2 is a modifier of w_1 (or a Synt-attribute) if w_1 depends on w_2 semantically.
- Coordination: w_2 is a conjunct of w_1 if, and only if, w_2 depends syntactically on w_1, and neither of these wordforms depends semantically on the other (but are dependent of semantemes such as "and", "or", etc.).

Constructions containing complementation are called exocentric, while the ones containing modification and coordination are named endocentric (Mel'čuk, 2009).

Figure 2.6. "Major subtypes of syntactic dependency" (Mel'čuk, 2009)



2.7. Dependency Parsing

2.7.1. Historical background

According to Jurafsky and Martin (2021), the development of dependency parsing was mostly influenced by the following dependency grammar frameworks: Meaning-Text Theory (MTT) (Mel'čuk, 1988), Word Grammar (Hudson, 1984), and Functional Generative Description (FDG) (Sgall et al., 1986). The main differences in these works are related to the approach concerning morphological, syntactic, semantic, and pragmatic factors, as well as, different usage of multiple layers of representation and set of dependency relations. These contemporary theories have been developed under the influence of Tesnière (1959) but their bases were set by the Indian grammarian Pāṇini between the 7th and 4th century B.C., and by the Greek grammar traditions.

They have been used in the 1960's together with constituency parsing in works such as the machine translation project developed by the RAND Corporation led by David Hays (Pierce, 1966). Nevertheless, the main usage of these formalisms started in the late 1990's with the creation of dependency-based treebanks (annotated text with dependency parsing information) and the development of data-driven approaches such as: the deterministic word by word approach which was the base for transition-based methods (Covington, 2001), for the the shift-reduce paradigm, and for the usage of supervised machine learning introduced by Yamada and Matsumoto (2003) and Kudo and Matsumoto (2002).

These works have been followed by the deterministic transition-based approach to dependency parsing defined by Nivre (2003), who continued working on this domain developing different transition systems and training methods. Furthermore, the application of a graph-based maximum spanning tree approach to dependency parsing was introduced by McDonald et al., 2005, while a neural classifier was developed in 2016 by Kiperwasser and Goldberg.

In terms of treebanks which are the source of data for training and evaluating dependency parsing, some examples are: the Prague Dependency Treebank project (Hajič, 1998) for Czech language; and the Universal Dependencies (de Marneffe et al., 2021) which is a framework for dependency annotation across languages and that will be described in details in section 4.1.1. as it is the source of the data used in this study.

Dependency parsing formalism has also been incentivized by the several Conferences on Natural Language Learning (CoNLL) which have been organized together with a series of shared tasks related to dependency parsing over the years (e.g.: Buchholz and Marsi, 2006, Nivre et al. 2007, Hajič et al. 2009, etc.) focusing on parser robustness and evaluation of dependency parsing performance also concerning non-canonical language forms (e.g.: spoken and social media texts).

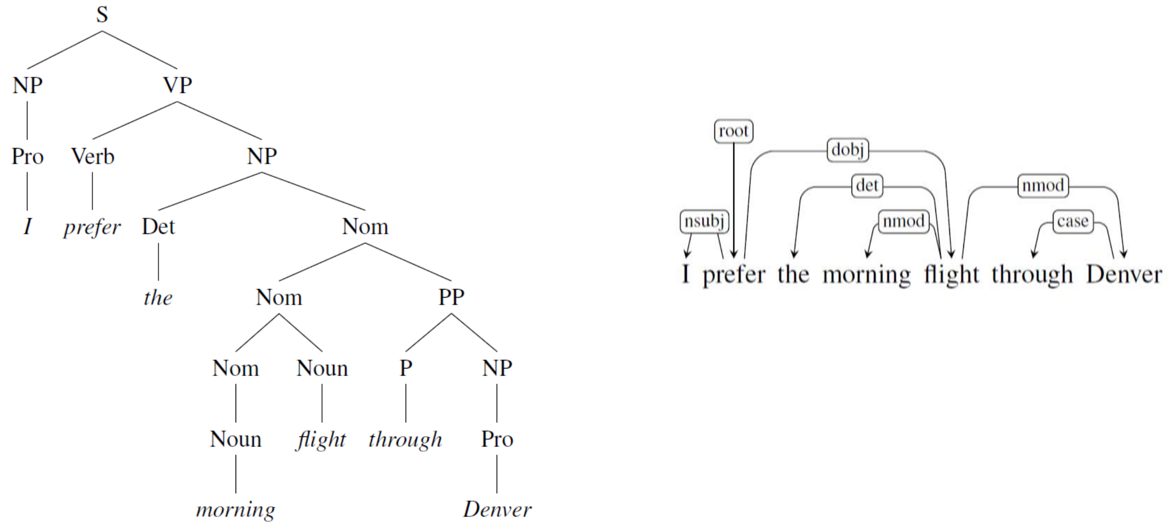
2.7.2. Dependency parsing formalisms

Dependency parsing is the natural language processing task concerning dependency grammar formalism. "In dependency formalisms, phrasal constituents and phase-structure rules do not play a direct role as the syntactic structure of a sentence is totally described using directed binary relations between the words" (Jurafsky and Martin, 2021).

The figure 2.7 shows on the right the constituent analysis for the sentence “I prefer the morning flight through Denver”, while on the left side the dependency analysis of this sentence is presented.

Thus, in dependency analysis, the relation between two words is illustrated with a directed and labelled arc going from the head to the dependent. It is a typed dependency structure as the labels come from a fixed inventory of syntactical relations (the set of possible labels may vary in different treebanks but is intended to be unique for the corpora from the Universal Dependencies collection). The word labelled as “root” marks the head of the entire structure of the sentence (Jurafsky and Martin, 2021).

Figure 2.7. Constituency analysis (on the left) and dependency analysis (on the right) of the sentence “I prefer the morning flight through Denver” (Jurafsky and Martin, 2021).



One advantage of dependency analysis when compared to the constituency one is that it directly displays relevant information which are often more difficult to be decrypted in constituency analysis, especially for more complex phrase-structures. Furthermore, dependency grammars do not consider directly word order information, and it can simplify the treatment of languages with relatively free word order. For constituency grammars, separate rules for each possible place of the words have to be defined.

As previously mentioned, the dependency structure is “formed by binary dependency relations from the traditional linguistic notion of grammatical relation between words. In dependency-based approaches, the head-dependent relationship is presented by links between the heads and the words that are immediately dependent on them, thus, without the usage of constituent structures” (Jurafsky and Martin, 2021) (which are present in constituency analysis where the head word is the central organizing word of a larger constituent).

The translation process from constituent to dependency structures involves two steps: first, all head-dependent pairs are recognized, and, secondly, the right dependency relation for each one of them is assigned. Xia and Palmer (2001) developed an algorithm capable of transforming constituent trees (with annotated grammatical relations, as is the case of Penn Treebank, Marcus et al., 1993) to dependency structures. However, this algorithm fails in representing non-projecting structures or in integrating morphological information when necessary. The lack of internal structure inside noun-phrases are usually are solved by attributing the label “flat” to the relations occurring between words inside these nodes. Consequently, manual annotation of treebanks or automatic annotation followed by manual correction are still the most efficient way to obtain dependency corpora (Jurafsky and Martin, 2021).

As the links between heads and dependents are characterized in terms of grammatical relations (or functions) that the dependents play with respect to the head, a dependency structure can be represented as a directed graph:

$$(2.1)\ G = (V, A)$$

Where V is a set of vertices and A, a set or ordered pairs of vertices A (which are most commonly named as arcs). Usually, the set of vertices corresponds to the set of words in a sentence (but can also include punctuation, stems, and affixes in morphologically rich languages). The set of arcs (A) is formed by the head-dependent information together with the grammatical function between elements of V (Jurafsky and Martin, 2021).

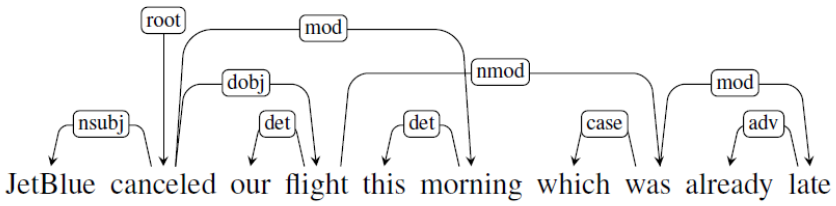
Most of formalisms concerning dependency grammars used in computational linguistics (such as Universal Dependencies framework) adopt the concept that a dependency tree is a directed graph that satisfies the following conditions (Jurafsky and Martin, 2021):

1. “There is a single designated root node that has no incoming arcs”.
2. “With the exception of the root node, each vertex has exactly one incoming arc”.
3. “There is a unique path from the root node to each vertex V”.

Consequently, the dependency structure is a connected graph: from the “single root node one can follow a unique and direct path to each of the words in the sentence, and each word (with the exception of the root) has a single head” (Jurafsky and Martin, 2021).

In dependency grammar formalisms, “an arc from a head is considered projective if there is a trail from the head to each word that lies between the head and the dependent in the sentence” (Jurafsky and Martin, 2021). If all arcs are projective, the tree is defined as projective. However, specially concerning languages with flexible word order, non-projective trees can occur. It is the case of the tree presented in the figure 2.8, where the arc going from the word “flight” to its modifier “was” (labeled as “nmod”) is a non-projective one as there is no path from the head intervening the words “this” and “morning”.

Figure 2.8. Example of a non-projective tree. Dependency analysis of the sentence “JetBlue cancelled our flight this morning which was already late” (Jurafsky and Martin, 2021).


root
mod
dobj
nmod
mod
nsubj
det
det
case
adv
JetBlue canceled our flight this morning which was already late

Graphically, it means “that a dependency tree is projective if it can be drawn with no crossing edges” (Jurafsky and Martin, 2021). The notion projectivity can be problematic in cases where dependency treebanks are derived from phrase-structure treebanks by using head-finding rules as, in these cases, all the generated trees are projective. Moreover, some systems based in transition-based approaches can only produce projective trees and, thus, present errors when non-projective examples are to be examined. This fact led to the development of graph-based parsing methods which can deal better with such specific structures.

2.7.3. Dependency parsers

In this subsection, the two main computational strategies that are used in dependency parsing tools are described: transition-based and graph-based.

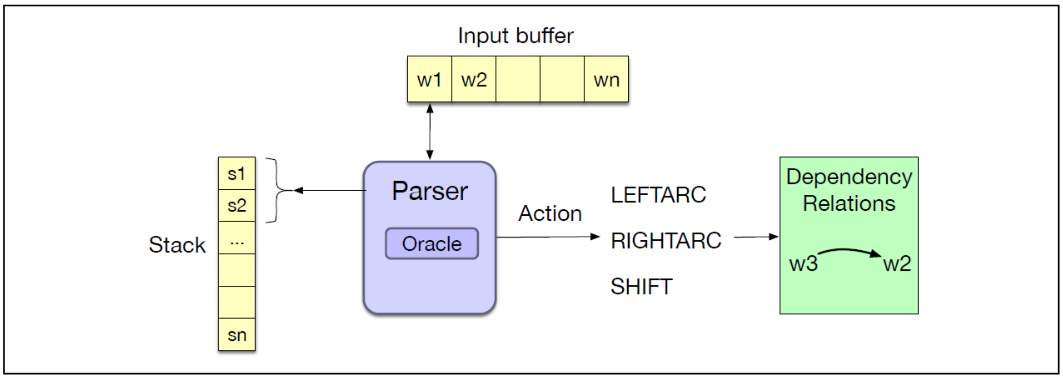
Concerning the Transition-Based Dependency Parsing, this approach is derived from the algorithm that was first developed for the analysis of programming languages (Aho and Ullman, 1973). Essentially, it is composed by a stack on which the parser is built, a buffer of tokens to be parsed, and a predictor (or oracle), in which the action of assigning the right head and dependency relation is done. The main architecture of transition-based parsers is schematized in the Figure 2.9.

In this method, the parser examines each word (going from left to right), and sequentially shifts items from the buffer (input) onto the stack.

For each word, the top two elements of the stack are verified and the oracle decides which transition to apply (Jurafsky and Martin, 2021):

- "Assign the current word as head of a previous seen word"
- "Assign some previously seen word as the head of the current word"
- "Postpone the decision, storing the information for later decision"

Figure 2.9. Basic architecture of transition-based parsers (Jurafsky and Martin, 2021).



These transitions correspond to concrete possible operations that are applied to the top two elements of the stack (Jurafsky and Martin, 2021), and they are known as the arc standard approach (Covington, 2001, Nivre, 2003):

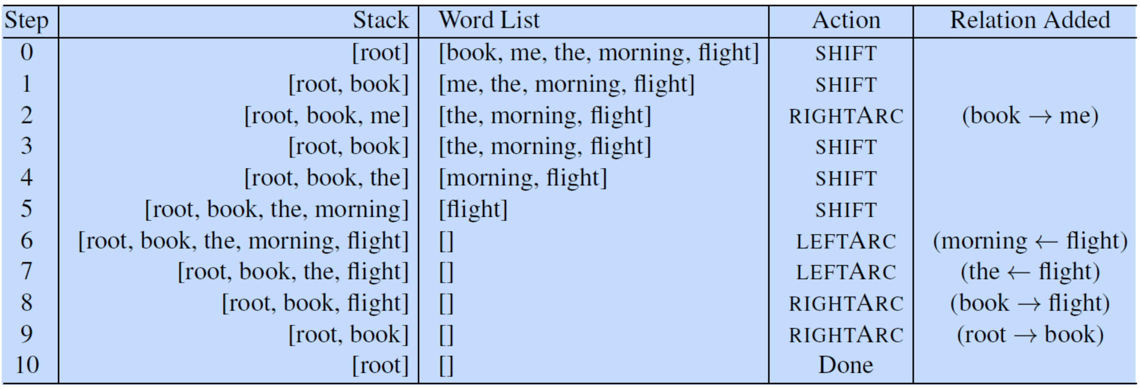
- LEFTARC: Assert a head-dependent relation between the top-word (head) and the second word (dependent) of the stack, removing the second word from it. This operation cannot be applied if the second word is the root.
- RIGHTARC: Assert a head-dependent relation between the second word and the top one, removing the top-word from it.
- SHIFT: Skip the input-word, storing it in the stack.

Once one element is assigned to a head, "it is removed from the stack and is no longer available for further processing. LEFTARC and RIGHTARC operators can be parameterized to include the specific type of dependency relation, multiplying the final set of possible operations that can be applied for each word" (Jurafsky and Martin, 2021). This approach is a straightforward greedy algorithm: the oracle passes through the sentence only once, deciding, for which word, which action is applied. In the end of the process, only a single parsed tree is provided, meaning that if an erroneous relation is assigned at one point, it is reviewed throughout the other steps of the process.

The Figure 2.10 exemplifies the set of ten operations applied to build the dependency tree of the sentence “Book me the morning flight”.

The decisions are made by the oracle are obtained mostly via machine learning methods trained with annotated data. The usual corpora containing dependency parsing trees (such as the data provided by Universal Dependencies) do not provide explicitly the set of transitions that the system has to apply, only the tokens and the associated head and dependency label are available. Thus, the model must use classifiers (usual neural ones) that represent the possible configurations of the dependency trees associated to transitions (and operations) using embeddings. During the training step of the oracle, the algorithm tries different set of operations learning in a deterministic way which are the correct ones to apply for each pair of tokens (Jurafsky and Martin, 2021).

Figure 2.10. Ensemble of operations concerning the transition-based approach for dependency parsing of the sentence “Book me the morning flight”, in this example, the type of dependency relation is not detailed (Jurafsky and Martin, 2021).



Step	Stack	Word List	Action	Relation Added
0	[root]	[book, me, the, morning, flight]	SHIFT	
1	[root, book]	[me, the, morning, flight]	SHIFT	
2	[root, book, me]	[the, morning, flight]	RIGHTARC	(book → me)
3	[root, book]	[the, morning, flight]	SHIFT	
4	[root, book, the]	[morning, flight]	SHIFT	
5	[root, book, the, morning]	[flight]	SHIFT	
6	[root, book, the, morning, flight]	[]	LEFTARC	(morning ← flight)
7	[root, book, the, flight]	[]	LEFTARC	(the ← flight)
8	[root, book, flight]	[]	RIGHTARC	(book → flight)
9	[root, book]	[]	RIGHTARC	(root → book)
10	[root]	[]	Done	

Classifiers in transition-based tools can be built using feature-based algorithms or can be neural, in this case built with embedding features. Feature-based classifiers use information such as word forms, lemmas, part-of-speech, and morphosyntactic in addition to the information concerning the dependency structures. Most important features for the decision-making process come from the top levels of the stack, the words near the front of the buffer.

On the other hand, in neural classifiers, the sentence is passed through an encoder, the representation of the top 2 words in the stack is concatenated with the first word of the buffer and the result is presented to a feedforward network which, then, predicts the transition to be applied (Kiperwasser and Goldberg, 2016; Kulmizev et al., 2019).

The arc standard approach “can be replaced by the arc eager one: the main difference is that in the standard method, operators are applied to the top two elements of the stack, then the front of the buffer and dependents are removed from the stack as soon as a head is assigned to them, whilst in the arc eager approach, the operators act only at the top level of the stack and the front of the buffer, and the dependent is added to the stack, not removed, thus being available to serve as head of other words” (Jurafsky and Martin, 2021). This improvement allowed the development of tools using transition-based methods for dependency parsing of non-projective structures (Nivre, 2009) and dependency parsers for multilingual texts (Bhat et al., 2017).

Finally, beside the improvement provided by the arc eager method, transition-based systems can also profit from the beam search strategy which can be applied to supress the limitation concerning the single pass through the sentence. Beam search methods use a breadth-first search strategy with a heuristic filter that reduces the search frontier to maintain its borders in a fixed-size width. This way, the model no longer chooses the best transition operator at each interaction, instead, all possible operators are applied and each possible state is stored in an agenda of limited size (defined by the beam width). The iterations continue until there are only final states left in the agenda (yet, new states can be added to the agenda if they are estimated better than the ones present in it, therefore, removing the worst ones). The scoring of states is applied throughout the whole process to define the set of elements in the agenda, and in the end of the task, to select the final dependency structure provided by the model (Jurafsky and Martin, 2021).

The other possible possible strategy for dependency parsing is the Graph-Based one. This approach tends to be more accurate than transition-based algorithms especially for long sentences and in cases where the head is distant from the dependent (McDonald and Nivre, 2011). It can also produce non-projective structures without the need of implementation of other complementary strategies. The main reason for these advantages is that decisions are made observing the whole syntactic tree while transition-based tools make decisions locally with a greedy approach. “Graph-based dependency parsers are built using graph theory, it means that these tools search through the space of all possible solutions (G_s) concerning a sentence (S) and check which tree maximizes some score (T)” (Jurafsky and Martin, 2021):

$$(2.2)\ T(S) = argmaxScore_{t \in G_S}(t, S)$$

Usually, the score is considered edge-factored, meaning that the overall score for a tree is the sum of each score of the comprising edges (e):

$$(2.3)\ Score(t,S) = \sum_{e \in t} Score(e)$$

Therefore, the system assigns first a score for each edge and, then, finds the best tree given the score of all possible edges. The assignment of a score for each edge can be done using feature-based algorithms (i.e.: the final score being a sum of the available features). Like in the case of transition-based parsers, the most common used features include: wordforms, lemmas, part-of-speech, and dependency relations information such as the type of relation, its direction (left or right), and the distance between head and dependent. To learn the set of weights corresponding to each feature that allows the system to evaluate the set of possible trees, the chosen algorithms must determine the weights that guarantee that the highest scores are associated with best solutions.

One possible method is to use inference-based learning combined with the perceptron learning rule where a parsing tree (from the training set) is inferred using a set of initially random values of weights. If the resulting parse corresponds to the solution, weights are not changed. If not, the values corresponding to the features of edges mistakenly assigned are decreased by a small amount. This operation is done for all sentences of the training data until the weights converge.

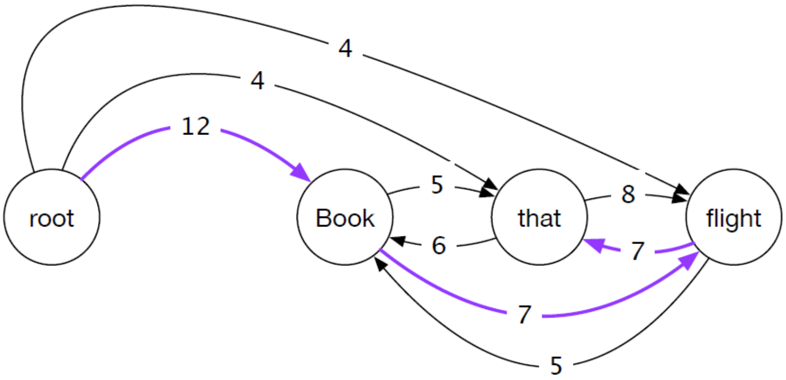
Yet, the state-of-the-art concerning graph-based parsers is based on neural networks. In this approach, first sentences are encoded, then, their representations are processed through a network that estimates the score of the edges between them. It is the case of the algorithm developed by Dozat and Manning (2016) that was first built to identify heads and dependents, but that can be easily transformed in a way to provide the type of relation by adding a second classifier of the same type (trained with dependency type information), allowing the system to provide the most probable label for each edge assigned in the first step (Jurafsky and Martin, 2021).

Once the weights are defined, given a sentence S, the system generates a direct graph G which is totally connected. “In this representation, the vertices are the input words and the directed edges represent all possible head-dependent assignments (a root node is also added with outgoing edges to every other vertices of the sentence)” (Jurafsky and Martin, 2021). Each subset of G covering all the vertices (starting from the root) is a spanning tree.

Consequently, the way to find the best dependency tree correspond to identifying the maximum spanning one over G. The Figure 2.11 illustrate the directed graph for the sentence “Book that flight” with the maximum spanning tree corresponding to the path showed in purple (Jurafsky and Martin, 2021).

In the generated graph, for each spanning tree, the vertices have only one incoming edge. Also, the absolute values of the edge scores are not critical to determining its maximum spanning tree as what is really considered are the relative weights of edges arriving in each vertex. Basically, what is done by the algorithm is finding the maximum spanning tree via a greedy edge selection, followed by a re-scoring of the edge costs, and by recursive clean-up phases which is responsive for dealing with possible cycles (vertex with two entering edges) coming from the greedy strategy.

Figure 2.11. Graph with possible head and dependent assignments for the sentence “Book that flight”.



2.7.4. Dependency parsing evaluation

The way to evaluate dependency parsers requires testing the developed systems on test sets to check how well they perform the task of assigning the dependency trees. One possible metric corresponds to the calculation of the exact matches, corresponding to the percentage of sentences parsed correctly. This method is not usually performed as scores tend to be low.

Instead, the most common method for evaluating dependency parsers consider the overall accuracy at the token level.

Two possible values can be calculated:

- Labelled attachment score (LAS): this metric refers to percentage of the correct assignment of a word to its head along with the correct dependency relation.
- Unlabelled attachment score (UAS): in this case, only the head assignment is verified.

Therefore, as LAS metric accounts for more information, values can only be inferior or equal to UAS (being equal only if all dependency relations are well identified). When more than one sentence is concerned, it is possible to calculate the macro or the micro-average of these metrics:

- Micro-average: UAS and LAS are calculated for the ensemble of the tokens in the test set.
- Macro-average: UAS and LAS values are measured for each sentence and the final metrics are obtained calculating their average.

For example, considering two sentences for which:

- Sentence 1: 9 out of 10 tokens are well analysed (head and label)
- Sentence 2: 15 out of 45 tokens are well assigned (head and label)

Sentence 1 has a LAS equivalent to 90.00, while sentence 2 has 33.33. The LAS micro-average in this case would be 43.63, while the macro-average, 61.67.

The CoNLL 2017 Shared Task (Zeman et al., 2017) used LAS (micro-average) as the main criterion for the evaluation of dependency parsers as it is the most usual approach for dependency parsing evaluation. Moreover, Choi et al. (2015) developed a tool that allows not only the calculation of UAS and LAS (micro-average) but also the analysis of the accuracy concerning each label, and in relation to the distance between head and dependent. Thus, the ensemble of results presented in this thesis concerns micro-average scores.

The UAS and LAS metrics seem to be a good way for evaluating the performance of dependency parsing systems when languages are studied individually. For cross-lingual studies, the MLAS metric is more recommended (as seen in the CoNLL 2018 shared task, Zeman et al., 2018). MLAS stands for “Morphology-aware Labelled Attachment Score”, and combines the evaluation of the correctness of dependency parsing labels (heads and relations, LAS) with part-of-speech and morphological features. The difference from LAS is that some types of relations are not directly evaluated, instead, the words comprised by them are not considered as independent words and are treated as features of the content words they belong to[10].

The system-produced[11] word (S) is considered correct if all the conditions below are respected:

- It is aligned to the correspondent gold-standard word (G);
- Its head is aligned with G's head;
- There is a match between the list of "content relations"[12];
- The part-of-speech tags of words S and G are the same;
- The morphological features (from a specific list[13]) of S and G are the equal (other possible features not present in the list are ignored);
- "Functional children" of a node are the child nodes attached via one of the "Function relations"[14]. Thus, when child nodes are present:
    - Each functional child of S must be aligned to a child of G and vice-versa.

And, for each and every pair of aligned functional children:

- The universal part of the label of their relation must be the same;
- The part-of-speech tags must be equal;
- The values of listed morphological features must match analogically to how the features of the content words are compared.

Precision is, then, calculated using the number of correct words divided by the total number of system-produced content words (which are those attached via a content relation), while, recall value is obtained by dividing the number of correct words by the total number of gold-standard content words.

Due to the fact of comprising more elements to be evaluated, MLAS values are always inferior to LAS. As mentioned before, it is useful for cross-lingual studies for the comparison of languages with different degrees of analyticity and syntheticity. When only LAS is considered, synthetic languages are jeopardized as information is condensed in fewer tokens when compared to analytical ones. Thus, one dependency parsing mistake has a bigger impact in UAS and LAS metrics for synthetic languages such as Finnish and Hungarian. In this thesis, both LAS (due to its mainstream usage in dependency parsing studies) and MLAS (due to its relevance for cross-lingual comparison) are considered.

---

[10] The official script for the LAS and MLAS evaluation is available at: https://github.com/UniversalDependencies/tools/blob/master/eval.py

[11] A parsing system in the shared task was also responsible for word segmentation; hence, it was not guaranteed that the sets of system-produced and gold-standard words would be identical.

[12] "Content relations": nsubj, obj, iobj, csubj, ccomp, xcomp, obl, vocative, expl, dislocated, advcl, advmod, discourse, nmod, appos, nummod, acl, amod, conj, fixed, flat, compound, list, parataxis, orphan, goeswith, reparandum, root, and dep.

[13] "Morphological features": PronType, NumType, Poss, Reflex, Foreign, Abbr, Gender, Animacy, Number, Case, Definite, Degree, VerbForm, Mood, Tense, Aspect, Voice, Evident, Polarity, Person, Polite.

[14] "Function relations": aux, cop, mark, det, clf, case, and cc.

3. Objective and Hypotheses of Research

The aim of this thesis is to propose an extensive analysis of the influence of syntactic typological features when languages are combined to improve dependency parsing results for disfavoured languages in terms of annotated data. Thus, the main outcome of this research is an optimized method to classify languages regarding syntactic typology.

More precisely, the goal is to propose a new typological classification of the 24 official European Union (EU) languages in a scenario also containing non-EU ones. The EU languages will be classified via optimized quantitative automatic methods based on the identification of syntactic rules from annotated corpora. These typological analyses will also serve as the base for multilingual corpora association to improve NLP dependency parsing tools in low resource scenarios.

As it has been presented in the previous section, many studies have proven the efficacy of combining corpora from different languages to improve LAS and UAS metrics. One of the limitations of the articles proposing this type of method is their emphasis on using solely the information provided by typological databases. This strategy has proven to be efficient in different scenarios, however, for each analysed syntactic typological feature, only the most frequent attested order is considered for each language, thus, less frequent phenomena are ignored. Therefore, it raises the question of how languages can be compared when less frequent phenomena are also integrated in the analysis. To overcome this limitation, we propose to compare languages using quantitative methods (corpus-based typology) which allow the extraction of word order patterns (and their respective frequency) from annotated corpora and relate the obtained classifications to the state-of-the-art ones, which are obtained from the information of the abovementioned databases. Moreover, the new classifications are compared with the classic genealogical classification, which has also been used in experiments concerning dependency parsing improvement.

While most studies regarding quantitative typology focus on different methods of language comparison in terms of language complexity and on different approaches to prove typological universals, there has been no specific examination of how specifically corpus-based typology can be used in natural language processing tasks. Thus, what is proposed here is an application of corpus-based typological approaches which are tested for the specific task of automatic syntactic annotation. The aim is to understand which possible syntactic features play major roles when models are trained using deep learning methods. The idea is to use correlation measures calculated between language distances (obtained via the extraction from corpora of syntactic patterns and their frequency) and the improvement (or deterioration) in terms of dependency parsing metrics (LAS and MLAS).

Quantitative word order analysis has been applied in some studies concerning multilingual corpora association, however, most of them use simply the quantification of part-of-speech trigrams and do not present a large variety of corpus-based typological methods. Furthermore, the main objective of the studies concerning dependency parsing improvement is to parse languages without any annotated data (thus, called unannotated languages). Therefore, there is a lack of understanding of how word order analysis and corpora combination can be applied for low-resourced languages (with a small amount of annotated data) which also suffer from low scores in terms of LAS and MLAS.

It is also pertinent to mention that most of the research which has been conducted regarding methods for language association uses machine and deep learning algorithms that are not associated with language models. However, the state-of-the-art concerning dependency parsing algorithms is now based on methods using this type of resource (Otter et al., 2019), as is the case of UDify (Kondratyuk and Straka, 2019) and UDPipe 2.0 (Straka, 2018). Therefore, the focus of this thesis is the definition of corpus-based typological methods which are pertinent for state-of-the-art dependency parsing algorithms. Moreover, the strategy here is to use lexicalized corpora (and not delexicalized ones that have been tested in previous works) to facilitate further implementation of the methods as many tools require lexicalized corpora in their training step. This choice also enables some analysis on how independent from the lexicon is the learning phase of dependency parsing relations.

The objective of this thesis is also to confront the application of syntactic typology in dependency parsing experiments with its theoretical frame. Greenberg (1963) and Dryer (1992) focused on the position of the verb and object to describe universals and correlations pairs respectively, while Hawkins (1983) based his universals on a higher variety of components based on the “Head and Dependent Theory” (HDT). As described in the previous section, Hawkins’ method was further criticized by Dryer (1992) because of the existing contradictions among different authors regarding some head and dependent relations. However, Dryer himself (1992) considered that his branching theory can be more elegantly described in terms of heads and dependents if the inconsistencies are surpassed, which is precisely the case when dealing with dependency parsing using a unified annotation framework (such as Universal Dependencies).

Therefore, the theoretical background is challenged using different quantitative methods that allow languages to be compared using: verb and object relative position patterns, head and dependent surface ordering, and more complex deeper structures and possible properties between components (e.g.: if the presence of a particular element excludes another one). Moreover, these new approaches are compared to the state-of-the-art method of language classification using typological databases and genealogy.

Therefore, it is possible to summarize the objectives of this thesis in the two following pairs of research question (RQ) and hypothesis (H):

RQ1: Is it possible to typologically classify European languages in terms of syntactic rules quantitatively extracted from annotated corpora?

H1: A new way of classifying European languages can be achieved by determining the syntactic typological distance between languages using statistical information obtained from annotated corpora which will also allow the identification of syntactic features that have not been considered so far in qualitative typological analysis.

RQ2: Is the new quantitative typological classification of languages a better way of selecting corpora to improve deep learning systems which perform automatic syntactic annotations from raw text?

H2: The typological classification using the quantitative syntactic typological distance between languages is an efficient way to identify related languages whose corpora can be combined to optimize the performance of deep learning tools in terms of automatic syntactic annotation.

To answer the research questions and test the hypothesis, the idea is, first, to test different corpus-based typological strategies with a multi-lingual collection containing 20 parallel corpora (with 10 European Union languages). This analysis, in a more controlled scenario, allows a strict and precise comparison between the possible strategies and the identification of possible bias in each of them. In the second step, all these 20 languages are associated to train a deep-learning tool and the final LAS and MLAS metrics are, then, correlated with the proposed corpus-based classifications. The typological method which correlates the best with the dependency parsing results is used, in the third step, to classify all the 14 other EU languages which are not part of the parallel collection used in the first two steps. Furthermore, with this final typological classification, it will be possible to check how much this strategy improves parsing metrics for some EU low-resourced languages.

The first hypothesis is developed, using the parallel data, in section 4 where the different corpus-based methods are fully described and the different language classifications are presented and analysed. Then, in section 5, the dependency parsing results of the parallel corpora combinations are displayed, together with the correlation examination which is the main criteria for determining which classification method is the most pertinent when dealing with multilingual corpora association for improving automatic syntactic annotation concerning low-resourced languages. Finally, in section 6, the most optimized methods are applied to all the other EU languages for which no parallel data is available, thus, providing a classification of 34 worldwide languages (24 EU and 10 non-EU) which allows the improvement of LAS and MLAS for EU languages with lack of annotated data in terms of dependency parsing.

4. Syntactic Typological Classifications

As mentioned in the previous section, the first research question and hypothesis concern the development and analysis of different quantitative typological approaches involving syntactic structures and patterns extracted for annotated corpora in comparison to existing language classifications such as genealogical and the ones obtained from syntactic features available in typological databases.

In this section, first, the material that has been selected for this study regarding datasets and software is described. Then, the methodology adopted concerning the corpus-based typological approaches is detailed. After that, the obtained results (i.e.: possible syntactic typological classifications) are presented. Finally, a comparative analysis of the different new strategies is conducted in relation to the more traditional typological approaches.

4.1 Language Resources and Tools

As pointed out by Levshina (2022), the choice of the type of data to be used in corpus-based typological studies is crucial. For this thesis, the Parallel Universal Dependencies (PUD) collection has been chosen as it provides a set of parallel corpora of 20 different languages: each corpus is composed of 1,000 sentences annotated following the Universal Dependencies framework. It has the advantage of having, for each language, the same size (in terms of sentences) and, also, equal semantic content, thus, enabling the focus to be on the cross-lingual syntactic comparison.

Due to the importance of the data selection step, in the following sub-sections, the Universal Dependencies framework will be detailed, followed by the characterization of the PUD dataset. The dataset description is followed by a detailed presentation of two tools that will be used throughout this thesis: the lang2vec Python library (based on URIEL typological database, Littell et al., 2017) and the MarsaGram software (Blache et al., 2016).

4.1.1 Universal Dependencies

Universal Dependencies[15] (UD) is a framework developed by an open community whose aim is to develop cross-linguistically consistent treebank annotation by providing a robust guideline for annotation of grammar (i.e.: parts of speech, morphological features, and syntactic dependencies) that can be applied across different human languages, while still allowing language-specific extensions when necessary (de Marneffe et al., 2021). Moreover, this community is also responsible for creating morphosyntactically annotated corpora, continually feeding a growing repertoire of data.

---

[15] https://universaldependencies.org/

The UD annotation scheme is based on previous works established in this direction: the Stanford dependencies (de Marneffe and Manning, 2008), the Google universal part-of-speech tags (Petrov et al., 2011), and the Interset interlingua for morphosyntactic tag-sets (Zeman, 2008). Thus, the Universal Dependencies is a product of merging these previous enterprises into a unique and homogeneous framework.

The main objective of UD is to provide a linguistic representation that is useful "for morphosyntactic research, semantic interpretation, and for practical natural language processing across different human languages" (de Marneffe et al., 2021). Thus, UD focuses on simple surface representations that enable parallelism between similar phenomena in diverse languages.

The UD framework organizes its linguistic description around two fundamental linguistic elements: nominal units (i.e.: entities) and clauses (i.e.: events). Moreover, both of these elements can be further refined by the presence of units called modifiers (i.e.: attributes). As defined by de Marneffe et al. (2021): "Clauses can contain nominals, modifiers, and other clauses, nominals can also contain all three phrasal units, and modifiers can contain modifiers". These concepts are expressed in UD via a dependency grammar perspective (such as the one described by Mel'čuk, 2009). Thus, the words of a sentence are represented as a tree structure with the main predicate being the root.

The grammatical relations between heads and dependents are equivalent to Synt-D as described in Mel'čuk dependency grammar, 2009 (e.g.: the head of a nominal is canonically a noun, while the head of a clause – i.e.: predicate – is generally, but not always, a verb). Occasionally, "linguistic head functions are split between a structural centre and a semantic one (usually an auxiliary or function word and a lexical or content word respectively)" (de Marneffe et al., 2021). In these cases, the lexical word is annotated as being the head according to UD framework. Thus, "a UD tree represents a sentence's observed surface predicate-argument structure rather than necessarily accurately capturing phrase-internal syntactic constituency" (de Marneffe et al., 2021). In UD framework, the clear distinction between nominals and clauses is a fundamental concept as, in its guidelines, different types of dependency relations are defined for each one of these structures.

This annotation scheme has been developed following the lexicalist hypothesis in syntax, which means that grammatical relations should be considered as being between whole words (or lexemes). Thus, words are taken as the basic elements (with specific morphological properties) which are connected by dependency relations, in agreement with Mel'čuk dependency grammar (2009) and with the lexical integrity principle (Chomsky, 1970; Bresnan and Mchombo, 1995; Aronoff, 2007). This principle can be described as: "words are product of different structural elements and by different principles of composition other than syntactic constructions" (de Marneffe et al., 2008). This approach is more appropriate for practical computational models than the approach which considers that both words and phrases are built up using the same compositional syntactic mechanisms (and in which "the notion of a word has minimal privileged existence") (de Marneffe et al., 2014). Nevertheless, the abovementioned notion of word does not necessarily coincide with the orthographical or phonological units in all cases. For example, clitics (e.g.: the English genitive "'s") need to be separated from the hosts to be treated independently.

The UD framework provides an extended universal part-of-speech tag-set with a clear definition of categories. Moreover, the UD morphological features correspond to a common set of characteristics across human languages. In addition, the UD dependency representation (based on Stanford Dependencies) follows the ideas of grammatical relations-focused description: it is centrally organized around notions of subject, object, clausal complement, noun determiner, noun modifier, etc. Some modifications have been implemented throughout time to better account the grammatical structures of typologically different languages, thus, avoiding being centred in English and other European languages.

In its first release (version 1.0, from January 2015), 10 treebanks for 10 different languages were created, while in its version 2.10, from May 2022, the number of available treebanks is 228, corresponding to 130 languages.

As mentioned by Kondratyuk and Straka (2019) in their article presenting UDify tool, the Universal Dependencies corpora are adapted for analysing syntactic knowledge transfer across languages as they are based on a consistent and homogeneous annotation framework. It is also mentioned by De Lhoneux et al. (2018) as a useful resource for multilingual studies, specially concerning cross-lingual methods such as parameter sharing to improve results of dependency parsing for low-resourced languages.

The Universal Dependencies framework uses the CoNLL-U format as the standard for text annotation. It is an enhanced version of the CoNLL-X format, which was created for the 10th "Computational Natural Language Learning" (CoNLL) shared task (Buchholz and Marsi, 2006). Text and annotations are encoded in UTF-8 (Unicode Transformation Format 8), a standard variable-width character encoding worldwide used for electronic communication.

Each CoNLL-U file has three different types of lines:

1. Word-lines containing the annotation of a word-form/token in 10 fields separated by a single tab character.
2. Blank lines corresponding to sentence boundaries.
3. Comment lines (marked with a starting hash "#") containing information such as sentence identification and plain text, as well as possible translation in other languages.

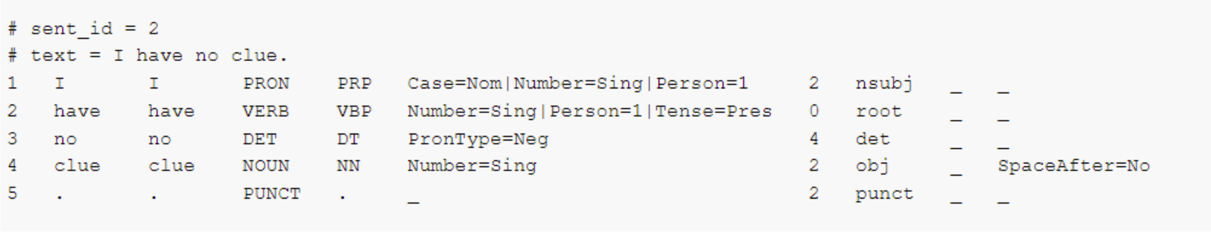
Each word line corresponds to one word-form, therefore, sentences are composed of one or more word-lines. Every word line is composed by the following fields (10 in total):

1. ID: Word index, integer starting at 1 for each new sentence and following its word order. It can be a range (e.g. 2-3) in cases of multiword tokens (e.g. "au" contraction between the preposition "à" and the article "le" in French), followed by lines with annotations for each component of the multiword element. In some corpora, there may be empty nodes (with decimal ID), which are used in the enhanced UD representation. These cases are ignored in this study.
2. FORM: Word-form or punctuation symbol (i.e.: the token, as it appears in the sentence).
3. LEMMA: Lemma or stem of the word-form.
4. UPOS: Universal part-of-speech tag (from a limited and determined tag-set).
5. XPOS: Language-specific part-of-speech tag.
6. FEATS: List of morphological features (from the Universal Dependencies inventory or from a defined language-specific extension).
7. HEAD: Head of the word-form, either a value of ID or 0.
8. DEPREL: Universal dependency relation to the HEAD (if HEAD is 0, DEPREL is "root"). It can also be followed by a language-specific subtype.

9. DEPS: Enhanced dependency graph in the form of a list of HEAD/DEPREL pairs. This feature is optional in Universal Dependencies treebanks, it will not be considered in this study.
10. MISC: Any other annotation that may be relevant, also not pertinent to this thesis.

In CoNLL-U files, all fields presented above must be fulfilled, and only FORM, LEMMA, and MISC columns accept space characters. When a specific item is unspecified, the underscore character "_" is used, however, UPOS, HEAD, and DEPREL are not allowed to be left undetermined (only in the case of multiword tokens). Figure 4.1 is an example of a sentence in English annotated following the CoNLL-U format.

Figure 4.1. Example of the sentence "I have no clue." annotated following Universal Dependencies framework in CoNLL-U format (figure extracted from the Universal Dependencies website[16]).



```
# sent_id = 2
# text = I have no clue.
1   I       I       PRON    PRP   Case=Nom|Number=Sing|Person=1     2   nsubj   _   _
2   have    have    VERB    VBP   Number=Sing|Person=1|Tense=Pres   0   root    _   _
3   no      no      DET     DT    PronType=Neg                      4   det     _   _
4   clue    clue    NOUN    NN    Number=Sing                       2   obj     _   SpaceAfter=No
5   .       .       PUNCT   .     _                                 2   punct   _   _
```

In the following subsections, further details on how the Universal Dependencies framework defines its annotation scheme are presented regarding: a) tokenisation and word segmentation, b) part-of-speech labels, c) morphological features, and d) syntactic relations.

a) Tokenisation and word segmentation:

As previously mentioned, the UD framework is based on a "lexicalist view of syntax which considers that dependency relations hold between words" (Nivre et al., 2020). Therefore, tokens are not segmented into morphemes, and morphological features are encoded and analysed as word properties. The basic units of annotation are syntactic words (not phonological or orthographic), thus, multiword tokens (a single orthograph token corresponding to multiple syntactic words) are systematically scrutinized, and each one of their components is annotated.

Word segmentation strongly depends on the properties of the language and the specific writing system. In this way, the presence of white space and punctuation facilitates the task, however, it can still be problematic in the case of languages in which the mapping between white-space delimited tokens and syntactic words is ambiguous (e.g.: Arabic and Hebrew).

---

[16] https://universaldependencies.org/

For each annotated corpus, the Universal Dependencies framework demands a precise explanation of choices concerning tokenisation and word segmentation (with references to standard tokenisation schemes, if possible).

b) Part-of-Speech:

The UPOS field (column 4) corresponds to the annotation in terms of part-of-speech (or word class), using the universal part-of-speech tag-set which has been determined by the contributors of the framework to be applicable to any human language. Tags are divided into three groups: open class words (e.g. verbs, nouns, and adjectives), closed class words (e.g. determiners, pronouns, and subordinating conjunction), and others (e.g. punctuation and symbol). The entire list of UPOS tags (17 in total) is presented in Annex 1. The UD framework does not assume that all possible UPOS must appear in all languages, but that every word in every language can be represented by one of these labels. Several different criteria (e.g.: morphological, syntactic, etc.) are necessary to describe word classes cross-linguistically, thus, the definition of word categories is not universal.

The XPOS field also corresponds to the part-of-speech annotation, however, most generally it is filled with language-specific part-of-speech tags, usually from more fine-grained tag-sets when compared to the UD list of tags. The UD framework requests that if the XPOS field is used, a mapping from XPOS to UPOS must be defined and described in the corpus documentation. As the XPOS tag-set may vary from language to language (and between corpora from the same language), this field is less relevant than UPOS regarding multilingual comparative studies.

c) Morphological annotation:

Some classes of words in several languages present paradigms of forms that express certain features (e.g.: number, tense, etc.). The list of morphological features of each token is presented in the 6th column of CoNLL-U files, named “FEATS”. Each one of them is represented as an attribute-value pair coming from the universal feature tag-set provided by the framework (24 features in total with a specific set of correspondent values), however, if some features are not present in this inventory, new ones can be created as long as they are completely described in the corpus documentation.

The full universal inventory of features, and their possible values, is presented in the Annex section (2 to 4). The universal features are classified into two different groups: lexical and inflectional ones, the latter subdivided into nominal and verbal. Yet, this classification is approximate as borders between these classes and sub-classes cannot be universally defined. Regarding lexical features, the same value applies to the entire paradigm (all forms with the same common lemma), while, concerning inflectional ones, different forms in a word's paradigm may have different values of the corresponding feature.

In the UD framework, there are no constraints in terms of compatibility between features and UPOS categories, even though specific restrictions may be observed in different languages. Moreover, the same feature may be "marked more than once on the same word, thus, defining several layers of the feature (e.g.: possessive adjectives, determiners, and pronouns which mark both gender and number of the possessor and the possessed entities). In these cases, the different layers are indicated by specific identifiers in square brackets after the feature label" (de Marneffe et al., 2021).

d) Syntactic annotation:

Syntactic annotation concerns the fields HEAD, DEPREL, and DEPS in the word lines, encoding the dependency tree of the sentence. The HEAD value of a word corresponds to the ID of the word governing the dependency relation. As is the case for UPOS and FEATS, the UD framework proposes a universal dependency relation tag-set (presented in Annexes 5 to 7), composed by 63 labels (from which 26 are composed of a type and a subtype which are separated by a colon) to be used as values for DEPREL. This list can be completed with other language-specific subtypes if needed, and the new labels must be detailed in the corpus documentation.

The HEAD and DEPREL values express the basic syntactical relations which rigorously determine a tree. The DEPS values correspond to an enhanced dependency representation, providing additional syntactical information about dependencies (in cases of propagation of dependencies over coordinate structures, for example). The DEPS values generate a graph, not a tree. As mentioned previously, this specific feature is not considered in our study as it is mostly absent in the Universal Dependencies corpora.

The Universal Dependency relations are classified into types of functional categories in relation to the head: core arguments of clausal predicates, non-core dependents of clausal predicates, and nominal dependents. Each category is characterized in terms of the dependent structural category (nominals, clauses, modifier words, and function words). The Universal Dependency relations tag-set also contains a list of relations that are not dependency ones in the strict sense, and which can be classified as: coordination, multiword expressions, loose, special, and other.

Moreover, the UD framework distinguishes "the core arguments of a predicate, essentially subjects and objects, from all other dependents at the clause level, collectively referred to as oblique modifiers" (de Marneffe et al., 2021). It is assumed that all languages have a method to identify the subject and object relations and that the status of a core argument is decoupled from the semantic role of a participant.

Furthermore, "UD does not assume the traditional argument–adjunct distinction found in many linguistic theories, which we take to be sufficiently subtle and hard to apply consistently both within and across languages, thus, the best solution is to avoid it" (de Marneffe et al., 2021).

Usually, the criteria to identify core arguments are specific for each language, however, the following principles are present in most cases:

- "Verbs usually only agree with core arguments".
- "Core arguments often appear as bare nominals while obliques are marked by adpositions or other grammatical markers".
- "Core arguments often appear in certain cases, traditionally called nominative, accusative, and absolutive".
- "Core arguments in many languages occupy special positions in the clause, often adjacent to the verb".
- "Properties such as being the controller of a subordinate clause argument are often limited to core arguments".
- "Valency-changing operations such as passive, causative, and applicative are often restricted to the promotion or demotion of core arguments".

The Universal Dependencies framework focus on grammatical relations, and its representations are in the midway between surface constituency and argument structure in multistratal theories (e.g.: f-structures in LFG, Bresnan et al., 2015; and deep syntactic representations in multi-stratal versions of dependency grammar, Mel'čuk, 1988). More precisely, "UD captures the observed surface predicate-argument structure rather than any sort of abstracted or underlying deeper structure. However, although being a monostratal theory, UD also needs to incorporate aspects of surface realization, such as word order, function words, and morphological inflections, which typically belong to a separate surface-oriented representation in multistratal theories" (De Maneffre et al., 2016).

Thus, UD represents a classic surface constituency solely by delimiting clauses, nominals, and modifiers. The inner structure of every single phrase represents predicates and grammatical relations which are similar to LFG f-structure and Synt-R in Mel'čuk dependency grammar (2009).

An alternative to the Universal Dependencies framework was presented by Gerdes et al. (2019a) under the name of Surface-Syntactic Universal Dependencies (SUD)[17]. The aim was to provide "a new surface-syntactic annotation scheme based on purely syntactic criteria (Mel'čuk, 2009), providing dependency structures closer to traditional dependency syntax (favouring functional heads)" (Gerdes et al., 2019a). Universal Dependencies corpora can be converted to SUD data using grammars such as the one developed by Bonfante et al., 2018.

The conversion from SUD to UD is also possible with some loss of information as the first schema uses a more succinct set of dependency labels to characterize the syntactic relations. The authors claim that SUD annotations are less redundant and more economical than UD annotations.

For instance, SUD uses a simple "subj" relation because the nominal character of a subject should be indicated only once (as a POS), while UD framework proposes two labels: "nsubj" (nominal subject), "csubj" (clausal subject). Figure 4.2 presents the specificities of SUD in comparison to UD framework.

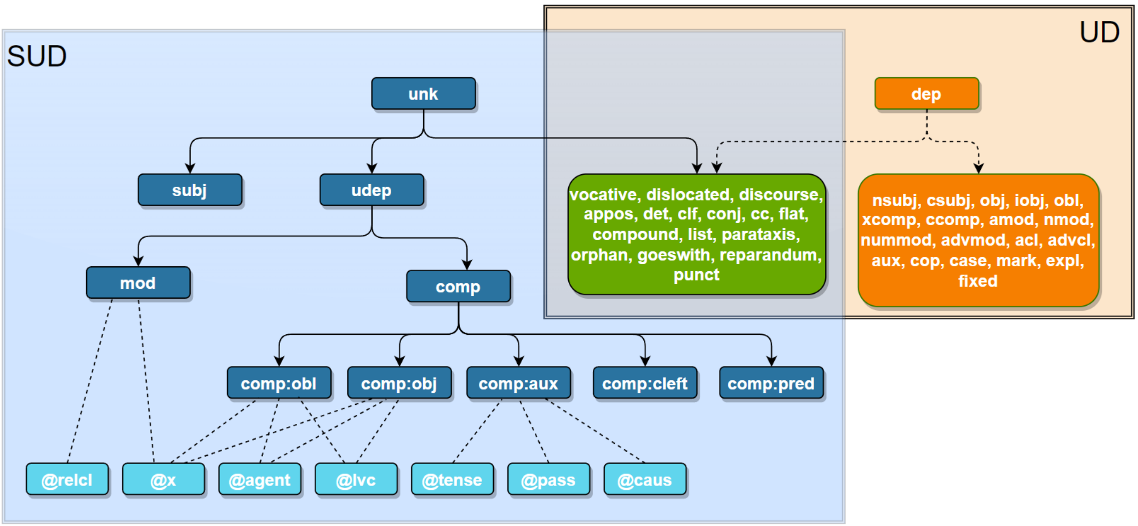
Furthermore, SUD creators affirm that the UD goal of "maximizing parallelism between languages might be of use for parser development of neighbouring languages, but reducing language differences makes the resulting treebank, by definition, less interesting for typological research on syntax. For example, UD does not account for the hierarchy between functional words and tends to flatten syntactic structures" (Gerdes et al., 2019a).

Although SUD approach seems interesting concerning the typological approach of this study, the UD framework is more pertinent to this study as its aim is to connect both dependency parsing usage of annotated corpora with quantitative typology and, until the present time[18], the UD framework is the main standard regarding multilingual dependency parsing projects.

---

[17] https://surfacesyntacticud.github.io/

Figure 4.2. Comparison between SUD and UD frameworks. The hierarchy of relations specific to SUD is presented in blue, the relations shared with UD are presented in green and in orange, the UD relations not used in SUD.


SUD
UD
unk
dep
subj
udep
vocative, dislocated, discourse, appos, det, clf, conj, cc, flat, compound, list, parataxis, orphan, goeswith, reparandum, punct
nsubj, csubj, obj, iobj, obl, xcomp, ccomp, amod, nmod, nummod, advmod, acl, advcl, aux, cop, case, mark, expl, fixed
mod
comp
comp:obl
comp:obj
comp:aux
comp:cleft
comp:pred
@relcl
@x
@agent
@lvc
@tense
@pass
@caus

It is also important to mention that other syntactic theories have been used to provide annotated corpora for corpus-based linguistic studies. It is the case of Role-and-Reference and Lexical Function Grammars (RRG and LFG respectively):

1) RRG: Bladier et al. (2019) and Evang et. al (2021) proposed automatic methods to transform Universal Dependencies data into corpora (ud2rrg) composed of RRG trees in accordance with Van Valin's standard formulation of this grammar.
2) LFG: Rosén et al. (2020) presented the INESS database composed of multilingual corpora following the LFG concepts as described by Bresnan and Kaplan in the 1970s. They also described some specific linguistic phenomena for which the search possibilities are increased when LFG representation is used compared to the shallower representation of UD corpora. Moreover, Przepiórkowski and Patejuk (2020), proposed a method to convert an LFG Polish corpus to one following Universal Dependencies guidelines, showing that only few information is lost during this conversion.

---

[18] A search on https://scholar.google.com/ (all dates and all languages) on 19/09/2022, using "Universal Dependencies" as keywords, displays more than 5,000 results while with the keywords "Surface-Syntactic Universal Dependencies" returns only 68 entries.

Even though RRG and LFG provide different insights which could be used for the typological study presented in this thesis, dependency syntax is still the most mainstream theory which is adopted in NLP projects, thus justifying our choice of focusing on Universal Dependencies corpora.

4.1.2 Parallel Universal Dependencies (PUD) corpora

The Parallel Universal Dependencies collection is an ensemble of treebanks (parallel annotated corpora following Universal Dependencies guidelines) that was developed for the CoNLL 2017 shared task on Multilingual Parsing from Raw Text to Universal Dependencies[19] for twenty languages[20] (Zeman et al., 2017): Arabic, Chinese, Czech, English, Finnish, French, German, Hindi, Icelandic, Indonesian, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish, Thai, and Turkish.

Regarding the genealogical classification of languages, we consider the one proposed by the World Atlas of Language Structures (WALS) which is, as presented previously, a large and popular online[21] database of structural properties (phonological, grammatical, lexical) compiled by a collaborative team from descriptive materials (Dryer et al., 2013).

The International Organization for Standardization (ISO) has issued in 2007 the latest international standard for language codes, named ISO 639-3, whose author is SIL[22]. The aim of this third part of ISO 639 was to establish an identifier for every distinct human language (spoken, written, or signed), living or extinct. Throughout this study, languages will be referred to in tables and figures by their ISO 639-3 identifier. Table 4.1 presents the list of languages inside the PUD collection of corpora, their respective ISO 639-3 code, their genealogical classification according to WALS database (family and genus), and the geographical area (Dryer, 1993).

Thus, according to the WALS database (Dryer et al., 2013), the twenty PUD languages correspond to 9 distinct linguistic families: twelve Indo-European languages (distributed in 4 different genera: Slavic, Germanic, Romance, and Indic), and eight languages from eight other families (Afro-Asiatic, Altaic, Austronesian, Japanese, Korean, Sino-Tibetan, Tai-Kadai, and Uralic). Although showing some variety in terms of linguistic families, when considering geographical distribution, the PUD collection contains only 3 of the 6 areas proposed by Dryer (1993): Africa, Eurasia; and Southeast Asia and Oceania.

---

[19] http://universaldependencies.org/conll17/

[20] The original language-set was composed of 18 languages. Polish and Icelandic were added afterwards.

[21] https://wals.info/

[22] https://www.sil.org/

Language	**ISO 639-3**	**Family**	**Genus**	**Geographical area (Dryer)**	**Geographical area (WALS)**
Arabic	arb	Afro-Asiatic	Semitic	Africa	Eurasia
Chinese	cmn	Sino-Tibetan	Chinese	Southeast Asia and Oceania	Eurasia
Czech	ces	Indo-European	Slavic	Eurasia	Eurasia
English	eng	Indo-European	Germanic	Eurasia	Eurasia
Finnish	fin	Uralic	Finnic	Eurasia	Eurasia
French	fra	Indo-European	Romance	Eurasia	Eurasia
German	deu	Indo-European	Germanic	Eurasia	Eurasia
Hindi	hin	Indo-European	Indic	Eurasia	Eurasia
Icelandic	isl	Indo-European	Germanic	Eurasia	Eurasia
Indonesian	ind	Austronesian	Malayo-Sumbawan	Southeast Asia and Oceania	Papunesia
Italian	ita	Indo-European	Romance	Eurasia	Eurasia
Japanese	jpn	Japanese	Japanese	Eurasia	Eurasia
Korean	kor	Korean	Korean	Eurasia	Eurasia
Polish	pol	Indo-European	Slavic	Eurasia	Eurasia
Portuguese	por	Indo-European	Romance	Eurasia	Eurasia
Russian	rus	Indo-European	Slavic	Eurasia	Eurasia
Spanish	spa	Indo-European	Romance	Eurasia	Eurasia
Swedish	swe	Indo-European	Germanic	Eurasia	Eurasia
Thai	tha	Tai-Kadai	Kam-Tai	Southeast Asia and Oceania	Eurasia
Turkish	tur	Altaic[23]	Turkic	Eurasia	Eurasia

Table 4.1. List of languages inside PUD collection, their respective ISO 639-3 three-character code, their phylogenetic (WALS), and geographical information (Dryer, 1992 and WALS).

It has been pointed out (e.g.: Moravcsik, 2012 and Dryer, 1993) the importance of having a large variety of languages in terms of geographical areas and linguistic families when conducting typological studies. Thus, it justifies the choice of focusing, in this first step, on parallel corpora (worldwide languages) to decrease some bias concerning the size and the semantic content. Once the best strategy is defined, it can be applied to wider scenarios (i.e.: to all EU languages). In terms of word order typology, it is possible to characterize almost all of the PUD languages with the analysis provided by Hawkins (1983). The exception is Polish, for which no information concerning its word order characteristics can be found in this specific reference.

---

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

For each language of Hawkins' extended sample, the author presented in his 1983's book: a) the typological status concerning the position of the subject, verb, and object; b) if the language has prepositions or postpositions; c) the word ordering of different components of noun phrases, d) the number of the typological class[24]. Table 4.2 presents the PUD corpora characterization according to Hawkins' typology.

Language	**Hawkin's Word Order Summary**			**Type**
arb	VSO	Pr	NumN/nnum, DN, NPoss, NA, NG, Nrel	1
cmn	SOV/SVO	Pr/Po	DN, AN, GN, RelN	-
ces	SVO	Pr	NumN, DN, AN, NG, NRel	10
eng	SVO/v-1	Pr	NumN, DN, PossN, AN, GN/NG, NRel	-
fin	SVO	Po	NumN, DN, AN, GN, reln/Nrel, AdvAdj, SMAdj/AdjMS	15
fra	SVO	Pr	NumN, DN, PossN, an/NA, NG, NRel	9
deu	SOV/v-1,V-2	po/Pr	NumN, DN, PossN, AN, GN/NG, reln/NRel	-
hin	SOV	Po	NumN, DN, AN, GN, NRel/RelNRel, AdvAdj, SMAdj	23
isl	SVO	Pr	DN, AN, NG, NRel	10
ind	SVO	Pr	NumN, ND, NPoss, NA, NG, NRel	9
ita	SVO	Pr	NumN, DN, an/NA, NG, NRel	9
jpn	SOV	Po	NumN/NNum, DN, AN, GN, RelN, AdvAdj, SMAdj	23
kor	SOV	Po	NumN, DN, PossN, AN, GN, RelN	23
pol	-	-	-	-
por	SVO	Pr	NumN/NNum, DN, PossN/NPoss, an/NA, NG, NRel	9
rus	SVO	Pr	NumN, DN, AN, NG, NRel	10
spa	SVO	Pr	NumN/NNum, DN, PossN/NPoss, an/NA, NG, NRel	9
swe	SVO	Pr	NumN, DN, PossN, AN, GN, NRel	11
tha	SVO	Pr	NumN, ND, NPoss, NA, NG, Nrel, AdjAdv, AdjMS	9
tur	SOV	Po	NumN, DN, AN, GN, RelN, AdvAdj, SMAdj	23

Table 4.2. Typological characteristics and classification of PUD languages according to Hawkins (1983). When components are written in lower cases, it means that the phenomenon is less frequent than the other possible word order structure involving the same elements.

It is possible to notice that there is a predominance of VO languages (12 in total) in the PUD collection, and all these languages have prepositions, while the few OV languages have postpositions. Chinese, German and English have a more complex subject, object, and verb orderings, and the first two possess both types of adposition, thus, they are not characterized by Hawkins (1983) with an associated typological type.

---

[24] Hawkins defined different typological classes combining the ordering of subject, verb, and object with the adposition strategy of the language, and with the relative position of qualifying adjectives and genitives in relation to the noun. No type has been provided for Chinese, English, German, or Polish.

Regarding the other 17 PUD languages, they are divided into 6 different classes:

1) Type 1: Arabic;
2) Type 9: French, Indonesian, Italian, Portuguese, and Spanish;
3) Type 10: Czech, Icelandic, and Russian;
4) Type 11: Swedish;
5) Type 15: Finnish;
6) Type 23: Hindi, Japanese, Korean, and Turkish.

The above classification considers only word-ordering phenomena. Thus, it allows the presence of languages, that are not genealogically related, in the same class-type (such as the ones from type 23). Hawkins (1983) showed how this type of analysis can provide interesting results in terms of prediction (and quantification) of possible languages. However, it is quite limited for NLP applications as it does not allow fine-grained comparison among languages classified as the same type. Thus, the usage of typological databases with a higher number of word order features is more appropriate for specific applications such as the improvement of dependency parser results via typological strategies.

In terms of composition, each PUD corpus contains 1,000 sentences strictly in the same order (the sentence alignment is 1-1, although in some cases a sentence-level segment is formed by two real sentences). The sources of the PUD sentences are texts from the news domain (sentence id starts with “n”) and from Wikipedia (sentence id starts with “w”). The sentences were selected randomly from a great variety of documents, therefore, most of them come from different texts and when sentences belong to the same document, they are not necessarily adjacent.

The vast majority of PUD sentences are originally in English, but some of them come from German, French, Italian, or Spanish texts. In table 4.3, the distribution of the sentences considering the source language is presented. Most translations[25] were provided by DFKI[26] and executed by professional translators (except for German) via English. Morphological and Syntactical annotations were performed by Google following its universal annotation guidelines and, then, translated to Universal Dependencies labels by members of the Universal Dependencies community. Morphological features and lemmas were added automatically using Stanford CoreNLP (Manning et al, 2014).

---

[25] Czech, Swedish, and Finnish corpora were prepared outside this workflow as they were not included in the original plans. Similarly, Poslish and Icelandic were added later and translations were performed by other teams.

[26] https://www.dfki.de/web

Language	**PUD Reference**	**Number of sentences (news)**	**Number of sentences (Wikipedia)**	**Total**
eng	01	375	375	750
deu	02	50	50	100
fra	03	25	25	50
ita	04	25	25	50
spa	05	25	25	50

Table 4.3. Distribution of PUD sentences concerning the original language and type of text.

Although the number of sentences is the same for each PUD corpus, the numbers of tokens and words[27] per dataset vary due to the morphosyntactic characteristics of each language. Table 4.4 presents the detailed information regarding the amount of these elements for each PUD corpus.

Language	**Number of tokens**	**Number of words**
arb	20,751	20,751
cmn	21,415	21,415
ces	18,565	18,610
eng	21,176	21,176
fin	15,807	15,813
fra	24,137	24,734
deu	21,001	21,329
hin	23,829	23,829
isl	18,831	18,833
ind	19,030	19,440
ita	22,182	23,731
jpn	28,784	28,784
kor	16,584	16,584
pol	18,338	18,389
por	21,917	23,407
rus	19,355	19,355
spa	22,822	23,287
swe	19,076	19,076
tha	22,322	22,322
tur	16,536	16,882

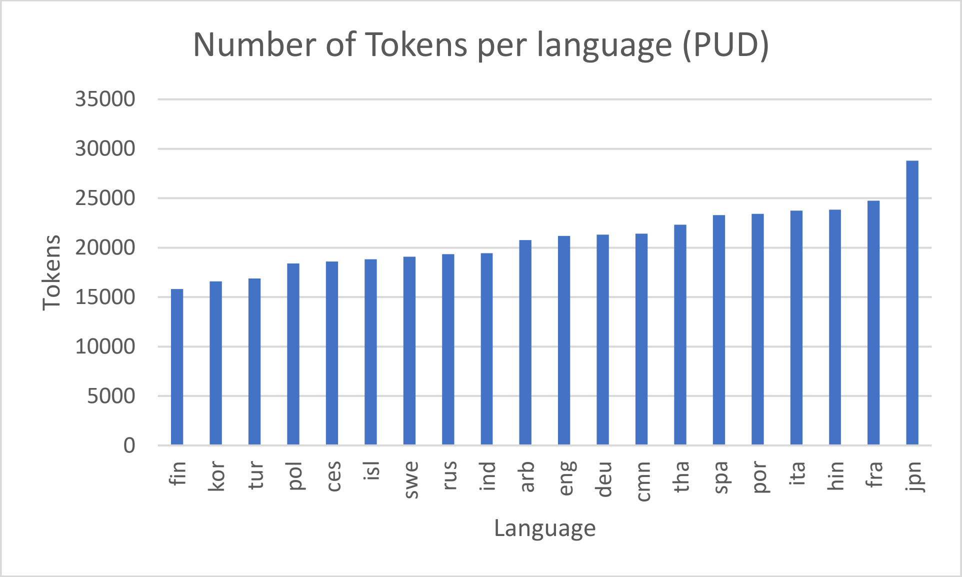
Table 4.4. Number of tokens and words for each PUD language. Languages in yellow are the ones for which the number of tokens is identical to the number of words.

---

[27] Tokens mean the surface tokens (e.g.: vámonos in Spanish) while words mean syntactic words (e.g.: the Spanish token vámonos is split into two words “vamos” and “nos”).

As expected, there is a large discrepancy in the number of tokens when comparing PUD languages (from 15,813 for Finnish to 28,784 for Japanese). The graph in Figure 4.3 shows in crescent order the differences between PUD languages regarding their sizes.

Figure 4.3. Size of each PUD corpus regarding the number of tokens.



As it is noticeable, Finnish and Korean present the least amounts of tokens due to their agglutinative aspect. Morphological-rich languages such as Polish and Czech also present fewer tokens than Romance languages and Japanese (analytical languages). Regarding the annotation scheme, as presented previously, the Universal Dependencies framework established a universal part-of-speech tag-set (UPOS) with seventeen possible labels. The table 4.5 provides information concerning the number of these UD tags in each PUD language.

Interjections are absent in 9 out of the 20 PUD corpora. The English corpus contains only one sentence with this part-of-speech: “Luckily, someone in Sony Australia was like, 'Hey, by the way, did you guys notice this?'" says Pall.”, the interjection being the token “Hey”. It has been translated by “Tiens” in French, which is erroneously tagged as verb due to its form being the same as the 3rd person singular of the present tense (indicative) of the verb “tenir”.

Moreover, the “X” tag is defined by the UD framework as a label to be “used for words that for some reason cannot be assigned a real part-of-speech category”. In the case of the English corpus, it has been assigned to foreign terms (a total of 17 tokens). The other tags which are absent in some corpora represent some specificities of the respective languages (e.g.: the absence of particles in Portuguese and Spanish, and determiners in Finnish).

Language	Number of UPOS labels	Labels not present in the corpus
arb	16	INTJ
cmn	15	INTJ, SYM
ces	15	INTJ, X
eng	17	-
fin	15	DET, PART
fra	16	INTJ
deu	16	INTJ
hin	16	INTJ
isl	17	-
ind	17	-
ita	16	INTJ
jpn	16	X
kor	13	ADP, INTJ, SCONJ, SYM
pol	16	INTJ
por	16	PART
rus	17	-
spa	15	INTJ, PART
swe	16	X
tha	15	INTJ, X
tur	16	PART

Table 4.5. Number of labels from the universal part-of-speech tag-set for each PUD language and labels from this inventory that are not present in each corpus.

In terms of UPOS, only 10 out of the 17 labels used for PUD annotation are present in all languages of this ensemble of corpora (list presented in table 4.6).

UPOS tag	Part-of-speech
ADJ	Adjective
ADV	Adverb
AUX	Auxiliary
CCONJ	Coordinating conjunction
NOUN	Noun
NUM	Numeral
PRON	Pronoun
PROPN	Proper noun
PUNCT	Punctuation
VERB	Verb

Table 4.6. Labels from the universal part-of-speech tag-set present in every PUD language.

As explained earlier, the XPOS labels serve as language-specific part-of-speech tags. The analysis of PUD XPOS tag-set shows that 13 corpora follow Penn Treebank guidelines for this specific annotation (Santorini, 1990). Finnish and Indonesian corpora do not have any XPOS annotations, while Czech, Icelandic, Japanese, Polish, and Swedish corpora have language-specific sets of labels. Due to this discrepancy and to the fact that UPOS labels are available for all languages providing enough part-of-speech information, XPOS tags are not considered in this thesis.

In terms of FEATS labels, the Universal Dependencies framework suggests a list of 24 features for which different values can be attributed with the possibility of adding language-specific ones to the list. In PUD corpora, there is a total of 43 different features, 23 of which come from the universal features tag-set (the only one not present in PUD corpora is “NounClass”).

As PUD is composed of languages with different morphological complexity, it is possible to observe a great discrepancy in terms of the features used to describe each corpus, some of them presenting many different labels (maximum of 30 features for the Polish language, and minimum of 1 for the Japanese language). This observed difference in terms of morphological description may also be due to diverse criteria utilized by annotators of each language. No feature is present in all languages. In table 4.7, the PUD corpora are described regarding the number of features and the language-specific labels.

Concerning dependency relations, the Universal Dependencies framework established a set of 63 DEPREL labels[28] (see Appendix 5 to 7), yet, this tag-set is not closed and specific relations can be used if needed. The PUD collection contains, in total, 110 DEPREL labels, 50 tags appear in only one corpus among the twenty PUD languages, and, from these specific labels, only 3 are present in the original universal dependencies tag-set: “expl:impers”, “expl:pass”, and “list”. All the others DEPREL in PUD collection are specific tags, mostly created by combining a DEPREL label with a subtype representing specific syntactic relations of the respective language.

The table 4.8 presents the list of PUD languages with the respective total number of DEPREL labels and the language-specific relations present in each corpus.

From the 110 DEPREL labels present in the PUD collection, only 15 are used in all PUD languages. This list is presented in the table 4.9.

---

[28] This number considers both types and subtypes. The number of main types is 37.

Language	Number of FEAT tags	Language-specific features
arb	12	-
cmn	8	-
ces	26	NameType, NumValue
eng	15	-
fin	21	Connegative, PartForm, Derivation, InfForm
fra	10	-
deu	13	-
hin	14	-
isl	14	-
ind	14	Clusivity*, Typo*
ita	14	-
jpn	1	-
Kor	12	Form
pol	30	VerbType, ConjType, PartType, PunctType, Pun
por	13	-
rus	15	-
spa	19	-
swe	15	-
tha	6	-
tur	16	Evident*, Register

Table 4.7. Number of FEATS labels present in each PUD corpus and specific tags used only for each respective corpus. The "*" symbol represents labels present in the original universal features tag-set.

When analysing PUD languages in terms of types and subtypes of DEPREL (as displayed in table 4.10), it is possible to notice that the number of types varies from 25 (Japanese) to 36 (English). Regarding subtypes, the discrepancy is even higher, Japanese language does not require the usage of any subtype while there are 31 for Polish. Most languages have between 8 to 14 DEPREL subtypes.

As presented in the previous section, DEPREL tags are checked when calculating the efficiency of dependency parsers using the LAS metric. For the calculation of MLAS, UPOS and FEATS are also considered, thus, it is important to have this precise view on how PUD languages differ concerning these annotations as it may play a role when applying the strategy of combining languages for improving dependency parsing results.

Language	Number of DEPREL labels	Specific DEPREL label
arb	42	-
cmn	44	case:loc, discourse:sp, mark:adv, mark:prt, mark:relcl, obl:patient
ces	43	expl:pass*
eng	48	nmod:npmod, obl:npmod
fin	44	compound:nn, cop:own, csubj:cop, nsubj:cop, nmod:gsubj, nmod:gobj, xcomp:ds
fra	45	aux:caus, aux:tense, expl:comp, expl:subj, nsubj:caus, obj:agent, obl:mod
deu	45	-
hin	38	compound:conjv, list*
isl	36	-
ind	47	case:adv, compound:a, nmod:lmod
ita	40	expl:impers*
jpn	25	-
kor	34	dep:prt
pol	59	advcl:relcl, advmod:arg, advmod:neg amod:flat, aux:clitic, aux:cnd, ccomp:cleft, ccomp:obj, nmod:arg, nmod:flat, nmod:pred, obl:cmpr, parataxis:insert, parataxis:obj, xcomp:pred, xcomp:subj
por	42	-
rus	39	nummod:entity
spa	41	-
swe	42	acl:cleft
tha	41	obl:poss
tur	43	aux:q

Table 4.8. Number of DEPREL labels present in each PUD corpus and specific tags used only in the respective corpus. The “*” symbol represents labels which appear in the global UD documentation where the main types are described.

The aforementioned objectives of this thesis concern typological strategies for improving dependency parsing results for low-resourced languages. Yet, not all PUD languages can be classified as such, as can be observed in Table 4.11 where the size (in terms of the number of sentences) of the largest UD corpus (v.2.10) for each PUD language is presented.

DEPREL tag	Dependency Relation
advcl	adverbial clause modifier
advmod	adverbial modifier
amod	adjectival modifier
appos	appositional modifier
cc	coordinating conjunction
ccomp	clausal complement
cop	copula
det	determiner
fixed	fixed multiword expression
nsubj	nominal subject
nummod	numeric modifier
obj	object
obl	oblique nominal
punct	punctuation
root	root

Table 4.9. Labels from the universal dependencies tag-set present in every PUD language.

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

Table 4.10. Distribution of the number of DEPREL labels (types and sub-types) for each language in the PUD data-set.

Among PUD languages, only Thai can be considered a real low-resourced language as the only UD corpus available is the PUD itself. Other languages such as Chinese, Swedish and Indonesian have at least one corpus with a size higher than 4,500 sentences. It is important to mention that the size of annotated data is just one possible way for defining low-resourced languages which correspond to the "Speech and Text Resources" criterion[29] defined by the META-NET Language Whitepaper series (Rehm et al. 2012). Although the languages chosen for this study are not low-resourced ones, the ensemble of experiments is conducted in a low-resourced scenario (1,000 sentences). Truly low-resourced languages[30] could have been selected, nevertheless, the choice of using parallel corpora for the reasons aforementioned was privileged. Once the best method for combining languages is defined, it can be, then, applied in real low-resource scenarios, which is the case of the application in section 6 regarding EU low-resourced languages.

Languages	Name of Treebank	Number of sentences
arb	PADT	7,644
cmn	GSD	4,997
ces	PDT	87,913
eng	EWT	16,621
fin	TDT	15,136
fra	GSD	16,341
deu	HDT	189,928
hin	HDTB	16,647
isl	IcePaHC	44,029
ind	GSD	5,598
ita	ISDT	14,167
jpn	GSD	8,071
kor	Kaist	27,363
pol	PDB	22,152
por	GSD	12,019
rus	SynTagRus	87,336
spa	AnCora	17,662
swe	Talbanken	6,026
tha	PUD	1,000
tur	Kenet	18,687

Table 4.11. Largest available UD corpus (v.2.10) for each PUD language and the respective size in terms of the number of sentences.

---

[29] Together with "Speech and Text Resources", the META-NET consortium also defined "Machine Translation", "Speech Processing" and "Text Analysis" as criteria for classifying languages from "excellent support" to "weak/no support".

[30] E.g.: Languages with less than 1,000 sentences in UD (v.2.10) such as Yoruba (Niger-Congo family) and Tupinamba (Tupian family).

4.1.3 URIEL and lang2vec

As previously described, URIEL is a structured collection of information on language typology which compiles linguistic data from many different sources, such as WALS (Dryer and Haspelmath, 2013), PHOIBLE (Moran et al., 2014), Ethnologue (Lewis et al., 2015), and Glottolog (Hammarström et al., 2015). It has been released along with lang2vec[31], a querying tool specifically conceived to easily extract information from this database (Littel et al., 2017).

Languages are represented as typological, phylogenetic, and geographical vectors in homogeneous and consistent formats which allow languages to be linguistically compared according to the user's need. Languages are identified by their ISO 639-3 code, each feature composing the language vector receives a value from 0 (generally indicating the absence of the phenomenon) to 1 (when the feature is normally observed in the language). As lang2vec provides standardized and normalized information coming from a great variety of sources, it allows better replication and language comparison.

The representation of languages as vectors has been proven to be more effective for NLP tasks when compared to language representation in one-single dimension, as shown by Tsvetkov et al. (2016) and Bharadwaj et al. (2016).

For each language, four types of vectors are available:

1. Typological vectors: these vectors concern either syntax or phonology, thus, being divided into three sub-categories.
    a. Syntactic vectors describing languages morphosyntactically whose information comes from WALS database (Dryer and Haspelmath, 2013), the Syntactic Structures of World Languages (SSWL) (Collins and Kayne, 2009), and Ethnologue (Lewis et al., 2015).
    b. Phonological vectors whose sources are WALS (Dryer and Haspelmath, 2013) and Ethnologue (Lewis et al., 2015).
    c. Inventorial vectors which provide the whole available phonological information following PHOIBLE (Moran et al., 2014), a normalized representation of phonological features applied to the ensemble of phonological databases available in URIEL.

---

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

2. Phylogenetic vectors: they express shared membership in language families, according to the Glottolog world language family tree (Hammarström et al., 2015). The different existing linguistic families and genera compose the list of features of the language phylogenetic vectors and a value of 1.0 is assigned if the language belongs to the correspondent family/genus, and 0.0 if not.
3. Geographical vectors: this type of vector expresses geographical location with a fixed number of dimensions representing the orthodromic distance ("great circle" distance) to a fixed point on the Earth's surface. The distances are described as a fraction of the Earth's antipodal distance (values between 0.0 and 1.0). Language distance points are obtained from Glottolog (Hammarström et al., 2015), WALS (Dryer and Haspelmath, 2013), and the Syntactic Structures of World Languages (Collins and Kayne, 2009).
4. Identity vectors: these vectors serve as an identifier for each language to be used as a control in experiments or combined with other vectors whose information may not identify the language uniquely.

Phylogeny, geography, and identity vectors do not have missing values; however, typological vectors may not be complete due to the possible lack of information of certain features in the listed sources. This can be problematic in typological studies focused on morphosyntactic structures which is the case of this thesis. All typology vectors have the same dimensionality within the same sub-category to allow straightforward comparison of languages. When values are missing, they are represented in the vectors as "--".

Thus, under-resourced languages, which usually have less descriptive studies, have emptier typological vectors in URIEL when compared to major languages. This tool proposes a functionality that allows the prediction of values regarding certain features for incomplete languages. However, in this thesis, only values that have reliable bibliographical references are considered.

The lang2vec tool allows users to access determined information coming from a specific database, but it also proposes average vectors ("avg") which are generated via the combination of the data from all databases, being, therefore, more complete than single-source vectors. Table 4.12 shows the coverage (in terms of languages and number of features) concerning syntactic features of each source (typological databases: WALS, SSWL, and Ethnologue) and the resulting average vector.

Vector type	Number of languages	Number of features
syntax_wals	1,808	98
syntax_sswl	230	33
syntax_ethnologue	1,336	30
syntax_avg	2,654	103

Table 4.12. Coverage of syntax vectors available in the URIEL database.

Since the coverage of the average vector is higher, it is the one considered for this study. Syntactic information provided by vectors comes from qualitative sources.

Each feature can receive the following values:

- 0.00 – 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.
- 0.67 – the phenomenon is relatively common.
- 1.00 – the phenomenon is normally encountered in the language.

These values are determined via the compilation of the information provided by the different sources and have not been calculated with quantitative methods using annotated corpora. The URIEL database and lang2vec tool have been used in many experiments concerning typological strategies for improving NLP tasks (Levshina 2022). Therefore, it is considered as the standard of comparison for the proposed quantitative methods in this study.

4.1.4 MarsaGram

MarsaGram is an open-source software (Blache et al., 2016) created by a specialized team from the Laboratoire Parole et Langue (Speech and Language Laboratory) and is available[32] as a downloadable tool in the ORTOLANG (Outils et Ressources pour un Traitement Optimisé de la LANGue) repository containing more than five hundred linguistic resources (Pierrel, 2014).

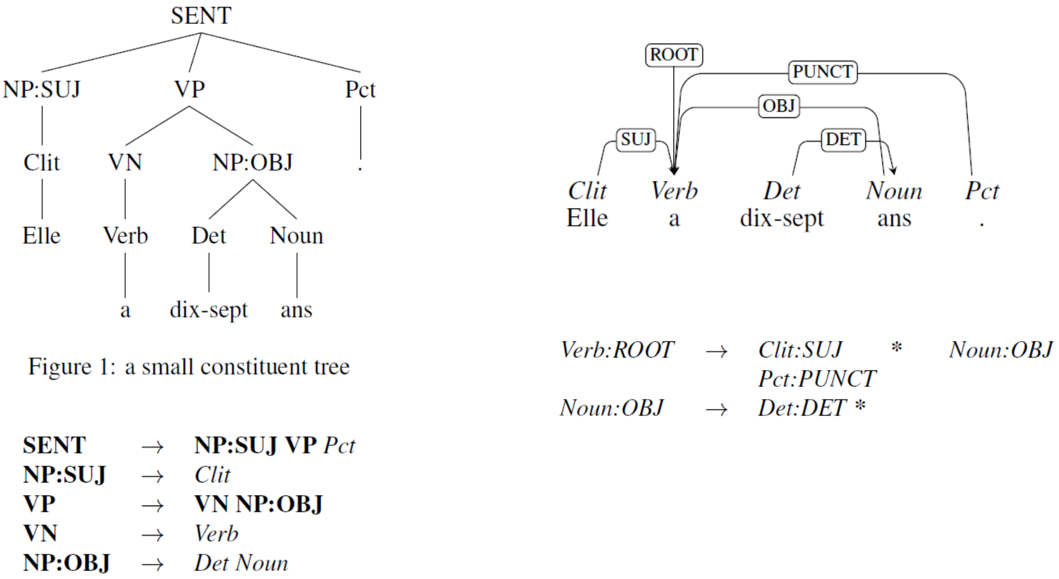
This tool was developed to allow languages to be compared syntactically using a quantitative method of extraction of context-free grammars (CFG) from treebanks. MarsaGram extracts parameters regarding syntax by navigating through annotated corpora. Quantitative analysis, such as the one proposed by MarsaGram, allows “the identification of regularities and the description of specific realization of syntactic constructions independently of the formalism (constituency or dependency)” (Blache et al., 2016). This resource is capable of identifying syntactic patterns which are implicit in treebanks, thus, describing “finer-grained information such as government phenomena, linear order, cooccurrence, etc” (Blache et al., 2016). This type of information is valuable for typological studies where detailed information such as relations between verbs and arguments and head and modifiers are necessary, as is the case of this thesis.

---

[32] https://www.ortolang.fr/market/tools/ortolang-000917/v1

MarsaGram’s method consists of the extraction of a context-free grammar (CFG) from a constituency or a dependency treebank. For constituency trees, each internal node is represented as a rule: “the left-hand side (LHS) being the node’s constituent tag and the right-hand side (RHS), the sequence of children’s nodes labels. In the case of dependency trees, internal nodes (LHS) are composed of the head category and the children nodes (RHS) are the list of the dependents, according to the projection order and with an extra node noted as “*” that corresponds to the head projection” (Blache et al., 2016). This set of extracted rules (quantified the by number of occurrences and frequency) compose the implicit grammar associated with the corpus. Figure 4.4 shows an example of the parsing tree for the sentence in French “Elle a dix-sept ans.” (“She is seventeen.”) and the inferred rules (Blache et al., 2016).

Figure 4.4. Example of parsing tree and CFG rules extracted from the sentence “Elle a dix-sept ans.”, on the left, the constituency tree, and on the right, the dependency one (Blache et al., 2016). The arrows in the dependency represention presented by these authors do not follow the usual convention (i.e.: arcs from the heads to the dependents).


SENT
NP:SUJ
VP
Pct
Clit
VN
NP:OBJ
.
Elle
Verb
Det
Noun
a
dix-sept
ans
Figure 1: a small constituent tree
SENT → NP:SUJ VP Pct
NP:SUJ → Clit
VP → VN NP:OBJ
VN → Verb
NP:OBJ → Det Noun
ROOT
PUNCT
OBJ
SUJ
DET
Clit
Elle
Verb
a
Det
dix-sept
Noun
ans
Pct
.
Verb:ROOT → Clit:SUJ * Noun:OBJ Pct:PUNCT
Noun:OBJ → Det:DET *

The MarsaGram tool extracts four different types of patterns from the CFG inferred rules:

1. Linearity: two components (in a subtree governed by a specified node) have a linear relationship when one occurs before the other at the surface level. It is noted as "precede" in MarsaGram analysis.
2. Requirement: two components have a requirement relation if the presence of one requires the presence of the other.
3. Exclusion: two components have this relationship when they do not occur together.
4. Unicity: this property is defined for one component when it never occurs multiple times in the RHS of a rule with the same LHS.

The identified patterns are, then, filtered regarding:

- Tag granularity: As seen in Figure 4.4, tags are composed by the part-of-speech and the syntactic function (e.g. NP:SUBJ). MarsaGram keeps only the first level of information (POS).
- None elements: these are components that express relations between constituents but which have no projection (e.g. ellipsis). This type of phenomenon is not considered by MarsaGram.
- Coordination: the authors consider that coordination is frequent but does not provide much information on relations between constituents. Thus, this filter removes rules extracted from pre-defined coordination patterns.
- Frequency: To avoid noise, a minimal number of occurrences is required for the rule to be considered.

Thus, by identifying the linear relations, MarsaGram recognizes patterns concerning the word order at the surface level of the sentence between two components that are part of the same subtree. On the other hand, requirement, exclusion, and unicity relations do not provide information about the word order itself but provide analytical elements concerning possible (and impossible) combinations of components inside a subtree ruled by a specific node.

Each pattern (p) is identified as a 4-tuple described below:

$$(4.1)\ p = < C, rel, A, B >$$

Where C is the LHS component, rel one of the four possible relations, and A and B (from the RHS) are the two components for which the property is defined (B = A for unicity relation).

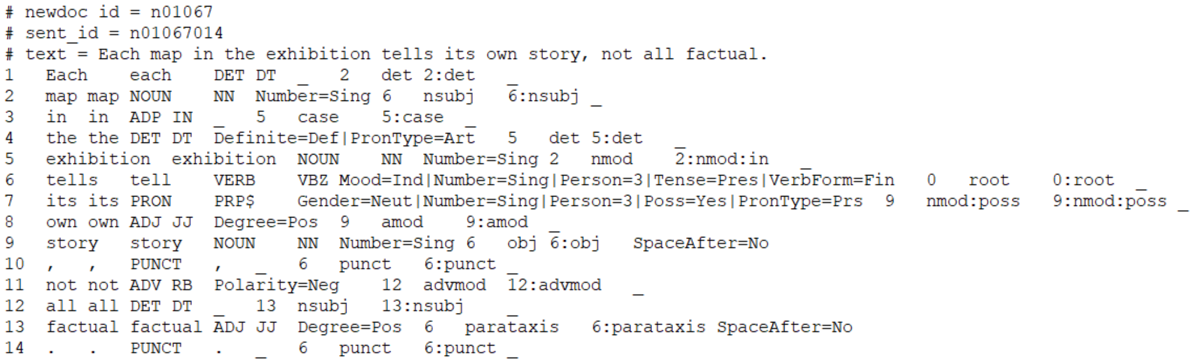
For each pattern, the system counts the number of times that the rule is respected or violated and, then, applies weights that balance these frequencies in relation to the ensemble of occurrences, thus, determining the final set of MarsaGram patterns.

This software has been used in different linguistic studies such as the one proposing to establish a link between linguistics and fuzzy phenomena (Urrutia et al., 2018) and the article proposing new approaches concerning Spanish syntax (Urrutia, 2017). Nevertheless, it has never been used for typological strategies for dependency parsing improvement.

In practice, by applying MarsaGram perl scripts to a dependency parsing treebank, the outcome is a set of patterns extracted from the identified CFG rules in the format of a tsv file. Below, four examples of patterns extracted from the English PUD corpus are presented (representing each possible MarsaGram relation):

1. $< NOUN, precede, DET - det, NOUN - nmod >$: An example a of sentence illustrating this property is presented in Figure 4.5. It means that the element DET-det (token 4) appears before the element NOUN-nmod (token 5) in the sentence and that they are both parts of a subtree governed by a NOUN (token 2).
2. $< VERB, require, NOUN - nsubj{:}pass, AUX - aux{:}pass >$: An example of a sentence illustrating this property is presented in Figure 4.6. This relation (require) means that in the whole corpus, component 1 (NOUN – nsubj:pass, token 2) is always in the presence of component 2 (AUX – aux:pass, token 3) inside the subtree governed by the head VERB (token 5).
3. $< VERB, exclude, NOUN - nsubj, PRON - nsubj >$: It means that inside the English PUD corpus, the element NOUN-nsubj never appears together with a PRON-nsubj in a subtree governed by a VERB. The relation “exclude” means that this pattern cannot be seen in the corpus, thus, no example a of sentence can be presented.
4. $< ADJ, unicity, NOUN - obl{:}npmod >$: An example a of sentence illustrating this property is presented in Figure 4.7. It means that the component (NOUN – obl:npmod, token 13) never occurs multiple times when the head of the subtree has the part-of-speech tag “ADJ” (token 15).

Figure 4.5. Sentence from which property 1 presented above has been extracted. Source: PUD English corpus. The head of the subtree is the token “map” (id=2), and the components 1 and 2 are respectively the tokens “the” (id=4) and “exhibition” (id=5).

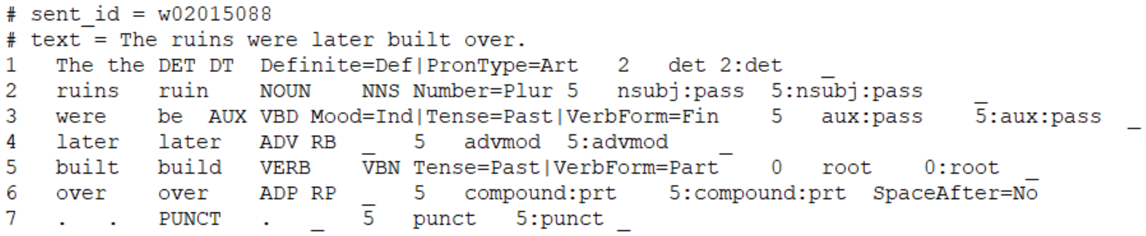


```
# newdoc id = n01067
# sent_id = n01067014
# text = Each map in the exhibition tells its own story, not all factual.
1   Each    each    DET DT  _   2   det 2:det   _
2   map map NOUN    NN  Number=Sing 6   nsubj   6:nsubj _
3   in  in  ADP IN  _   5   case    5:case  _
4   the the DET DT  Definite=Def|PronType=Art   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 _
```

The analysis provided by MarsaGram allows the identification and quantification of word order phenomena that occur in specific syntactic constructions combined with other specific properties. Blache et al. (2016) showed that the linear relations seem to be more adapted to be used in typological studies as the results excluding the other patterns were more coherent when compared to the genealogical classification. As no study concerning the usage of MarsaGram patterns has been conducted with the aim of improving dependency parsing results via corpora association, both scenarios will be considered in this thesis: a) all MarsaGram patterns and b) only linear relations.

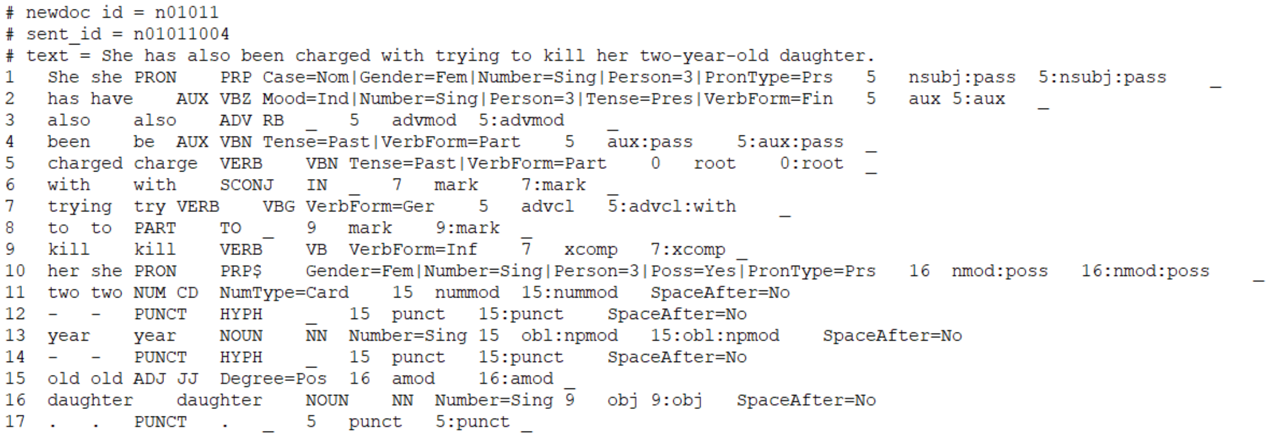
The ensemble of patterns obtained with MarsaGram does not correspond to any of the typological theories previously mentioned. Although it assumes the existence of heads and dependents, the extracted patterns regarding word ordering at the sentence level concern elements which are not necessarily a head and dependent pair. What the components of a linear pattern have in common is that they are both inside a subtree ruled by the same head (but elements are not necessarily directly governed by it). Thus, the linear patterns represent word order possibilities that can occur inside structures defined by a specific head (characterized by its part-of-speech), allowing, in this way, a fine-grained analysis of word order phenomena inside determined syntactical structures.

Figure 4.6. Sentence from which property 2 presented above has been extracted. Source: PUD English corpus. The head of the subtree is the token “built” (id=5), and the components 1 and 2 are respectively the tokens “ruins” (id=2) and “were” (id=3).



```
# 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 4.7. Sentence from which property 4 presented above has been extracted. Source: PUD English corpus. Columns 9 and 10 have been omitted for not being relevant to this study. The head of the subtree is the token “old” (id=15), and the component is the token “year” (id=13).



```
# newdoc id = n01011
# 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 _
```

4.2 Methods

The first hypothesis of this thesis is that a new way of classifying languages can be achieved by determining the syntactic typological distance between languages using statistical information obtained from annotated corpora.

Three new strategies are proposed:

1. Quantitative typological classification using MarsaGram:
    a. Considering all MarsaGram properties;
    b. Considering only linear MarsaGram properties.
2. Quantitative typological classification using head and dependent position (head/dependent).
3. Quantitative typological classification using verb and object position (VO/OV).

To define which method to classify languages is the most relevant, it is important to compare the possible new strategies for language comparison to the existing standard classifications which have been applied in NLP studies: genealogical and syntactic classification based on typological databases.

In this study, every PUD language is characterized as a vector for each method of comparison. It means that each language vector contains values for specific features which describe the languages. The different typological approaches presented here vary in terms of these features, and their total number defines the dimension of the language vectors.

Vectors can be compared in different ways via the calculation of distances or similarities between each pair of them. When all pairs are analysed, it is possible to build a dissimilarity or a similarity matrix with the obtained results which can, then, be used for language classification via clustering methods.

We have selected two different approaches to compare the PUD language vectors with the aim of obtaining distance (or dissimilarity) matrices. These methods consider different geometrical aspects and are highly used by other algorithms such as k-nearest neighbours' algorithm (k-NN) for machine learning (Chomboon et al. 2015), as well as UMAP (McInnes et al., 2018) and HDBSCAN (McInnes et al., 2017) for topological and clustering data-analysis.

The selected approaches are presented below:

1. Euclidean distance: in the Euclidean space, the Euclidean distance between two points corresponds to the length of the line segment connecting them. It can be extrapolated to multi-dimensional spaces where the distance (D) between two n-dimensional vectors (v_1 and v_2) can be calculated with:

   $$(4.2)\ D_E(v_1, v_2) = \sqrt{\sum_{i=1}^{n} ({v_1}_i - {v_2}_i)^2}$$

   Values are always positive. In this case, the magnitude of the vectors has an impact on the final result (Spencer, 2013).
2. Cosine similarity: This measure (S_C) corresponds to the cosine of the angle between two vectors (v_1 and v_2). It reflects how two n-dimensional vectors are positioned in terms of orientation in the space, and can be defined as:

   $$(4.3)\ S_c(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|} = \frac{\sum_{i=1}^{n} {v_1}_i {v_2}_i}{\sqrt{\sum_{i=1}^{n} {v_1}_i^{\,2}} \sqrt{\sum_{i=1}^{n} {v_2}_i^{\,2}}}$$

   Thus, if two vectors have the same orientation, their cosine similarity is 1 (the angle between them is 0°), while if they are diametrically opposed to each other (angle 180°) the similarity is equal to -1. In this specific case, the magnitude of the vector does not influence the result as only the orientation is considered (Spencer, 2013).

   As presented above, it corresponds to a measure of similarity (S_C), however, we can transform it into a distance metric (D_C) which can be used to generate a dissimilarity matrix:

   $$(4.4)\ D_C = 1 - S_C$$

As described, each method differs in terms of what is being considered when comparing vectors. It is possible to calculate the Euclidean distance without considering the magnitude of the vectors, for that, a normalization step for each vector is required (the vector norm being equal to 1 after this operation). However, it has been shown that when the Euclidean distance is calculated using normalized vectors, it is possible to convert it to the cosine distance, implying that the squared Euclidean distance is proportional to the cosine distance (Spencer, 2013). Thus, when using cosine and normalized Euclidean distances for ranking elements to compare them, the results of both methods are similar. For this reason, in this thesis, the Euclidean distance measures are calculated with non-normalized vectors to obtain a different ranking of language distances when compared to the cosine method. In this way, it is possible to verify the impact of the magnitude of the vectors when language vectors with syntactic features are analysed.

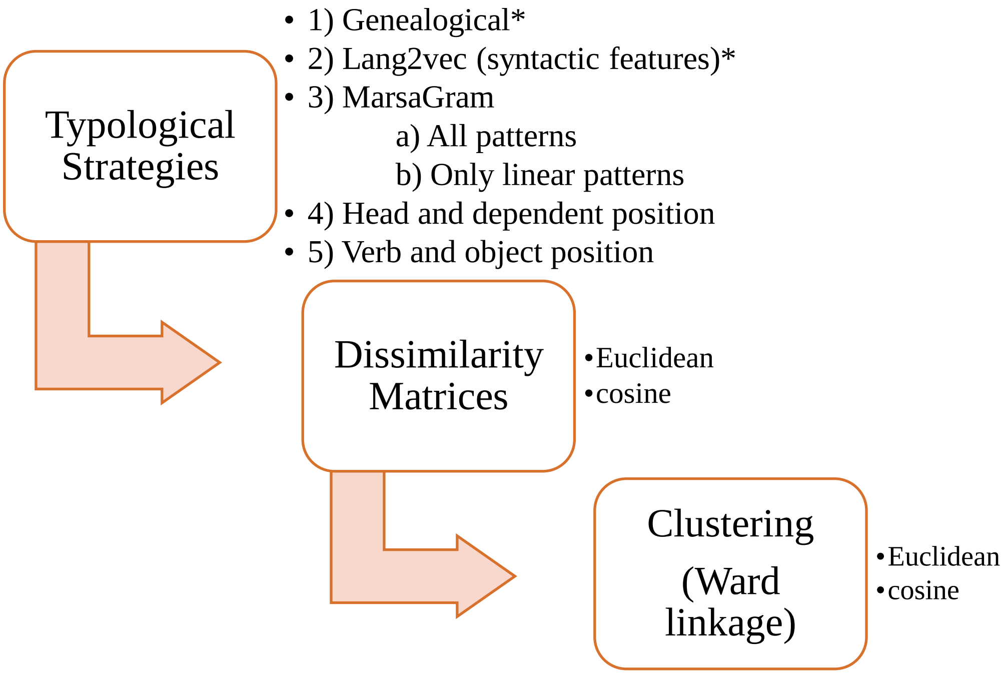
In this study, distance matrices are obtained using R programming language. The Euclidian dissimilarities matrices are obtained via the function dist(), while the cosine ones are calculated (with the equation 4.4 previously presented) from the similarity matrices generated with cosine() function. The distances between pairs of languages obtained for each typological method are used in further sections of this thesis for the calculus of correlation between language dissimilarities and dependency parsing improvement when corpora are combined. In the Annex section, all the dissimilarities matrices generated for this study are displayed.

While dissimilarity matrices provide numerical information, which enables certain analyses regarding language comparison, their format does not allow an intuitive identification of language clusters. The idea of classifying languages using corpora-based typology involves also the identification of groups of languages that are similar regarding some aspects. Thus, besides the distance matrices, a clustering method is also applied to the language vectors to automatically identify possible language clusters.

There exist many different algorithms widely used for data clustering. Each method is based on different arithmetic operations to compare the elements of an ensemble and group them according to defined clustering strategies in terms of the distance metrics between elements and linkage criteria. Concerning corpus-based typological studies, languages are usually classified in hierarchical diagrams (e.g.: Mayer and Cysouw, 2012; Blache et al., 2016; and Liu and Xu, 2012), thus this clustering approach is chosen for this thesis as a primary way for grouping languages with the data provided from the obtained language vectors.

In terms of hierarchical clustering methods, one must define the metric which is used for comparing elements and the linkage (or agglomeration) strategy. In this thesis, the Ward linkage method (Ward, 1963) is applied to both Euclidean and cosine dissimilarity matrices. Instead of minimizing possible distances between pairs of clusters, it 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 our experiments, we applied the method with the automatic identification of the 5 main clusters using a set of different colors. The obtained dendrograms also present dashed lines that correspond to nodes which contains a combination of labels/items, which are not present in the other tree following the solid lines. A simplified scheme of the abovementioned methodology is presented in Figure 4.8.

Figure 4.8. Simplified scheme regarding the methodological steps for language comparison and classification. The “*” symbol indicates the standard language classification methods which will be used as the reference for comparative analysis.


Typological Strategies
• 1) Genealogical*
• 2) Lang2vec (syntactic features)*
• 3) MarsaGram
a) All patterns
b) Only linear patterns
• 4) Head and dependent position
• 5) Verb and object position
Dissimilarity Matrices
•Euclidean
•cosine
Clustering (Ward linkage)
•Euclidean
•cosine

In the following sub-sections, the possible language classifications are described and analysed in detail: 1) the genealogical one obtained using URIEL database, 2) the syntactic typological one generated from lang2vec average vectors concerning syntactic features, 3) the two possible MarsaGram typological classifications, 4) the classification arising from the analysis of the relative position of heads and dependents, and 5) the one established when only verb and object order is considered. For each method, besides the results concerning the obtained language classification, some statistical information concerning the method-specific features and the differences among the generated language vectors is presented to check the existence of particularities and to verify possible biases. In this first step, 10 out of the 24 EU languages will be typologically compared and classified among themselves and among the other PUD languages. The complete analysis of all EU languages (with the other 14 ones) is presented in section 6.

4.3 Genealogical Classification of PUD languages

A brief description of the PUD languages regarding their genealogical families and genera (according to WALS) was presented in Table 4.1. Besides this classification, it is possible to use lang2vec genealogical features to propose a more complete genealogical characterization of them. Lang2vec provides genealogical vectors which are composed by information extracted from Glottolog (Hammarström et al., 2015). Each vector contains of 3,718 features, each one corresponding to a specific linguistic family or genus. The value 1.0 is attributed if the language belongs to the respective family or genus, otherwise, the feature receives the value 0.0.

Regarding PUD language vectors, only 85 different genealogical features present values different than 0.0. Each language is described by a different ensemble of features, varying from 1 for Korean (F_Koreanic) to 12 for French and Spanish. The complete list of descriptive genealogical features of each PUD language is presented in Annexes 8 and 9. One way to compare PUD languages in terms of genealogy is to check the number of similar features which are shared by every pair of languages as shown in Table 4.12.

	arb	**cmn**	**ces**	**eng**	**fin**	**fra**	**deu**	**hin**	**isl**	**ind**	**ita**	**jpn**	**kor**	**pol**	**por**	**rus**	**spa**	**swe**	**tha**	**tur**
arb	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
cmn	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ces	0	0	6	1	0	1	1	1	1	0	1	0	0	4	1	3	1	1	0	0
eng	0	0	1	9	0	1	4	1	3	0	1	0	0	1	1	1	1	3	0	0
fin	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
fra	0	0	1	1	0	12	1	1	1	0	7	0	0	1	9	1	9	1	0	0
deu	0	0	1	4	0	1	6	1	3	0	1	0	0	1	1	1	1	3	0	0
hin	0	0	1	1	0	1	1	7	1	0	1	0	0	1	1	1	1	1	0	0
isl	0	0	1	3	0	1	3	1	6	0	1	0	0	1	1	1	1	4	0	0
ind	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0
ita	0	0	1	1	0	7	1	1	1	0	9	0	0	1	7	1	7	1	0	0
jpn	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0
kor	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
pol	0	0	4	1	0	1	1	1	1	0	1	0	0	5	1	3	1	1	0	0
por	0	0	1	1	0	9	1	1	1	0	7	0	0	1	11	1	11	1	0	0
rus	0	0	3	1	0	1	1	1	1	0	1	0	0	3	1	4	1	1	0	0
spa	0	0	1	1	0	9	1	1	1	0	7	0	0	1	11	1	12	1	0	0
swe	0	0	1	3	0	1	3	1	4	0	1	0	0	1	1	1	1	6	0	0
tha	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0
tur	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6

Table 4.12. Comparison of PUD languages regarding the number of similar genealogical features. The diagonal values (in blue) correspond to the total number of features with the value 1.0 in the vector of each language.

Thus, from this table, it is possible to notice that Arabic, Chinese, Finnish, Indonesian, Japanese, Korean, Thai, and Turkish have no shared features with other PUD languages. On the other hand, Portuguese and Spanish have 11 features in common. If this table was to be used for choosing the best possible combination of similar languages, the results for each PUD language would be as presented in Table 4.13.

	Closest language in PUD collection
arb	-
cmn	-
ces	pol
eng	deu
fin	-
fra	por / spa
deu	eng
hin	ces / eng / fra / deu / isl / ita / pol / por / rus / spa / swe
isl	swe
ind	-
ita	fra / por / spa
jpn	-
kor	-
pol	ces
por	spa
rus	ces
spa	por
swe	isl
tha	-
tur	-

Table 4.13. Closest languages in terms of the number of shared lang2vec genealogical features.

Although providing valuable information to guide the choice of the best possible language association, this method does not allow languages to be compared in a more fine-grained way. For example, in this restricted scenario (PUD collection), there are 11 possible languages that are considered the closest to Hindi (all PUD Indo-European languages, thus sharing with it this specific genealogical feature). This method of selecting the closest language is also problematic for the abovementioned languages which do not share any phylogenetic feature with other PUD languages (such as Thai, Turkish, Korean, etc).

When applying the clustering method (with Ward linkage) for both Euclidean and cosine dissimilarity matrices (Annexes 10 and 11) obtained from the comparison of language vectors composed with genealogical features, the obtained clusters are presented in Figure 4.9 and 4.10 correspondingly.

In Figure 4.9, corresponding to the Euclidean dissimilarity matrix, both Romance and Germanic language clusters are easily noticeable on the left side. The Thai language is placed in an isolated cluster in the middle of the dendrogram. It is also possible to recognize the Slavic language sub-cluster together, in a large group (in purple), with the languages which do not share genealogical features with other PUD languages. Hindi is also located in this large cluster, although sharing one feature with all Indo-European languages. On the other hand, in Figure 4.10 (built with cosine distances), not only Romance and Germanic clusters are easily identified, but also the Slavic one. The PUD languages which have no shared genealogical features form a big cluster in the middle of the figure and are not grouped with any other PUD language with common features. The Romance languages are isolated from the other Indo-European languages most probably because of the clustering algorithm as the number of shared features between these languages is much higher when compared to others. The relatively small number of shared features between Slavic languages is also the reason why Hindi is closer to them, although it shares the same number of features with all Indo-European languages.

Figure 4.9. Cluster dendrogram obtained from the Euclidean dissimilarity matrix calculated with the comparison of the PUD genealogical language vectors.

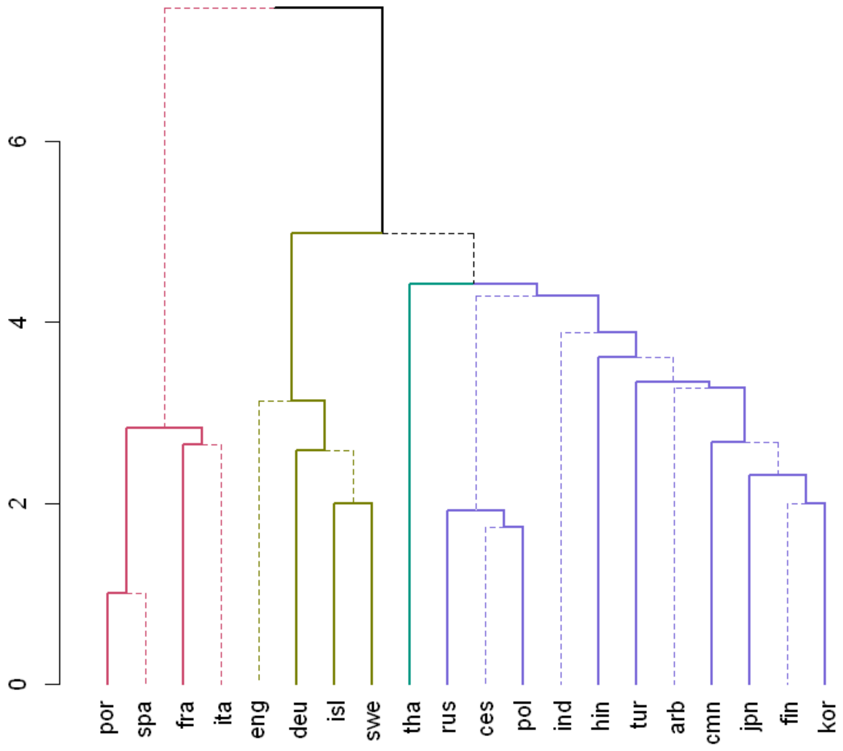
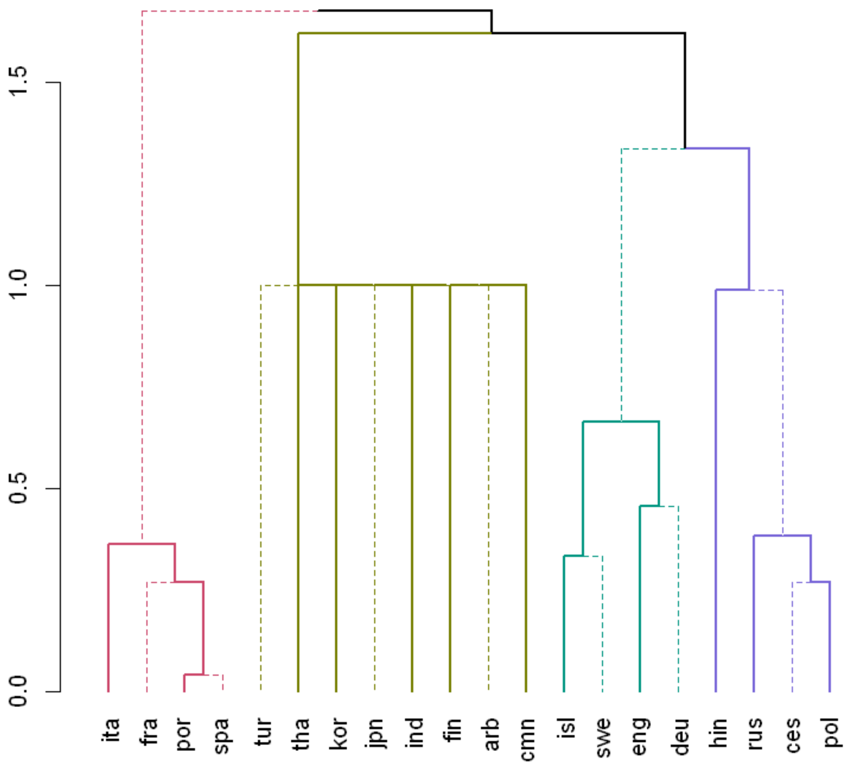


Figure 4.10. Cluster dendrogram obtained from the cosine dissimilarity matrix calculated with the comparison of the PUD genealogical language vectors.



This bias concerning the variety in terms of the number of features that describe each language is also the reason why some discrepancy is observed in the dendrogram obtained with Euclidean distances. It is caused by the way the language vectors are constructed (Table 4.12). The value of comparison of Thai to itself is 10, while for Arabic, it is only 6. It should be the same value in both cases because Thai is not more similar to Thai than Arabic is to Arabic. Thus, when using lang2vec genealogical features, it is important to compare normalized vectors (as it is the case for the cosine distances). Thus, regarding the genealogical classification of languages, the cosine dendrogram provides a more accurate overview of PUD languages.

Therefore, concerning the genealogical classification, the visualization through a dendrogram generated from dissimilarity matrices, especially the Euclidean one, may give the false impression that some languages are closer to others genetically while in reality, they are not. However, concerning all possible linkage strategies for the clustering algorithm, the Ward one gives the best results when compared to the expected language groups. All the other dendrograms obtained using the other available agglomeration methods in R did not provide coherent language clusters, thus, showing the choice of Ward strategy is more relevant for linguistic studies such as the one presented in this thesis.

4.4 Classification of PUD Languages From lang2vec Syntactic Vectors

As previously explained, using lang2vec syntactic features (“syntax_average”), it is possible to analyse languages regarding syntactic typology by comparing the language vectors fulfilled with values corresponding to each syntactic feature. We have chosen to consider the average vectors, which contain the ensemble of all syntactic information from URIEL (consisting of a total of 103 features), to conduct this examination as they encompass more complete information compared to vectors specifically built over single typological databases.

The focus of this thesis is the analysis of how languages can be classified regarding syntactic features, thus, the phonological vectors provided are not considered here. Some studies have proved the relevance of phonological information in strategies for combining languages (e.g.: De Lhoneux et al. (2018), Üstün et al., 2020), thus, this complementary typological material should be tested in future work with the best syntactic strategies defined in this study.

The analysis of the existing information concerning lang2vec syntactic vectors shows that there is a great discrepancy in terms of the availability of syntactic features among PUD languages. It varies from 66 features with a valid value for Arabic to 103 (all possible ones) for English. The distribution of the number of valid features per language is detailed in Table 4.14 and Figure 4.11.

The inequality in terms of valid features is one of the drawbacks of using vectors composed of information provided by typological databases. These resources are based on information extracted from grammars and other references, thus, as the number of described syntactic phenomena in the literature varies enormously, it is not possible to have complete vectors for all languages. This problem has been identified in previous studies (e.g.: Levshina, 2022), and is even more problematic for minority languages.

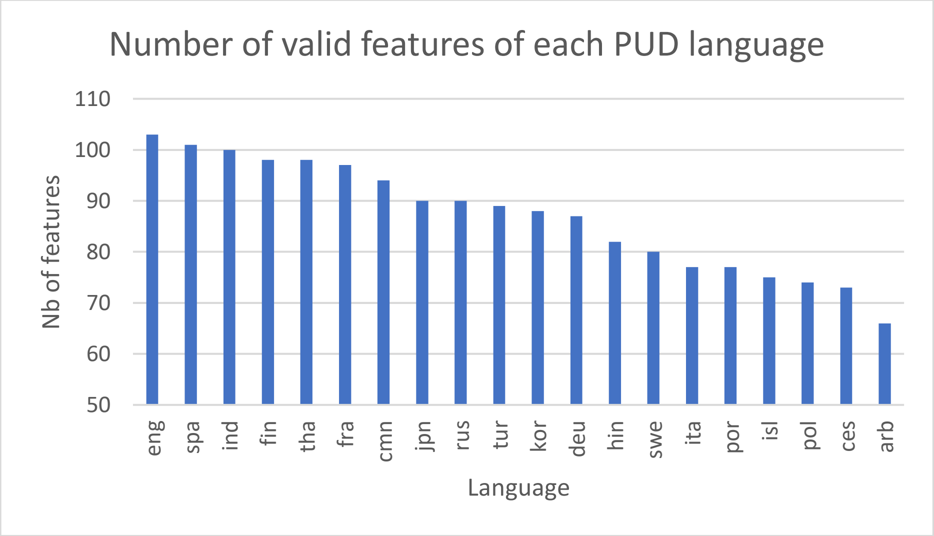
Regarding PUD languages, it is possible to notice that 8 languages have at least 90 valid features, while 6 of them have less than 80. Czech has a large amount of annotated data in Universal Dependencies but, regarding lang2vec syntactic vector, it is less described than Thai which has only PUD corpus in UD. Thus, in terms of syntactic description in typological databases, Czech is less resourced than Thai.

Even though the minimum number of valid features in PUD collection is 66 (Arabic), when checking the number of common features of all PUD languages with valid values, the final amount is 41. The complete list of common features is presented in Annex 12.

Language	Number of features with non-missing values
arb	66
cmn	94
ces	73
eng	103
fin	98
fra	97
deu	87
hin	82
isl	75
ind	100
ita	77
jpn	90
kor	88
pol	74
por	77
rus	90
spa	101
swe	80
tha	98
tur	89

Table 4.14. Number of lang2vec average syntactic features with non-missing values for each PUD language.

Figure 4.11. Graph with the number of lang2vec average syntactic features with non-missing values for each PUD language.



In terms of provided syntactic information, the forty-one common features express the most common observed word order phenomena between:

1. Subject, verb, and object (e.g.: SVO, SOV, SUBJECT_BEFORE_VERB);
2. Adposition and noun (e.g.: ADPOSITION_BEFORE_NOUN);
3. Possessor and noun (e.g.: POSSESSOR_AFTER_NOUN);
4. Adjective and noun (e.g.: ADJECTIVE_AFTER_NOUN);
5. Demonstrative and noun (e.g.: DEMONSTRATIVE_WORD_BEFORE_NOUN);
6. Numeral and noun (e.g.: NUMERAL_AFTER_NOUN);
7. Negative word and verb (e.g.: NEGATIVE_WORD_BEFORE_VERB);
8. Degree word and adjective (e.g.: DEGREE_WORD_BEFORE_ADJECTIVE);
9. Subordinator word and clause (e.g.: SUBORDINATOR_WORD_AFTER_CLAUSE);
10. Polar question particle position: initial or final (e.g.: POLARQ_MARK_INITIAL);
11. Existence of demonstrative prefix or suffix (e.g.: DEMONSTRATIVE_PREFIX);
12. Existence of negative prefix or suffix (e.g.: NEGATIVE_PREFIX);
13. Existence of TEND prefix or suffix (e.g.: TEND_SUFFIX);
14. Existence of case mark, enclitic, proclitic, prefix, and suffix (e.g.: CASE_ENCLITIC).

When comparing this information with the word order phenomena used by Hawkins (1983) in his proposal of language types, it is noticeable that the list of common PUD features provided by lang2vec is more complete, including information such as the existence of polar question particles and position of the negative word. However, the genitive position in relation to the noun is not described in this set of lang2vec features although being present in Hawkins' analysis.

It is possible to identify six common features, among the forty-one selected, for which all PUD languages have the same value, therefore, being less relevant when differentiating the languages typologically.

Table 4.15 presents the list of these features and their respective value.

Feature	**Value**
S_SUBJECT_BEFORE_OBJECT	1.0
S_SUBJECT_AFTER_OBJECT	0.0
S_DEMONSTRATIVE_PREFIX	0.0
S_DEMONSTRATIVE_SUFFIX	0.0
S_TEND_PREFIX	0.0
S_CASE_PREFIX	0.0

Table 4.15. Entries from the common list of features for which all PUD languages have the same value according to the URIEL database.

From the information provided in Table 4.15, it is also possible to notice the redundancy concerning some features which has been commented on by Ponti et al. (2019): both S_SUBJECT_BEFORE_OBJECT and S_SUBJECT_AFTER_OBJECT describe the same ordering when one feature receives the value 1.0, the other gets 0.0 and vice-versa.

It is also possible to identify 13 syntactic features for which most PUD languages (more than 15 out of the 20) have the same value and the languages which differ from the majority. These results are presented in Table 4.16.

Features	**Most common value**	**Differing language**
S_VSO	0.0	arb (1.0), tur (0.33)
S_SUBJECT_BEFORE_VERB	1.0	arb (0.0)
S_OBJECT_AFTER_VERB	1.0	hin (0.0), jpn (0.0), kor (0.0), tur (0.33)
S_ADPOSITION_BEFORE_NOUN	1.0	fin (0.5), jpn (0.0), kor (0.0), tur (0.0)
S_DEMONSTRATIVE_WORD_BEFORE_NOUN	1.0	ind (0.0), tha (0.0)
S_DEMONSTRATIVE_WORD_AFTER_NOUN	0.0	isl (0.5), ind (1.0), pol (0.5), tha (1.0)
S_NEGATIVE_SUFFIX	0.0	eng (0.5), jpn (0.5), tur (1.0)
S_TEND_SUFFIX	1.0	tha (0.0)
S_CASE_PROCLITIC	0.0	fra (1.0)
S_CASE_ENCLITIC	0.0	jpn (1.0)
S_DEGREE_WORD_BEFORE_ADJECTIVE	1.0	arb (0.0), tha (0.0)
S_SUBORDINATOR_WORD_BEFORE_CLAUSE	1.0	jpn (0.0), kor (0.5), tur (0.0)
S_SUBORDINATOR_WORD_AFTER_CLAUSE	0.0	jpn (1.0), kor (1.0), tur (1.0)

Table 4.16. List of common features for which most PUD languages have the same value and languages which differ from the majority.

As expected, in most cases, the languages which have different values when compared to the majority are the non-Indo-European ones. However, it is possible to see that some Indo-European languages also differ regarding some specific features, such as Polish regarding the position of the demonstrative and noun, English concerning the presence of negative suffixes, and French (S_CASE_PROCLITIC).

As previously mentioned, from the 103 possible syntactic features ("syntax_average"), only 41 of them have valid values for all PUD languages. Therefore, for the typological analysis, the first step was to use these values to generate the language vectors (41 dimensions) which can be represented as:

$$(4.5)\ vector_{language} = [value_{feature_1}, value_{feature_2}, \dots, value_{feature_{41}}]$$

With these PUD language vectors, the dissimilarity matrices (both Euclidean and cosine) were generated as formerly explained. The complete results are presented in Annexes 13 and 14. These matrices provide for each pair of languages the distance between the syntactic vectors; however, it does not provide an intuitive way for identifying possible language groups. Thus, cluster analysis, as described previously, is necessary for this aim. The dendrograms obtained using R hclust() function with Ward linkage method are presented in Figures 4.12 and 4.13.

Figure 4.12. Cluster dendrogram obtained from the Euclidean dissimilarity matrix calculated with the comparison of the PUD lang2vec "syntax_average" language vectors.

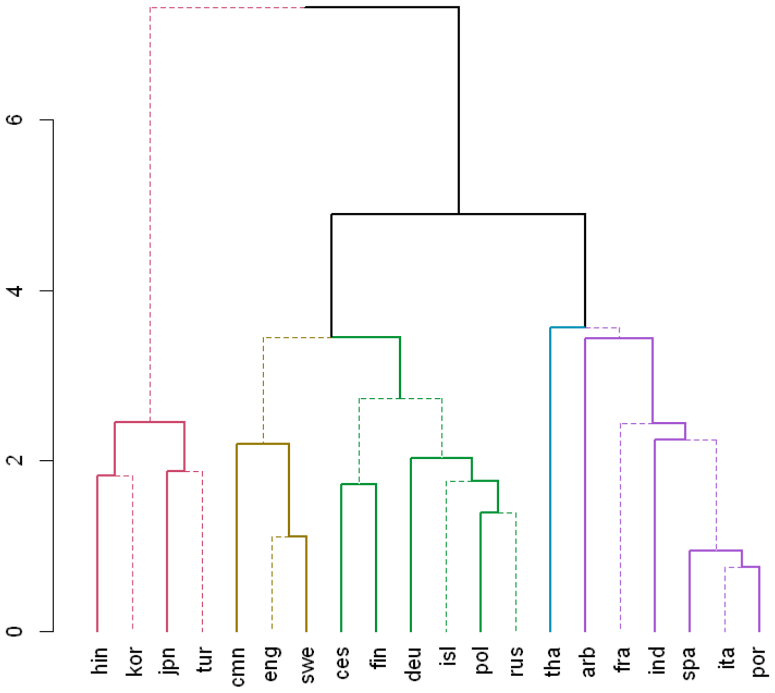
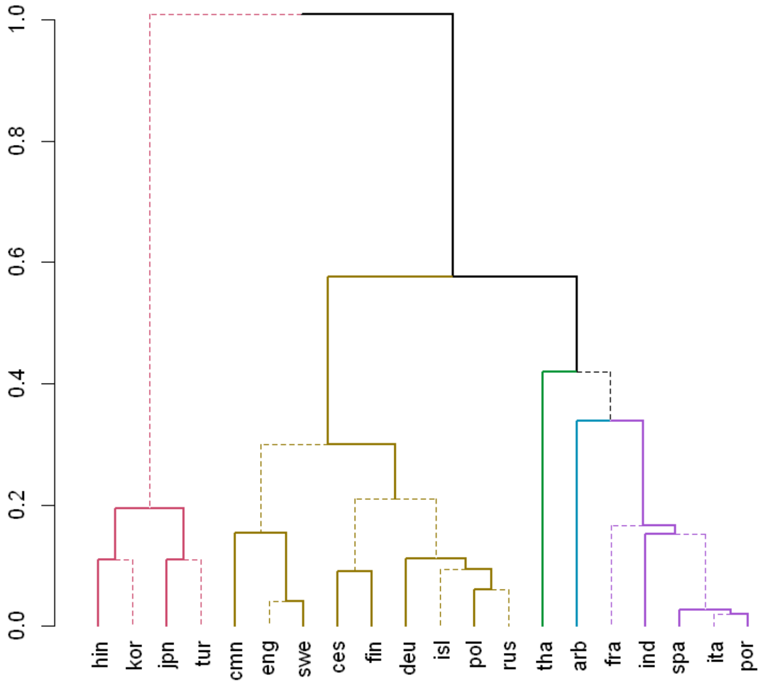


Figure 4.13. Cluster dendrogram obtained from the cosine dissimilarity matrix calculated with the comparison of the PUD lang2vec "syntax_average" language vectors.



When comparing the obtained dendrograms, one can observe that the language clusters are quite similar for both distance metrics. One different aspect is the magnitude of the dissimilarity values (y-axis), higher for the Euclidean graph. The other difference concerns the colour-clustering. While, in the Euclidean dendrogram, the central cluster is divided into two sub-groups (one composed of Chinese, English, and Swedish, and the other of Czech, Finnish, German, etc.), for the cosine dendrogram, there is no sub-group division. However, when the cosine distance is considered, Arabic forms a sub-group of the large cluster composed of Thai, Indonesian, and the Romance languages, while in the Euclidean dendrogram, it 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 by German and Icelandic (grouped with Polish and Russian). Regarding the Slavic languages, although Polish and 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 Romance languages when compared to the others in the PUD collection.

In Table 4.17, the three main clusters of both dendrograms are detailed in terms of the features which have the same value for all the composing languages. It is possible to observe that the isolated cluster formed by Hindi, Japanese, Korean and Turkish is composed of SOV languages with postpositions and 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.

Languages	Syntactic features with value 1.0	Syntactic features with value 0.0
hin, jpn, kor, tur	S_SOV, S_SUBJECT_BEFORE_VERB, S_OBJECT_BEFORE_VERB, S_ADPOSITION_AFTER_NOUN, S_POSSESSOR_BEFORE_NOUN, S_ADJECTIVE_BEFORE_NOUN, S_DEMONSTRATIVE_WORD_BEFORE_NOUN, S_NUMERAL_BEFORE_NOUN, S_TEND_SUFFIX, S_CASE_MARK, S_DEGREE_WORD_BEFORE_ADJECTIVE	S_ADPOSITION_BEFORE_NOUN, S_POSSESSOR_AFTER_NOUN, S_ADJECTIVE_AFTER_NOUN, S_DEMONSTRATIVE_WORD_AFTER_NOUN, S_NEGATIVE_PREFIX, S_CASE_PROCLITIC
cmn, ces, eng, fin, deu, isl, pol, rus, swe	S_SVO, S_SUBJECT_BEFORE_VERB, S_OBJECT_AFTER_VERB, S_ADJECTIVE_BEFORE_NOUN, S_DEMONSTRATIVE_WORD_BEFORE_NOUN, S_NUMERAL_BEFORE_NOUN, S_TEND_SUFFIX, S_DEGREE_WORD_BEFORE_ADJECTIVE, S_SUBORDINATOR_WORD_BEFORE_CLAUSE	S_VSO, S_VOS, S_OSV, S_CASE_PROCLITIC, S_CASE_ENCLITIC
arb, fra, ind, ita, por, spa, tha	S_OBJECT_AFTER_VERB, S_ADPOSITION_BEFORE_NOUN, S_POSSESSOR_AFTER_NOUN, S_ADJECTIVE_AFTER_NOUN, S_NEGATIVE_WORD_BEFORE_VERB, S_SUBORDINATOR_WORD_BEFORE_CLAUSE	S_OVS, S_ADPOSITION_AFTER_NOUN, S_POSSESSOR_BEFORE_NOUN, S_DEMONSTRATIVE_PREFIX, S_DEMONSTRATIVE_SUFFIX, S_NEGATIVE_PREFIX, S_NEGATIVE_SUFFIX, S_TEND_PREFIX, S_SUBORDINATOR_WORD_AFTER_CLAUSE

Table 4.17. List of common features of each large cluster from the dendrograms obtained with lang2vec syntactic vectors. The features presented in Table 4.16 which are common for all PUD languages are not present here.

4.5 Quantitative Typological Classification Using MarsaGram

As explained previously, MarsaGram allows the extraction of syntactic properties from CFG rules that are inferred from annotated corpora in terms of dependency relations. The tool provides a set of four types of properties: “precede”, “require”, “exclude” and “unicity” relating tokens inside a subtree of the syntactic structure (tokens being characterized by their part-of-speech and their dependency relation labels).

Only the patterns of the type “precede” (or linear) reveal word order phenomena at the surface level (sentence). The other three kinds of patterns concern the presence or absence of determined pair of tokens inside the subtrees but do not provide information concerning their order at the surface level.

Blache et al. (2016) conducted an experiment on quantitative typological classification of ten languages using MarsaGram. They compared two different approaches: one using all types of MarsaGram patterns, and one considering only the linear type. The obtained classifications were correlated to the genealogical one. In this specific trial, according to the authors, the second option was more convincing.

This thesis deals with a larger sample in terms of the number of languages and our aim is not to find the best fit between MarsaGram and the genealogical classification, but to analyse the possible correlations between quantitative approaches of typological classification and results obtained when languages are combined for better dependency parsing results. Additionally, this study differs from the one presented by Blache et al. (2016) as the analysis is conducted with parallel corpora, thus, using datasets with the same size and same semantic information, thus avoiding bias regarding genre and the quantity of extracted syntactic information from each corpus.

Therefore, both scenarios proposed by Blache et al. (2016) are considered, enabling the creation of two typological classifications with MarsaGram:

1. All properties approach, considering patterns of the four possible MarsaGram types.
2. Linear approach, considering only the patterns of the “precede” type.

For each corpus, MarsaGram provides a tsv file with all the extracted properties (components and frequency) which allows the construction of a language vector composed of n values, n corresponding to the number of extracted patterns, as schematized in:

$$(4.6)\ vector_{language} = [frequency_{property_1}, frequency_{property_2}, \dots, frequency_{property_n}]$$

Consequently, the final set of MarsaGram patterns is composed of all the observed ones in the PUD collection, which are not necessarily present in the individual analysis of each PUD language. If a language does not present a certain property in its MarsaGram results, the value of zero is attributed to the corresponding language vector. Regarding PUD languages, the total amount of extracted patterns is 158,755. Table 4.18 presents the distribution in terms of types and percentages concerning the whole ensemble of extracted MarsaGram properties from PUD collection.

Type of property	Number of properties	%
Precede (linear)	21,242	13.38
Exclude	129,180	81.37
Unicity	2,144	1.35
Require	6,189	3.90

Table 4.18. Distribution of the final set of MarsaGram patterns in terms of property types.

It is possible to notice that the great majority (more than 80%) of extracted properties is of type “exclude”: one component is never observed with the other one inside a subtree with a specified head. Furthermore, Table 4.19 shows the number of patterns extracted for each PUD corpus.

We can also observe that, even though only parallel corpora are considered, the number of extracted properties varies considerably among different languages: less than 10,000 for Japanese and Korean languages and more than 20,000 for English, Hindi, and Icelandic languages. This is probably due to the different number of tokens (more function words in some langs, more inflection in others). The other PUD languages have an amount of properties closer to the average (15,790).

Language	Number of properties
arb	16,460
cmn	18,070
ces	16,706
eng	20,517
fin	13,374
fra	13,656
deu	13,225
hin	22,106
isl	22,199
ind	10,889
ita	15,380
jpn	5,226
kor	8,860
pol	17,592
por	13,994
rus	16,827
spa	14,021
swe	19,795
tha	19,403
tur	17,508

Table 4.19. Number of extracted patterns for each PUD language using MarsaGram.

When only linear properties are considered (table 4.20), it is possible to notice that Japanese and Korean have the least amount of extracted properties (less than 1,500) and that English, Hindi, and Icelandic languages have higher amounts (more than 2,500), however, Chinese, Swedish and Thai languages are also well represented. The average number of properties is 2,130.

Concerning the final set of MarsaGram patterns (i.e.: the combination of all the extracted patterns). Its size is around 10 times higher than the extent of the set of properties extracted per language (Table 4.21), and that is valid for both scenarios (all properties and only linear).

When analysing the set composed of all types of MarsaGram information, it is possible to examine how these patterns are distributed among PUD corpora. Table 4.22 presents this quantitative investigation in terms of the amount: a) of properties occurring in all PUD corpora, b) of the patterns occurring in more than ten corpora, and c) of the ones with occurrences in only one PUD corpus.

Language	Number of linear properties
arb	2,208
cmn	2,552
ces	2,053
eng	2,599
fin	1,764
fra	1,928
deu	1,826
hin	2,842
isl	2,710
ind	1,664
ita	2,090
jpn	1,287
kor	1,418
pol	2,257
por	2,023
rus	2,072
spa	1,996
swe	2,508
tha	2,665
tur	2,144

Table 4.20. Number of extracted linear patterns for each PUD language using MarsaGram.

	Average	Final set size
All patterns	15,790	158,755
Linear patterns	2,130	21,242

Table 4.21. Comparison between the average number of patterns extracted from each PUD corpus, and the final set of concatenated properties.

	Number of properties	%
Occurring in only one corpus	108,006	68.03
Occurring in more than 10 corpora	3,128	1.97
Occurring in all corpora	78	0.05

Table 4.22. Distribution of properties (all types) inside PUD corpora and respective % of the total selected properties.

The same examination can be conducted considering only the linear patterns (Table 4.23).

	Number of properties	**%**
Occurring in only one corpus	14,569	68.59
Occurring in more than 10 corpora	467	2.20
Occurring in all corpora	10	0.05

Table 4.23. Distribution of linear properties inside PUD corpora and the respective % of the total selected properties.

It is noticeable that more than half of the patterns in the final sets (i.e.: higher than 68%) appears in only one PUD languages. The distribution in both cases is similar: very few properties occur in all corpora and only a small proportion is present more than 10 corpora (i.e.: lower than 3%).

Besides the analysis regarding the overall distribution of patterns, it is also possible to analyse the word order phenomena that occur in all PUD languages. The list of the 10 linear patterns common to all corpora from PUD collection is presented in Table 4.24.

Linear patterns in all PUD languages
VERB-+_precede_PRON-nsubj_NOUN-obj
VERB-+_precede_NOUN-nsubj_NOUN-obj
VERB-+_precede_PROPN-nsubj_NOUN-obj
NOUN-+_precede_*_NOUN-appos
VERB-+_precede_CCONJ-cc_*
PROPN-+_precede_*_NOUN-appos
NOUN-+_precede_*_PROPN-appos
NOUN-+_precede_CCONJ-cc_*
PROPN-+_precede_CCONJ-cc_*
PROPN-+_precede_*_PROPN-appos

Table 4.24. List of linear patterns which occur in all PUD languages (i.e.: frequency higher than 0.0). The “*” symbol indicates that the element corresponds to the head of the subtree.

The patterns which occur in all PUD corpora describe:

1) Subject (nsubj) and object (obj) positions in subtrees whose head is a verb. The subject preceding the object in all cases. These patterns are equivalent to the lang2vec features "S_SUBJECT_BEFORE_OBJECT" with value 1.0, and "S_SUBJECT_AFTER_OBJECT" with value 0.0, which are common to all PUD languages.
2) Coordinative conjunction (cc) and root (*) position in subtrees ruled by heads with different part-of-speech. The conjunction preceding the root in all patterns.
3) Appositional modifier (apos[33]) and root (*) ordering in different types of subtrees. The root preceding the appositional modifier for all the listed patterns.

Alongside the 12 common linear patterns, all PUD corpora share in common:

i) 48 "exclude" patterns;
ii) 20 "unicity" patterns.

All of these properties have a noun, a verb, or a proper noun as the head of the subtree. No "require" property is common to all PUD languages. The complete list of all shared patterns (82 in total) concerning all MarsaGram possible relations is presented in Annex 15.

The next two sub-sections present the dendrograms corresponding to: a) all MarsaGram patterns, and b) only linear ones.

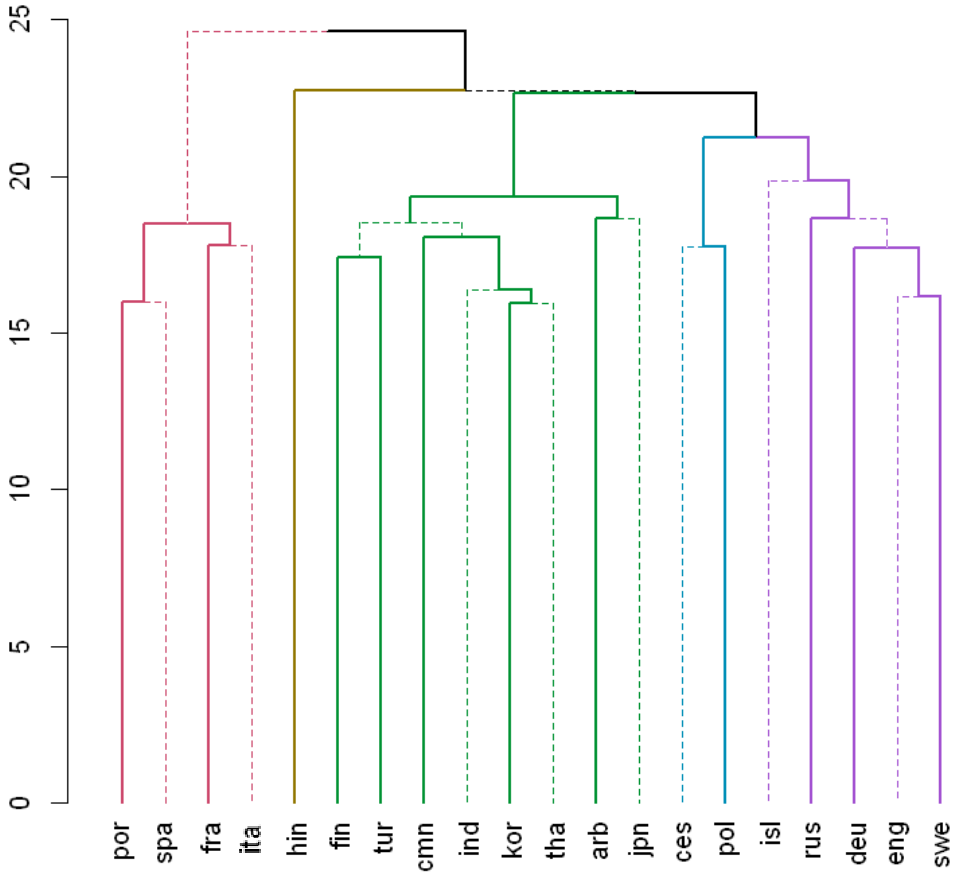
a) All patterns:

The Euclidean and cosine dissimilarity matrices obtained from the comparison of MarsaGram language vectors (all patterns) are presented in Annexes 16 and 17 respectively. And, with these matrices, the clustering analysis is conducted (Figures 4.14 and 4.15).

---

[33] The homogeneity concerning the appositional modifier position is explained by its definition in the Universal Dependencies framework: "An appositional modifier of a noun is a nominal immediately following the first noun that serves to define, modify, name, or describe that noun".

Figure 4.14. Cluster dendrogram obtained from the Euclidean dissimilarity matrix calculated with the comparison of the PUD MarsaGram language vectors (all patterns).



The dendrogram generated with the Euclidean distance matrix when all MarsaGram patterns are considered shows that all 4 Romance languages are grouped in a cluster on the left side of the figure. It is possible to identify a large cluster in the middle of the dendrogram (in green) composed mostly by languages which do not have close related ones in terms of genealogical features. Finnish is a language from the Uralic family but it is grouped with Turkish in a small sub-cluster (both languages that are known for sharing similar syntactic features due to their agglutinative characteristic). Hindi is positioned in between the Romance languages and the large green cluster. Furthermore, on the right side of the dendrogram, we can observe a large group formed by Slavic and Germanic languages. Czech and Polish are gathered in a sub-cluster, while Russian is positioned in between Icelandic and the other Germanic languages.

To better understand why some languages from distinct linguistic families were classed together, we conducted a detailed analysis of the ensemble of MarsaGram features. The idea is to identify, in each specific case, the main features for which the examined languages have similar values which differ from the majority of the language sample.

Thus, we created a python script that: a) identifies the features for which the values for the selected languages are higher than all the other PUD ones (i.e.: the phenomenon is more frequent in the selected languages when compared to the rest of the language-set), b) find the features for which the analysed languages have values lower than all the other PUD languages (i.e.: the phenomenon is less frequent in these languages when compared to the other PUD ones). The script also identifies the features for which the selected languages have 0 as value while all the other languages have positive frequency, and the cases where only the examined languages have values higher than 0.

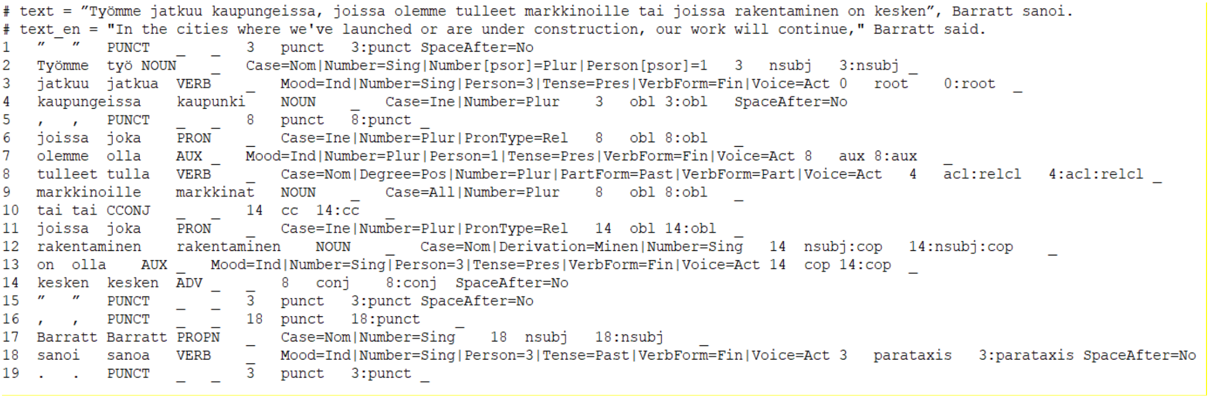
Thus:

1) Regarding Finnish and Turkish, the analysis shows that for 467 MarsaGram features, these two languages have values with higher frequency than the rest of the PUD sample. From them, 421 correspond to the property “exclude”, 14 describe “unicity” patterns, 3 concern “require” property, and 29 features concern the linear property (“precede”). Moreover, from these 467 features, in 184 cases (171 “exclude”, 3 “unicity”, 2 “require”, and 8 “precede”), all the other languages presented a frequency value of 0 for the specific phenomena.

   The “unicity” patterns exclusively present in the Finnish and Turkish corpora are: “NOUN-+_unicity_INTJ-discourse”, “ADJ-+_unicity_ADJ-nmod:poss”, and “ADV-+_unicity_NOUN-nmod:poss”. The “require” ones: “NOUN-+_require_INTJ-discourse_PUNCT-punct”, and “AUX-+_require_NOUN-obl_NOUN-obj”. Regarding the “precede” patterns, only 2 present a frequency higher than 0.02: “ADV-+_precede_PRON-obl_*” and “ADV-+_precede_NOUN-nmod:poss_*”. Each one of these two patterns are present in 3 sentences of the Finnish PUD corpus. Figures 4.15 and 4.16 present an example of each pattern.

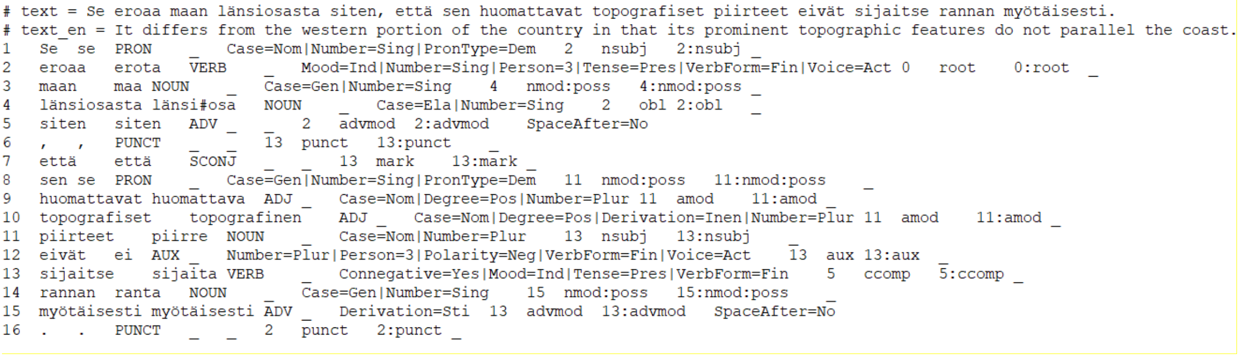
   On the other hand, Finnish and Turkish present only one feature with 0 as associated value while all the other languages have positive ones: “NOUN-+_exclude_NUM-nummod_NOUN-nsubj”.

Figure 4.15. Example of sentence from the Finnish PUD corpus with the pattern “ADV-+_precede_PRON-obl_*” where the token “joissa” (“id” = 11, UPOS = “PRON”) precedes the token “kesken” (“id” =14, UPOS= “ADV”) which is also the head of the sub-tree.



```
# text = ”Työmme jatkuu kaupungeissa, joissa olemme tulleet markkinoille tai joissa rakentaminen on kesken”, Barratt sanoi.
# text_en = "In the cities where we've launched or are under construction, our work will continue," Barratt said.
1   ”   ”   PUNCT   _   _   3   punct   3:punct SpaceAfter=No
2   Työmme  työ NOUN    _   Case=Nom|Number=Sing|Number[psor]=Plur|Person[psor]=1   3   nsubj   3:nsubj _
3   jatkuu  jatkua  VERB    _   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act 0   root    0:root  _
4   kaupungeissa    kaupunki    NOUN    _   Case=Ine|Number=Plur    3   obl 3:obl   SpaceAfter=No
5   ,   ,   PUNCT   _   _   8   punct   8:punct _
6   joissa  joka    PRON    _   Case=Ine|Number=Plur|PronType=Rel   8   obl 8:obl   _
7   olemme  olla    AUX _   Mood=Ind|Number=Plur|Person=1|Tense=Pres|VerbForm=Fin|Voice=Act 8   aux 8:aux   _
8   tulleet tulla   VERB    _   Case=Nom|Degree=Pos|Number=Plur|PartForm=Past|VerbForm=Part|Voice=Act   4   acl:relcl   4:acl:relcl _
9   markkinoille    markkinat   NOUN    _   Case=All|Number=Plur    8   obl 8:obl   _
10  tai tai CCONJ   _   _   14  cc  14:cc   _
11  joissa  joka    PRON    _   Case=Ine|Number=Plur|PronType=Rel   14  obl 14:obl  _
12  rakentaminen    rakentaminen    NOUN    _   Case=Nom|Derivation=Minen|Number=Sing   14  nsubj:cop   14:nsubj:cop    _
13  on  olla    AUX _   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act 14  cop 14:cop   _
14  kesken  kesken  ADV _   _   8   conj    8:conj  SpaceAfter=No
15  ”   ”   PUNCT   _   _   3   punct   3:punct SpaceAfter=No
16  ,   ,   PUNCT   _   _   18  punct   18:punct    _
17  Barratt Barratt PROPN   _   Case=Nom|Number=Sing    18  nsubj   18:nsubj    _
18  sanoi   sanoa   VERB    _   Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act 3   parataxis   3:parataxis SpaceAfter=No
19  .   .   PUNCT   _   _   3   punct   3:punct _
```

Figure 4.16. Example of sentence from the Finnish PUD corpus with the pattern “ADV-+_precede_NOUN-nmod:poss_*” where the token “rannan” (“id” = 14, UPOS = “NOUN”, DEPREL = “nmod:poss”) precedes the token “myötäisesti” (“id” =15, UPOS= “ADV”) which is also the head of the sub-tree.



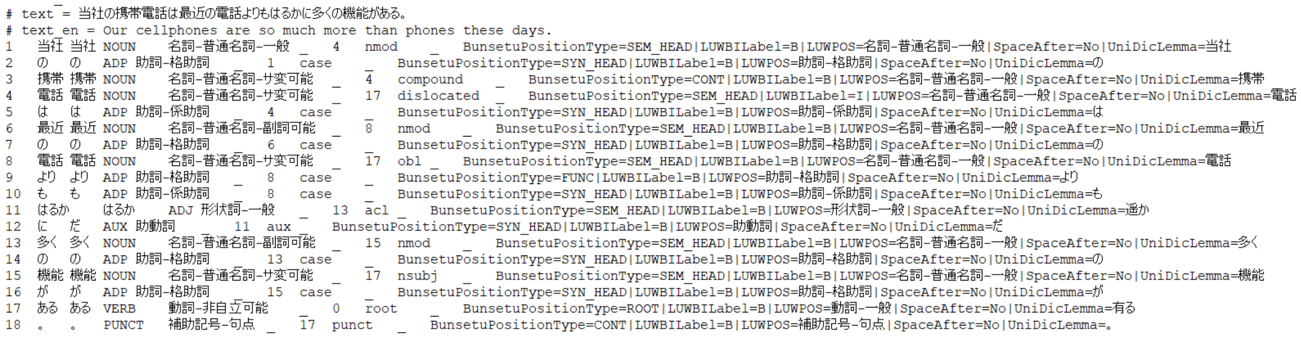
```
# text = Se eroaa maan länsiosasta siten, että sen huomattavat topografiset piirteet eivät sijaitse rannan myötäisesti.
# text_en = It differs from the western portion of the country in that its prominent topographic features do not parallel the coast.
1   Se  se  PRON    _   Case=Nom|Number=Sing|PronType=Dem   2   nsubj   2:nsubj _
2   eroaa   erota   VERB    _   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act 0   root    0:root  _
3   maan    maa NOUN    _   Case=Gen|Number=Sing    4   nmod:poss   4:nmod:poss _
4   länsiosasta länsi#osa   NOUN    _   Case=Ela|Number=Sing    2   obl 2:obl   _
5   siten   siten   ADV _   _   2   advmod  2:advmod    SpaceAfter=No
6   ,   ,   PUNCT   _   _   13  punct   13:punct    _
7   että    että    SCONJ   _   _   13  mark    13:mark _
8   sen se  PRON    _   Case=Gen|Number=Sing|PronType=Dem   11  nmod:poss   11:nmod:poss    _
9   huomattavat huomattava  ADJ _   Case=Nom|Degree=Pos|Number=Plur 11  amod    11:amod _
10  topografiset    topografinen    ADJ _   Case=Nom|Degree=Pos|Derivation=Inen|Number=Plur 11  amod    11:amod _
11  piirteet    piirre  NOUN    _   Case=Nom|Number=Plur    13  nsubj   13:nsubj    _
12  eivät   ei  AUX _   Number=Plur|Person=3|Polarity=Neg|VerbForm=Fin|Voice=Act    13  aux 13:aux  _
13  sijaitse    sijaita VERB    _   Connegative=Yes|Mood=Ind|Tense=Pres|VerbForm=Fin    5   ccomp   5:ccomp _
14  rannan  ranta   NOUN    _   Case=Gen|Number=Sing    15  nmod:poss   15:nmod:poss    _
15  myötäisesti myötäisesti ADV _   Derivation=Sti  13  advmod  13:advmod   SpaceAfter=No
16  .   .   PUNCT   _   _   2   punct   2:punct _
```

2) If we consider the cluster formed by Indonesian, Korean, and Thai, it is possible to notice that there are 17 features for which these three languages have higher values than all the other PUD languages: 15 “exclude” and 2 “precede”. The opposite is not observed (i.e.: features whose values for these 3 languages are lower than the ones of the other PUD languages). From the 15 identified “exclude” features, in 2 cases, the value is positive for Indonesian, Korean, and Thai, and 0 for all the rest of the PUD language-set: “VERB-+_exclude_ADJ-advmod_NOUN-ccomp” and “VERB-+_exclude_ADJ-advmod_ADJ-ccomp”. The two linear features with higher values for these 3 languages are: “NOUN-+_precede_PROPN-nsubj_ADJ-acl:relcl” and

"NOUN-+_precede_ADV-advmod_ADJ-acl:relcl" but the frequency values are quite low (i.e.: below 0.008).

3) Regarding the cluster formed by Japanese and Arabic, for 148 features, these two languages have higher values than the rest of the PUD collection: 126 "exclude", 15 "precede", 4 "require", and 3 "unicity". Of the 148 patterns, 82 have positive values for these two languages and 0 for the rest of PUD (11 out of the 15 "precede" features). From these 11 features, the most frequent one in the Japanese corpus is "VERB-+_precede_NOUN-dislocated_NOUN-nsubj". An example of a sentence presenting this feature is displayed in the Figure 4.17. For 3 patterns, Japanese and Arabic present lower values than the rest of the considered languages (2 "exclude" and 1 "precede"). For the "precede" pattern, the associated value is 0 for both languages (i.e.: pattern present in all the other PUD corpora but not in the Japanese and Arabic ones): "NOUN-+_precede_ADJ-amod_NOUN-conj".

Figure 4.17. Example of sentence from the Japanese PUD corpus with the pattern "VERB-+_precede_NOUN-dislocated_NOUN-nsubj" where the token "電話" ("id" = 4, UPOS = "NOUN", DEPREL = "dislocated") precedes the token "機能" ("id" =15, UPOS= "NOUN", DEPREL = "nsubj") in a sub-tree ruled by the VERB "ある" ("id" = 17).



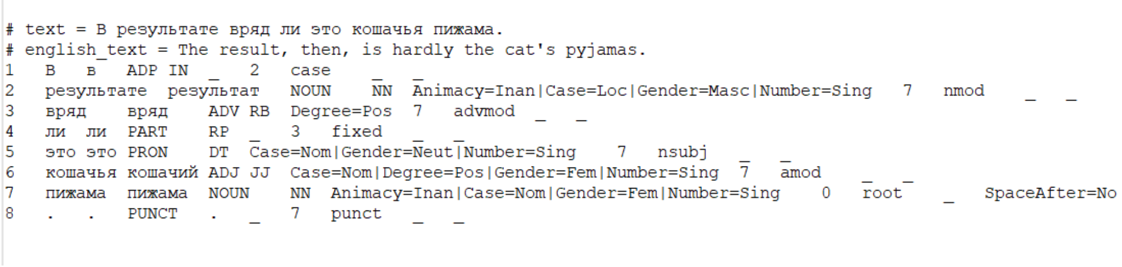
```
# text = 当社の携帯電話は最近の電話よりもはるかに多くの機能がある。
# text_en = Our cellphones are so much more than phones these days.
1	当社	当社	NOUN	名詞-普通名詞-一般	_	4	nmod	_	BunsetuPositionType=SEM_HEAD|LUWBILabel=B|LUWPOS=名詞-普通名詞-一般|SpaceAfter=No|UniDicLemma=当社
2	の	の	ADP	助詞-格助詞	_	1	case	_	BunsetuPositionType=SYN_HEAD|LUWBILabel=B|LUWPOS=助詞-格助詞|SpaceAfter=No|UniDicLemma=の
3	携帯	携帯	NOUN	名詞-普通名詞-サ変可能	_	4	compound	_	BunsetuPositionType=CONT|LUWBILabel=B|LUWPOS=名詞-普通名詞-一般|SpaceAfter=No|UniDicLemma=携帯
4	電話	電話	NOUN	名詞-普通名詞-サ変可能	_	17	dislocated	_	BunsetuPositionType=SEM_HEAD|LUWBILabel=I|LUWPOS=名詞-普通名詞-一般|SpaceAfter=No|UniDicLemma=電話
5	は	は	ADP	助詞-係助詞	_	4	case	_	BunsetuPositionType=SYN_HEAD|LUWBILabel=B|LUWPOS=助詞-係助詞|SpaceAfter=No|UniDicLemma=は
6	最近	最近	NOUN	名詞-普通名詞-副詞可能	_	8	nmod	_	BunsetuPositionType=SEM_HEAD|LUWBILabel=B|LUWPOS=名詞-普通名詞-一般|SpaceAfter=No|UniDicLemma=最近
7	の	の	ADP	助詞-格助詞	_	6	case	_	BunsetuPositionType=SYN_HEAD|LUWBILabel=B|LUWPOS=助詞-格助詞|SpaceAfter=No|UniDicLemma=の
8	電話	電話	NOUN	名詞-普通名詞-サ変可能	_	17	obl	_	BunsetuPositionType=SEM_HEAD|LUWBILabel=B|LUWPOS=名詞-普通名詞-一般|SpaceAfter=No|UniDicLemma=電話
9	より	より	ADP	助詞-格助詞	_	8	case	_	BunsetuPositionType=FUNC|LUWBILabel=B|LUWPOS=助詞-格助詞|SpaceAfter=No|UniDicLemma=より
10	も	も	ADP	助詞-係助詞	_	8	case	_	BunsetuPositionType=SYN_HEAD|LUWBILabel=B|LUWPOS=助詞-係助詞|SpaceAfter=No|UniDicLemma=も
11	はるか	はるか	ADJ	形状詞-一般	_	13	acl	_	BunsetuPositionType=SEM_HEAD|LUWBILabel=B|LUWPOS=形状詞-一般|SpaceAfter=No|UniDicLemma=遥か
12	に	だ	AUX	助動詞	_	11	aux	_	BunsetuPositionType=SYN_HEAD|LUWBILabel=B|LUWPOS=助動詞|SpaceAfter=No|UniDicLemma=だ
13	多く	多く	NOUN	名詞-普通名詞-副詞可能	_	15	nmod	_	BunsetuPositionType=SEM_HEAD|LUWBILabel=B|LUWPOS=名詞-普通名詞-一般|SpaceAfter=No|UniDicLemma=多く
14	の	の	ADP	助詞-格助詞	_	13	case	_	BunsetuPositionType=SYN_HEAD|LUWBILabel=B|LUWPOS=助詞-格助詞|SpaceAfter=No|UniDicLemma=の
15	機能	機能	NOUN	名詞-普通名詞-サ変可能	_	17	nsubj	_	BunsetuPositionType=SEM_HEAD|LUWBILabel=B|LUWPOS=名詞-普通名詞-一般|SpaceAfter=No|UniDicLemma=機能
16	が	が	ADP	助詞-格助詞	_	15	case	_	BunsetuPositionType=SYN_HEAD|LUWBILabel=B|LUWPOS=助詞-格助詞|SpaceAfter=No|UniDicLemma=が
17	ある	ある	VERB	動詞-非自立可能	_	0	root	_	BunsetuPositionType=ROOT|LUWBILabel=B|LUWPOS=動詞-一般|SpaceAfter=No|UniDicLemma=有る
18	。	。	PUNCT	補助記号-句点	_	17	punct	_	BunsetuPositionType=CONT|LUWBILabel=B|LUWPOS=補助記号-句点|SpaceAfter=No|UniDicLemma=。
```

4) In the dendrogam (Figure 4.14), Russian and Icelandic were clustered close to each other. The analysis of the MarsaGram features shows that in 396 cases, the values corresponding to these 2 languages are higher than for the rest of the language-set: 358 "exclude", 32 "precede", 4 "unicity", and 2 "require". The "require" patterns are "VERB-+_require_AUX-advcl_PUNCT-punct" and "PROPN-+_require_PROPN-parataxis_ADP-case", while the "unicity" patterns are "ADJ-+_unicity_ADV-obl",

"VERB-+_unicity_AUX-advcl", "PROPN-+_unicity_NUM-acl:relcl", and "VERB-+_unicity_SCONJ-obj".

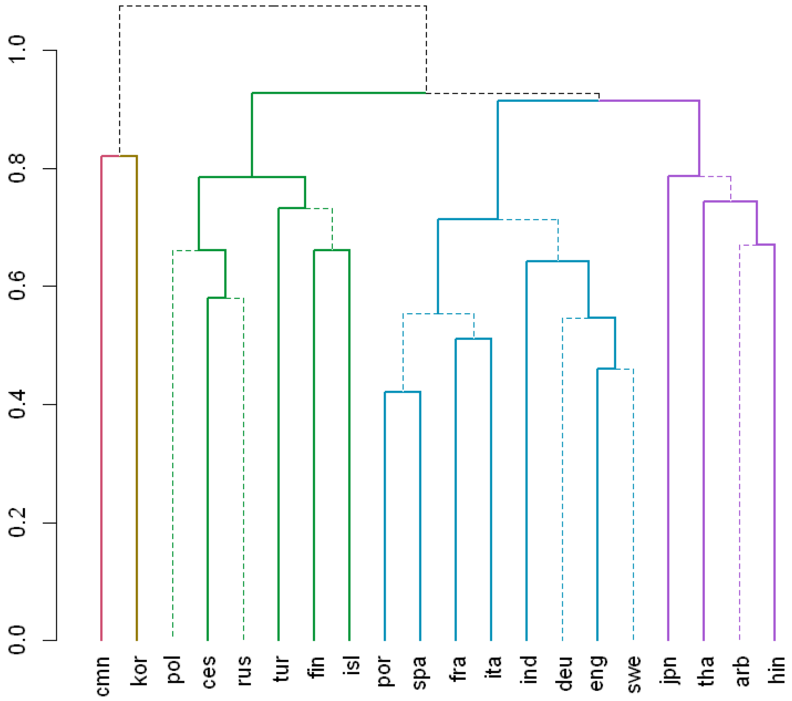
In total, for 256 patterns, Russian and Icelandic have positive values while the other PUD languages have 0 (22 of the "precede" type). These linear features involve mostly nouns and verbs as the heads of the sub-trees. Of these 22 features, the most frequent linear pattern is "ADV-+_precede_*_PART-fixed" (0.0115 for Icelandic and 0.0375 for Russian). An example of this phenomenon is presented in Figure 4.18.

Figure 4.17. Example of sentence from the Russian PUD corpus with the pattern "ADV-+_precede_*_PART-fixed" where the token "вряд" ("id" = 3, UPOS = "ADV") precedes the token "ли" ("id" =4, UPOS= "PART", DEPREL = "fixed") in a sub-tree ruled by the token "вряд" ("id" = 3, UPOS = "ADV").



```
# text = В результате вряд ли это кошачья пижама.
# english_text = The result, then, is hardly the cat's pyjamas.
1   В   в   ADP IN  _   2   case    _   _
2   результате  результат   NOUN    NN  Animacy=Inan|Case=Loc|Gender=Masc|Number=Sing   7   nmod    _   _
3   вряд    вряд    ADV RB  Degree=Pos  7   advmod  _   _
4   ли  ли  PART    RP  _   3   fixed   _   _
5   это это PRON    DT  Case=Nom|Gender=Neut|Number=Sing    7   nsubj   _   _
6   кошачья кошачий ADJ JJ  Case=Nom|Degree=Pos|Gender=Fem|Number=Sing  7   amod    _   _
7   пижама  пижама  NOUN    NN  Animacy=Inan|Case=Nom|Gender=Fem|Number=Sing    0   root    _   SpaceAfter=No
8   .   .   PUNCT   .   _   7   punct   _   _
```

Figure 4.18. Cluster dendrogram obtained from the cosine dissimilarity matrix calculated with the comparison of the PUD MarsaGram language vectors (all patterns).



The analysis of Figure 4.18 indicates that when cosine distances are considered, Chinese and Korean appear separated from the other PUD languages (while they were part of a same large cluster when Euclidean distances were considered).

Romance languages are clustered together and form a larger group together with Germanic languages (except for Icelandic) and Indonesian. The latter was also grouped with Romance languages in the dendrograms obtained from lang2vec syntactic vectors and is considered as the same language-type as them by Hawkins (1983).

Concerning the 3 Slavic PUD languages, they are clustered together forming a sub-group of a larger cluster which also contains the sub-group encompassing Turkish, Finnish, and Icelandic. The first two languages were not classified in the same group in the lang2vec syntactic dendrograms, thus, MarsaGram seems to recognize better the similarities which are shared by these two languages. However, in the cosine MarsaGram dendrogram, the position of Icelandic does not correspond at all with its genealogical characteristics.

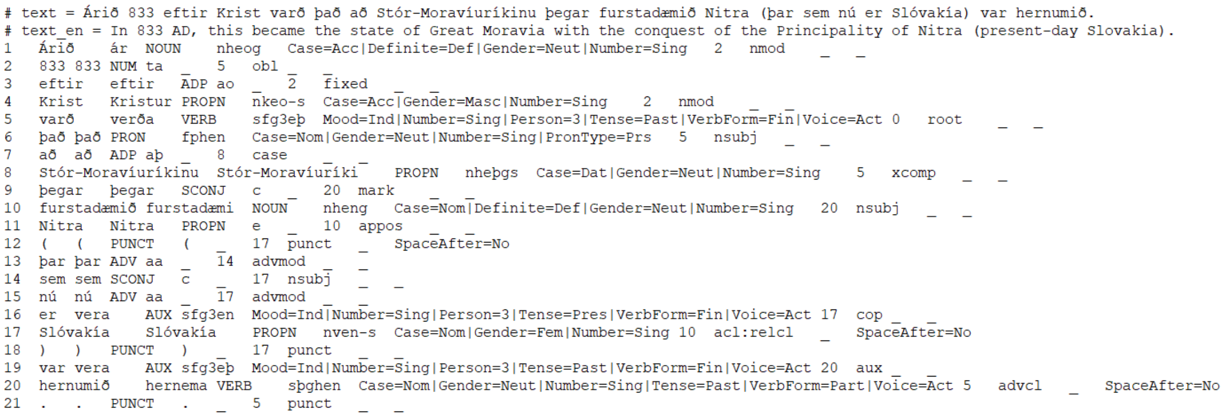
Finally, on the right side of Figure 4.18, there is a cluster composed of Japanese, Thai, Arabic, and Hindi. These languages are also isolated from other PUD languages; however, their dissimilarities values are a bit lower than the ones obtained with Chinese and Korean vectors, thus, they are clustered separately.

The analysis of the MarsaGram features shows that:

1) For the cluster formed of Indonesian, English, German, and Swedish, the values of these four languages are higher when compared to the other PUD ones for 34 patterns (31 “exclude” and 3 “precede”). There is no pattern whose values for these languages are lower than the other elements in our language-set. The 3 “precede” features occur only in the four abovementioned languages (i.e.: frequency equal to 0 to the other languages): “VERB-+_precede_PROPN-nsubj:pass_VERB-conj”, “VERB-+_precede_PRON-nsubj:pass_PROPN-obl”, and “VERB-+_precede_SCONJ-mark_PRON-nsubj:pass”. The first 2 have a frequency of 0.04 in the Indonesian corpus, and the third one, 0.02.
2) For the cluster composed of Icelandic and Finnish, these two languages have 601 features with higher values than the other PUD languages (511 “exclude”, 66 “precede”, 12 “unicity”, and 12 “require”). Only one feature is identified corresponding to lower values for these two languages in comparison to the others: “PROPN-+_unicity_DET-det”. A total of 228 patterns are exclusive of Icelandic and Finnish (191

"exclude", 3 "unicity", 8 "require", and 26 "precede"). Regarding the "precede" features, most of them concern sub-trees rules by nouns and adjectives. The most frequent linear pattern exclusive of these two languages is "SCONJ-+_precede_ADV-advmod_*" (0.833 for Finnish and 0.375 for Icelandic). The figure 4.19 presents an example of sentence in Icelandic with this pattern.

Figure 4.19. Example of sentence from the Icelandic PUD corpus with the pattern "SCONJ-+_precede_ADV-advmod_*" where the token "þar" ("id" = 13, UPOS = "ADV", DEPREL = "advmod") precedes the token "sem" ("id" =14, UPOS= "SCONJ") which is the head of the sub-tree.



```
# text = Árið 833 eftir Krist varð það að Stór-Moravíuríkinu þegar furstadæmið Nitra (þar sem nú er Slóvakía) var hernumið.
# text_en = In 833 AD, this became the state of Great Moravia with the conquest of the Principality of Nitra (present-day Slovakia).
1   Árið    ár  NOUN    nheog   Case=Acc|Definite=Def|Gender=Neut|Number=Sing   2   nmod    _   _
2   833 833 NUM ta  _   5   obl _   _
3   eftir   eftir   ADP ao  _   2   fixed   _   _
4   Krist   Kristur PROPN   nkeo-s  Case=Acc|Gender=Masc|Number=Sing    2   nmod    _   _
5   varð    verða   VERB    sfg3eþ  Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act 0   root    _   _
6   það það PRON    fphen   Case=Nom|Gender=Neut|Number=Sing|PronType=Prs   5   nsubj   _   _
7   að  að  ADP aþ  _   8   case    _   _
8   Stór-Moravíuríkinu  Stór-Moravíuríki    PROPN   nheþgs  Case=Dat|Gender=Neut|Number=Sing    5   xcomp   _   _
9   þegar   þegar   SCONJ   c   _   20  mark    _   _
10  furstadæmið furstadæmi  NOUN    nheng   Case=Nom|Definite=Def|Gender=Neut|Number=Sing   20  nsubj   _   _
11  Nitra   Nitra   PROPN   e   _   10  appos   _   _
12  (   (   PUNCT   (   _   17  punct   _   SpaceAfter=No
13  þar þar ADV aa  _   14  advmod  _   _
14  sem sem SCONJ   c   _   17  nsubj   _   _
15  nú  nú  ADV aa  _   17  advmod  _   _
16  er  vera    AUX sfg3en  Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act 17  cop _   _
17  Slóvakía    Slóvakía    PROPN   nven-s  Case=Nom|Gender=Fem|Number=Sing 10  acl:relcl   _   SpaceAfter=No
18  )   )   PUNCT   )   _   17  punct   _   _
19  var vera    AUX sfg3eþ  Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act 20  aux _   _
20  hernumið    hernema VERB    sþghen  Case=Nom|Gender=Neut|Number=Sing|Tense=Past|VerbForm=Part|Voice=Act 5   advcl   _   SpaceAfter=No
21  .   .   PUNCT   .   _   5   punct   _   _
```

3) Considering the isolated cluster on the left side of the dendrogram, Chinese and Korean have 747 features in common with values higher than the ones of the other PUD languages. Again, the "exclude" property is the most frequent one (603 cases), followed by "precede" (110), unicity (29), and "require" (5). Of these 747 patterns, 283 are only present in the corpora of these two languages.

   On the other hand, for 10 features, these two languages present values inferior than the rest of the PUD languages. Out of these 10 patterns, 3 "precede" ones are present in all the other corpora but not in the Chinese and Korean ones (i.e.: frequency equal to 0): "ADJ-+_precede_CCONJ-cc_ADV-advmod", "NOUN-+_precede_CCONJ-cc_ADP-case", and "VERB-+_precede_CCONJ-cc_NOUN-obl".

4) For the purple cluseter composed of Japanese, Thai, Arabic, and Hindi, only two patterns can be identified when features are analysed: "ADJ-+_exclude_ADP-case_PRON-obl" (more frequent in these 4 languages) and "VERB-+_exclude_ADV-advmod_NOUN-advcl" (less frequent in the listed languages in comparison to the other PUD ones).

When comparing both dendrograms obtained with MarsaGram data (all properties), it is possible to notice that they present some similarities for languages which are close in terms of phylogenetic features. Portuguese is grouped with Spanish, French with Italian, Czech with Polish, and English with Swedish in both dendrograms. However, Russian and Icelandic present a discrepancy in terms of classification when the different distance metrics are considered. In the Euclidean dendrogram, Icelandic is closer to the other Germanic languages but Russian is positioned in the same cluster, on the other hand, in the cosine dendrogram, Russian is part of the Slavic cluster, but Icelandic is grouped with Finnish and Turkish.

It is possible to notice via the analysis of the features that, in general, "exclude" patterns are more decisive when the distances are calculated. This is an expected result as the number of the features concerning this property is much higher than the other ones. Moreover, we could observe that in most cases, the number of identified patterns for which the languages in a cluster have higher values when compared to the other PUD languages is relatively higher than the cases where the pattern is less frequent.

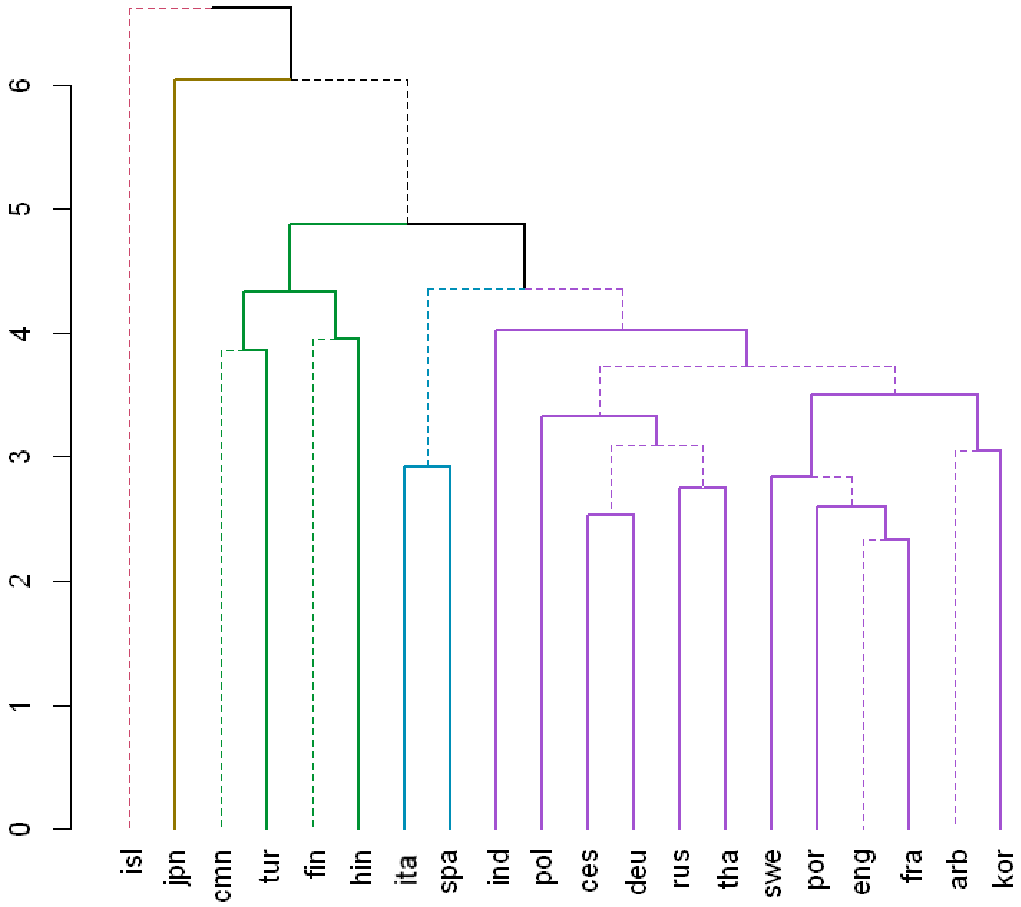
Moreover, the above MarsaGram typological classifications do not group languages with the same syntactic characteristics regarding specific word ordering occurrences such as the position of S, V, and O, or the adposition location in relation to the noun (important features in Hawkins' analysis). Even though the MarsaGram tool extracts this type of information, these phenomena become less relevant when the clusters are defined, due to the size of the language vectors.

b) Only linear patterns:

The dissimilarity matrices generated with MarsaGram language vectors (only linear patterns) regarding Euclidean and cosine distances are presented in Annexes 18 and 19 respectively. Moreover, the dendrograms produced via the clustering analysis of these matrices are displayed in Figures 4.20 and 4.21.

The result obtained from the Euclidean dissimilarity matrix regarding the comparison of the language vectors formed with only linear patterns 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 all together in the large purple group in Figure 4.20. 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 and MarsaGram all properties dendrograms, are not in the same sub-cluster when only linear patterns are considered).

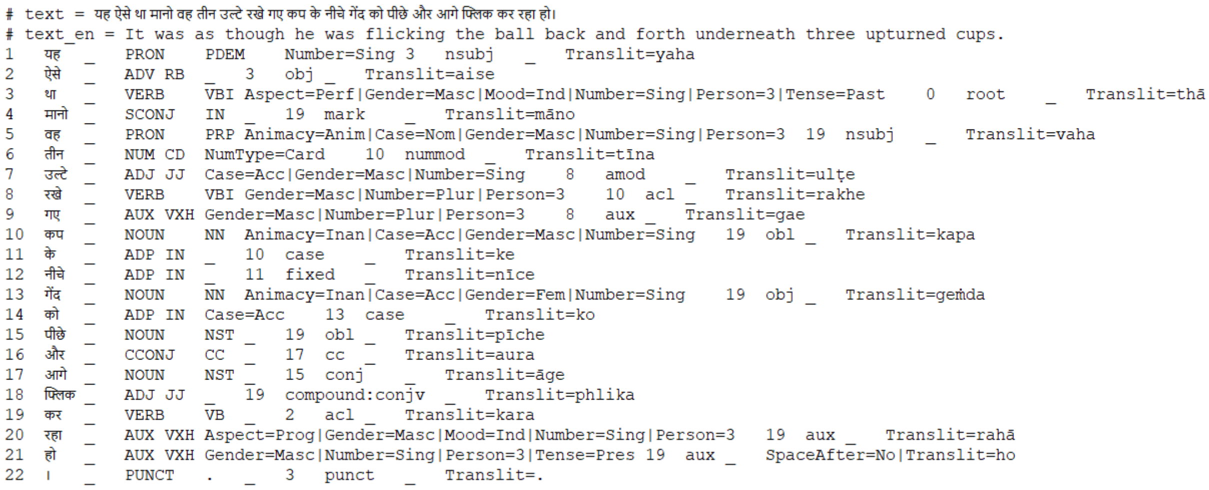
Figure 4.20. Cluster dendrogram obtained from the Euclidean dissimilarity matrix calculated with the comparison of the PUD MarsaGram language vectors (only linear patterns).



By analysing in details the MarsaGram linear features, it is possible to observe that:

1) For the green cluster composed of Chinese, Turkish, Finnish, and Hindi, there is no feature for which the corresponding values of these languages is higher or lower than all the other PUD languages. However, when Chinese and Turkish are analysed separately, 100 linear patterns are identified (16 of them being exclusive of these two languages). On the other hand, when we consider Finnish and Hindi, for 23 patterns these two languages have higher values (9 of them being exclusive, for example "VERB-+_precede_PRON-nsubj_ADV-obj"). An example of this pattern is presented in the Figure 4.21.

Figure 4.21. Example of sentence from the Hindi PUD corpus with the pattern “VERB-+_precede_PRON-nsubj_ADV-obj” where the token “यह” (“id” = 1, UPOS = “PRON”, DEPREL = “nsubj”) precedes the token “ऐसे” (“id” = 2, UPOS= “ADV”, DEPREL = “obj”) with the token “था” (“id” = 3, UPOS = “VERB”) as the head of the sub-tree.

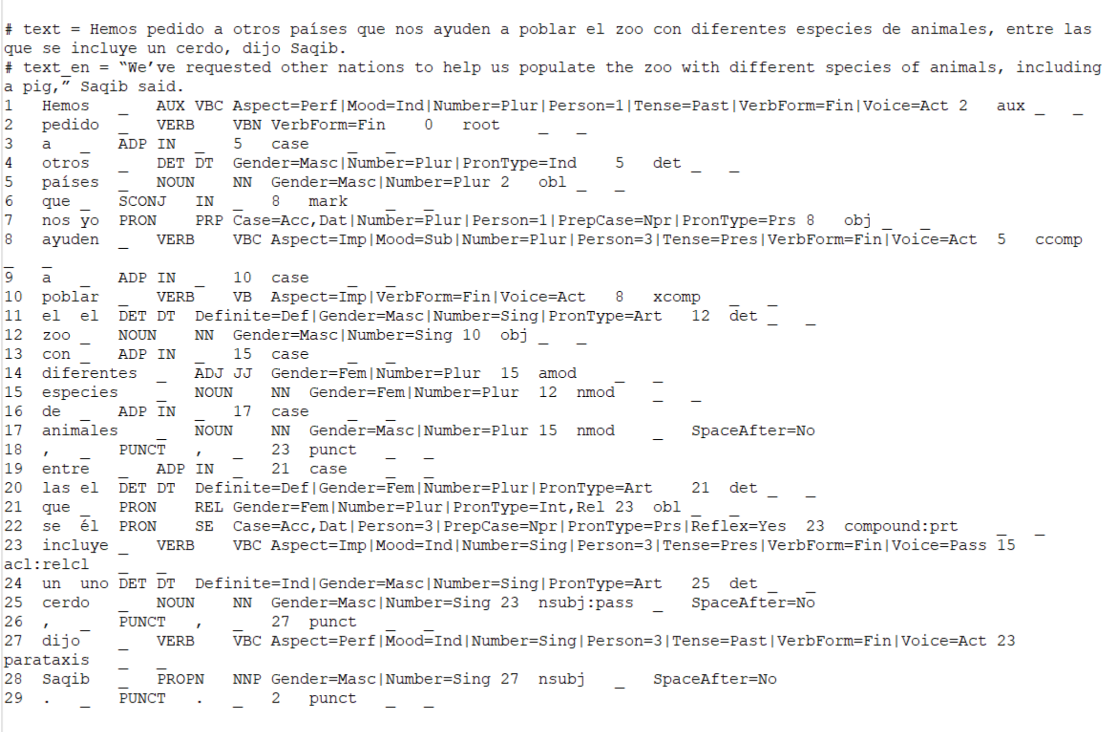


```
# text  =  यह ऐसे था मानो वह तीन उल्टे रखे गए कप के नीचे गेंद को पीछे और आगे फ्लिक कर रहा हो।
# text_en = It was as though he was flicking the ball back and forth underneath three upturned cups.
1    यह    _    PRON    PDEM    Number=Sing 3   nsubj    _    Translit=yaha
2    ऐसे    _    ADV RB  _    3   obj _   Translit=aise
3    था    _    VERB    VBI Aspect=Perf|Gender=Masc|Mood=Ind|Number=Sing|Person=3|Tense=Past    0   root    _    Translit=thā
4    मानो   _    SCONJ   IN  _   19  mark    _    Translit=māno
5    वह    _    PRON    PRP Animacy=Anim|Case=Nom|Gender=Masc|Number=Sing|Person=3  19  nsubj   _    Translit=vaha
6    तीन   _    NUM CD  NumType=Card    10  nummod  _    Translit=tīna
7    उल्टे   _    ADJ JJ  Case=Acc|Gender=Masc|Number=Sing    8   amod    _    Translit=ulṭe
8    रखे   _    VERB    VBI Gender=Masc|Number=Plur|Person=3    10  acl _    Translit=rakhe
9    गए    _    AUX VXH Gender=Masc|Number=Plur|Person=3    8   aux _    Translit=gae
10   कप    _    NOUN    NN  Animacy=Inan|Case=Acc|Gender=Masc|Number=Sing   19  obl _    Translit=kapa
11   के    _    ADP IN  _   10  case    _    Translit=ke
12   नीचे   _    ADP IN  _   11  fixed   _    Translit=nīce
13   गेंद   _    NOUN    NN  Animacy=Inan|Case=Acc|Gender=Fem|Number=Sing    19  obj _    Translit=gem̐da
14   को    _    ADP IN  Case=Acc    13  case    _    Translit=ko
15   पीछे   _    NOUN    NST _   19  obl _    Translit=pīche
16   और    _    CCONJ   CC  _   17  cc  _    Translit=aura
17   आगे   _    NOUN    NST _   15  conj    _    Translit=āge
18   फ्लिक  _    ADJ JJ  _   19  compound:conjv  _    Translit=phlika
19   कर    _    VERB    VB  _   2   acl _    Translit=kara
20   रहा   _    AUX VXH Aspect=Prog|Gender=Masc|Mood=Ind|Number=Sing|Person=3   19  aux _    Translit=rahā
21   हो    _    AUX VXH Gender=Masc|Number=Sing|Person=3|Tense=Pres 19  aux _    SpaceAfter=No|Translit=ho
22   ।     _    PUNCT   .   _   3   punct   _    Translit=.
```

2) Regarding the blue cluster formed by Italian and Spanish, these 2 languages have 33 features with higher frequencies than the other PUD languages, 14 of them being exclusive (e.g.: “VERB-+_precede_DET-obl_VERB-advcl”, “NOUN-+_precede_DET-det_VERB-ccomp”, “SYM-+_precede_DET-det_PUNCT-punct”). No pattern was identified where the correspondent values for Spanish and Italian are lower than the rest of PUD languages.

   An example of the pattern “NOUN-+_precede_DET-det_VERB-ccomp” from the Spanish PUD corpus is displayed in the Figure 4.22.

Figure 4.22. Example of sentence from the Spanish PUD corpus with the pattern "NOUN-+_precede_DET-det_VERB-ccomp" where the token "otros" ("id" = 4, UPOS = "DET", DEPREL = "det") precedes the token "ayuden" ("id" = 8, UPOS= "VERB", DEPREL = "ccomp") with the token "países" ("id" = 5, UPOS = "NOUN") as the head of the sub-tree.



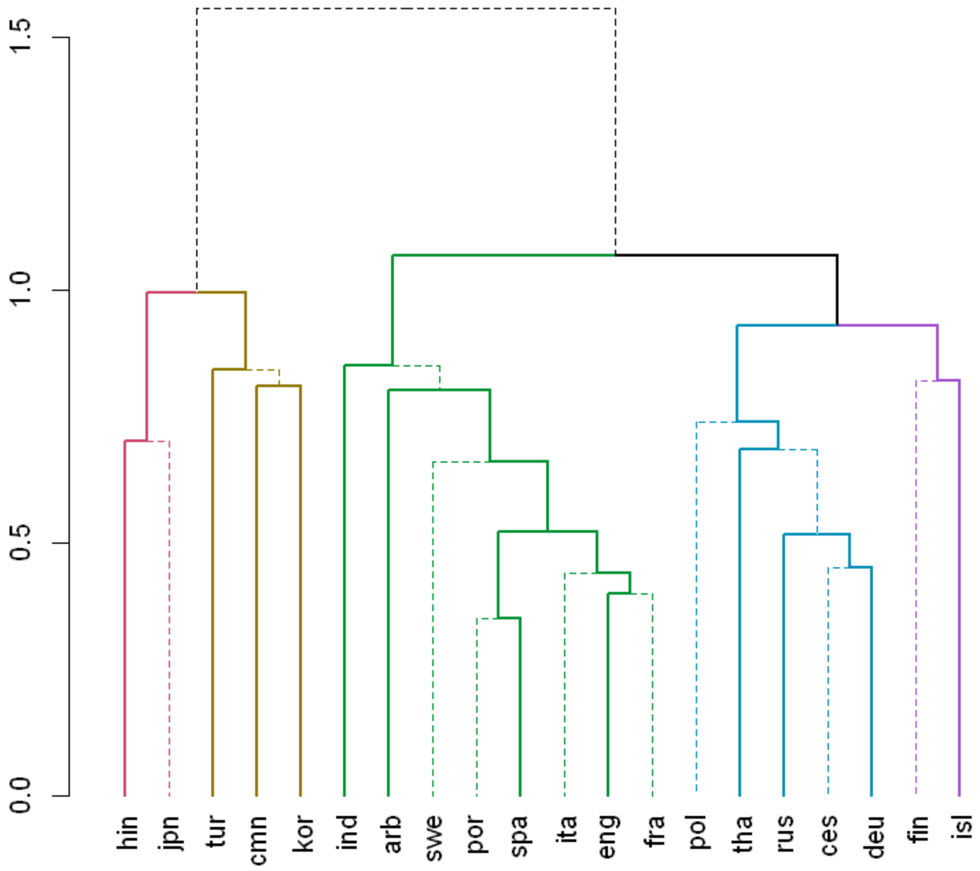
```
# text = Hemos pedido a otros países que nos ayuden a poblar el zoo con diferentes especies de animales, entre las
que se incluye un cerdo, dijo Saqib.
# text_en = "We've requested other nations to help us populate the zoo with different species of animals, including
a pig," Saqib said.
1   Hemos   _   AUX VBC Aspect=Perf|Mood=Ind|Number=Plur|Person=1|Tense=Past|VerbForm=Fin|Voice=Act 2   aux _   _
2   pedido  _   VERB    VBN VerbForm=Fin    0   root    _   _
3   a   _   ADP IN  _   5   case    _   _
4   otros   _   DET DT  Gender=Masc|Number=Plur|PronType=Ind    5   det _   _
5   países  _   NOUN    NN  Gender=Masc|Number=Plur 2   obl _   _
6   que _   SCONJ   IN  _   8   mark    _   _
7   nos yo  PRON    PRP Case=Acc,Dat|Number=Plur|Person=1|PrepCase=Npr|PronType=Prs 8   obj _   _
8   ayuden  _   VERB    VBC Aspect=Imp|Mood=Sub|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act  5   ccomp
_   _
9   a   _   ADP IN  _   10  case    _   _
10  poblar  _   VERB    VB  Aspect=Imp|VerbForm=Fin|Voice=Act   8   xcomp   _   _
11  el  el  DET DT  Definite=Def|Gender=Masc|Number=Sing|PronType=Art   12  det _   _
12  zoo _   NOUN    NN  Gender=Masc|Number=Sing 10  obj _   _
13  con _   ADP IN  _   15  case    _   _
14  diferentes  _   ADJ JJ  Gender=Fem|Number=Plur  15  amod    _   _
15  especies    _   NOUN    NN  Gender=Fem|Number=Plur  12  nmod    _   _
16  de  _   ADP IN  _   17  case    _   _
17  animales    _   NOUN    NN  Gender=Masc|Number=Plur 15  nmod    _   SpaceAfter=No
18  ,   _   PUNCT   ,   _   23  punct   _   _
19  entre   _   ADP IN  _   21  case    _   _
20  las el  DET DT  Definite=Def|Gender=Fem|Number=Plur|PronType=Art    21  det _   _
21  que _   PRON    REL Gender=Fem|Number=Plur|PronType=Int,Rel 23  obl _   _
22  se  él  PRON    SE  Case=Acc,Dat|Person=3|PrepCase=Npr|PronType=Prs|Reflex=Yes  23  compound:prt    _   _
23  incluye _   VERB    VBC Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Pass 15
acl:relcl   _   _
24  un  uno DET DT  Definite=Ind|Gender=Masc|Number=Sing|PronType=Art   25  det _   _
25  cerdo   _   NOUN    NN  Gender=Masc|Number=Sing 23  nsubj:pass  _   SpaceAfter=No
26  ,   _   PUNCT   ,   _   27  punct   _   _
27  dijo    _   VERB    VBC Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act 23
parataxis   _   _
28  Saqib   _   PROPN   NNP Gender=Masc|Number=Sing 27  nsubj   _   SpaceAfter=No
29  .   _   PUNCT   .   _   2   punct   _   _
```

3) For the big purple cluster:
    a. Czech and German: 29 features with higher frequency for these 2 languages (9 exclusive).
    b. Russian and Thai: 27 features with higher frequency for these 2 languages (16 exclusive). One feature is less frequent for these two languages when compared to the other PUD ones ("VERB-+_precede_PROPN-nsubj_NOUN-obj")
    c. Swedish, Portuguese, English, and French: only two patterns were identified which are more frequent in these four languages ("NOUN-+_precede_AUX-cop_DET-det" and "NOUN-+_precede_PRON-nsubj_DET-det").
    d. Arabic and Korean: 17 features with higher frequency for these 2 languages (10 exclusive).

The dendrogram obtained from the cosine dissimilarity matrix of MarsaGram considering only linear vectors (Figure 4.23) seems to be more coherent with the genealogical language classification: Romance languages are grouped in the same cluster, and Spanish and Portuguese form, again, a specific sub-group. However, Germanic and Slavic languages are split into different sub-groups: English and Swedish are part of the Romance languages sub-cluster, Icelandic forms a whole sub-group with Finnish, and German is positioned close to Czech, in a cluster together with other Slavic languages and Thai.

Hindi, Japanese, Turkish, Chinese, and Korean form a cluster that corresponds to the one obtained in lang2vec dendrograms. All these languages, with exception of Chinese, share specific syntactic features such as SOV and postpositions, and are classified as type 23 by Hawkins (1983). Moreover, Arabic and Indonesian are closer to Romance languages in this MarsaGram dendrogram, which was also the case in lang2vec clustering representations (also, Indonesian is classified as type 9 by Hawkins, 1983, same as other Romance languages in PUD).

Figure 4.23. Cluster dendrogram obtained from the cosine dissimilarity matrix calculated with the comparison of the PUD MarsaGram language vectors (only linear patterns).



Concerning PUD VO languages, Fnnish and Icelandic belong to a larger group in lang2vec dendrograms while this MarsaGram dendrogram analysis clusters them as one single group. Moreover, Thai is classified by Hawkins (1983) as the same type as Romance languages and

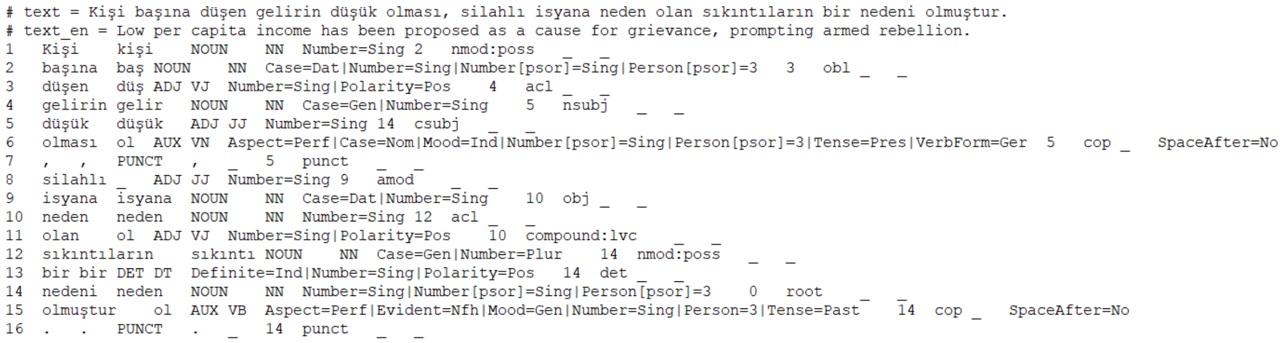
Indonesian (type 9), but in this dendrogram, it is placed together with PUD Slavic languages (type 10).

By analysing the distribution of the MarsaGram linear features, we observe that:

1) For the cluster composed of Hindi and Japanese, these two languages have 95 common features which are more frequent in their corpora in comparison to the other PUD languages. Of them, 53 are exclusive of these two languages. On the other hand, we identified 3 patterns whose frequency is lower in comparison to the other languages of the language-set (i.e.: “NOUN-+_precede_*_PROPN-appos”, “NOUN-+_precede_*_VERB-conj”, and “VERB-+_precede_*_NOUN-conj”).
2) For Turkish, Chinese, and Korean, the number of common features with higher frequency is 21 and correspond to patterns where the sub-tree is ruled either by a noun or a verb. Of these patterns, 4 are present only in the corpora of these 3 languages: “VERB-+_precede_NOUN-csubj_NOUN-obj”, “NOUN-+_precede_ADJ-csubj_*”, “NOUN-+_precede_ADJ-csubj_PUNCT-punct”, and “NOUN-+_precede_ADJ-csubj_AUX-cop”. An example of the last pattern (from the Turkish corpus) is displayed in Figure 4.24.

   One feature is present in these 3 languages but with a frequency lower than what is observed in the other PUD languages: “NOUN-+_precede_CCONJ-cc_ADP-case”.

Figure 4.24. Example of sentence from the Turkish PUD corpus with the pattern “NOUN-+_precede_ADJ-csubj_AUX-cop” where the token “düşük” (“id” = 5, UPOS = “ADJ”, DEPREL = “csubj”) precedes the token “olması” (“id” = 6, UPOS= “AUX”, DEPREL = “cop”) with the token “nedeni” (“id” = 14, UPOS = “NOUN”) as the head of the sub-tree.



```
# text = Kişi başına düşen gelirin düşük olması, silahlı isyana neden olan sıkıntıların bir nedeni olmuştur.
# text_en = Low per capita income has been proposed as a cause for grievance, prompting armed rebellion.
1   Kişi    kişi    NOUN    NN  Number=Sing 2   nmod:poss   _   _
2   başına  baş NOUN    NN  Case=Dat|Number=Sing|Number[psor]=Sing|Person[psor]=3   3   obl _   _
3   düşen   düş ADJ VJ  Number=Sing|Polarity=Pos    4   acl _   _
4   gelirin gelir   NOUN    NN  Case=Gen|Number=Sing    5   nsubj   _   _
5   düşük   düşük   ADJ JJ  Number=Sing 14  csubj   _   _
6   olması  ol  AUX VN  Aspect=Perf|Case=Nom|Mood=Ind|Number[psor]=Sing|Person[psor]=3|Tense=Pres|VerbForm=Ger  5   cop _   SpaceAfter=No
7   ,   ,   PUNCT   ,   _   5   punct   _   _
8   silahlı _   ADJ JJ  Number=Sing 9   amod    _   _
9   isyana  isyana  NOUN    NN  Case=Dat|Number=Sing    10  obj _   _
10  neden   neden   NOUN    NN  Number=Sing 12  acl _   _
11  olan    ol  ADJ VJ  Number=Sing|Polarity=Pos    10  compound:lvc    _   _
12  sıkıntıların    sıkıntı NOUN    NN  Case=Gen|Number=Plur    14  nmod:poss   _   _
13  bir bir DET DT  Definite=Ind|Number=Sing|Polarity=Pos   14  det _   _
14  nedeni  neden   NOUN    NN  Number=Sing|Number[psor]=Sing|Person[psor]=3    0   root    _   _
15  olmuştur    ol  AUX VB  Aspect=Perf|Evident=Nfh|Mood=Gen|Number=Sing|Person=3|Tense=Past    14  cop _   SpaceAfter=No
16  .   .   PUNCT   .   _   14  punct   _   _
```

3) For the sub-cluster composed of English and French, these two languages have 30 patterns in common with higher frequency than the rest of the PUD language-set (18 exclusive features).
4) For Thai and Russian, 17 exclusive features were identified (27 in total where the frequency is higher for these two languages). For only one pattern the frequency is lower for Russian and Thai when compared to the other PUD languages ("VERB-+_precede_PROPN-nsubj_NOUN-obj").
5) For the purple sub-cluster, Finnish and Icelandic have 66 common features with higher frequency than the other PUD languages (26 exclusive).

It is possible to notice that in the larger clusters, languages usually have in common around 20 to 30 features with higher frequencies. On the other hand, the languages in the extreme clusters (pink and purple) have more common patterns which distinguish them from the rest of the PUD language sample.

In conclusion, when using the different data extracted with the MarsaGram tool, as it was observed with the lang2vec analysis, the clustering analysis is sensitive to the type of distance metric (Euclidean or cosine). Also, it was noticeable that when all MarsaGram properties are considered, the dendrograms provide a classification more coherent with the genealogical one for the Indo-European PUD languages. Russian and Icelandic seem to present some specificities, thus being classed in different ways in respect to the distance metrics. The cosine dendrogram built with MarsaGram linear properties, on the other hand, is the only one which presents the ensemble of OV languages in a same isolated cluster (together with Chinese).

4.6 Quantitative Typological Classification Using Head and Dependents Ordering

In the section "Theoretical background and related work review", it has been presented that the concept of heads (or governor) and dependents are part of some typological theories. Vennemman (1973) assumed that the ordering of heads and dependents tend to follow the position of verb and object respectively, and that variations are explained by different evolutionary states of languages. Hawkins (1983), with the "Head and Dependent Theory" (HDT), analyses different pairs of constituents, showing that attested languages can be grouped in types concerning the specific observed ordering of these components.

On the other hand, Dryer (1992) presented an alternative to the HDT, as the definition of many pairs of heads and dependents can be controversial. However, even in his "Branching-Direction Theory" (BDT), Dryer proposes an alternative version considering these elements if they can be well-defined: “Verb patterners are heads and object patterners are fully recursive phrasal dependents”.

The concept of heads governing dependents is one of the main principles of dependency grammars which are the base for the Natural Language Processing task of dependency parsing. Moreover, the UD framework provides a consistent annotation in terms of syntactic relations for all languages, thus, minimizing the bias of divergent definitions of heads and dependents.

Hawkins (1983) provided a classification of languages regarding some word-ordering phenomena according to the most frequent occurrences (basic word order). What is proposed in this thesis is a more complete quantitative corpus-based approach: all head and dependent word order phenomena occurring in PUD corpora are considered and quantified. The adopted strategy consists of identifying all possible orderings of heads and dependents and using them as features to compose language vectors via python scripts. In this way, what is being quantitively analysed here is the head directionality parameter: whether the head precedes the dependent (right-branching) or if it is positioned after it (left-branching) on the surface level (Haider et al., 2015). Moreover, for every single PUD language, the value of each feature corresponds to the frequency of occurrence of the word order phenomenon in the specific corpus. The languages vectors can be schematized as:

$$(4.7)\ vector_{language} \\ = [frequency_{word\_order\_1}, frequency_{word\_order\_2}, \ldots, frequency_{word\_order\_3}]$$

The basic form of each extracted feature is:

$$(4.8)\ POS_{dependent}\_DepRel_{dependent}\_relative\ order\_POS_{head}$$

Thus, for each pattern, the part-of-speech (POS) and the dependency relation (DepRel) of the dependent is specified, together with its relative order observed in the sentence (i.e.: precedes or follows), and the part-of-speech of the respective head. Two examples of features observed in PUD corpora are presented below:

1) ADV_advmod_precedes_ADJ - 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. It is a case of a head-final

or left-branching feature. An attested example of this phenomenon in the English PUD corpus is presented in Figure 4.25.

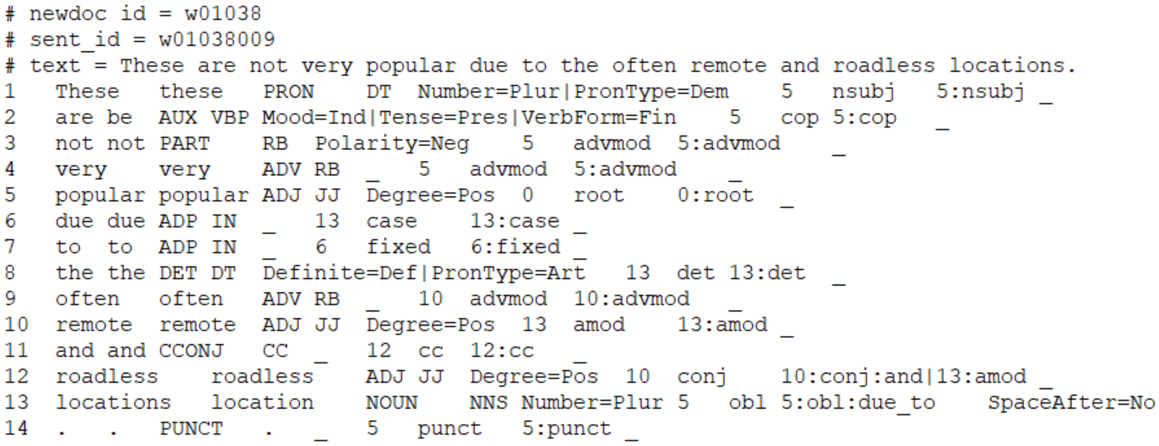
2) NOUN_obl_follows_VERB - In this case, the dependent (NOUN), comes after the head, which is a verb, and has the function of oblique nominal (obl). The dependent can be in any position after the head, not necessarily being right next to it. It is a head-initial or right-branching feature. In Figure 4.26, a sentence from the English PUD corpus containing this specific word order feature is displayed.

The typological analysis concerning heads and dependents position differs from the previous one regarding MarsaGram linear patterns as the word order phenomena extracted with this tool concern tokens which do not form necessarily a head and dependent pair. Thus, the study of features regarding the head directionality parameter focus on the surface order of elements that are directly linked syntactically and quantifies all attested orderings (dependent positioned before or after the head).

In total, 2,890 word order patterns were identified in all PUD corpora. 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. When a specific phenomenon was not observed in a language, the value 0 was attributed to the respective feature within the language vector.

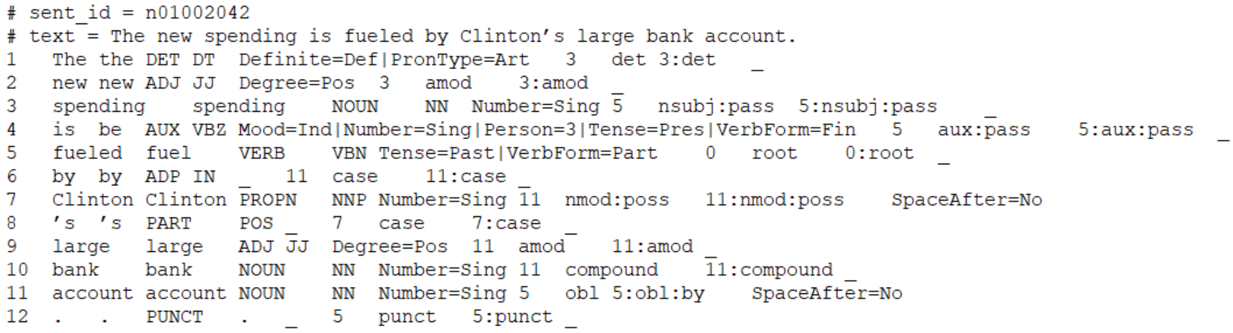
Table 4.25 shows the overall distribution of head directionality features attested in the corpora composing the PUD collection. It is possible to notice that the vast majority of phenomena occur in less than half of the languages.

Figure 4.25. Sentence from the English PUD corpus where the word order feature ADV_advmod_precedes_ADJ is attested. The dependent (ADV) corresponds to token 4 ("very") and the head (ADJ), to token 5 ("popular").



```
# newdoc id = w01038
# sent_id = w01038009
# 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 4.26. Sentence from the English PUD corpus where the word order feature NOUN_obl_follows_VERB is attested. The dependent (NOUN) corresponds to token 11 (“account”) and the head (VERB), to token 5 (“fueled”).



```
# sent_id = n01002042
# 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 _
```

	Number of properties	**%**
Occurring in only one corpus	98	3.39
Occurring in more than 10 corpora	268	9.27
Occurring in all corpora	28	0.97

Table 4.25. Distribution of head and dependent features inside PUD corpora and the respective percentage of the total number of patterns.

The 28 features attested in all corpora are detailed in the Annex 20, and correspond basically to:

- Adverbs with the dependency relation of adverbial modifier preceding the heads;
- Coordinative conjunctions (coordination) preceding the heads;
- Nouns and Proper Nouns as appositional modifiers following the heads;
- Subject (NOUN or PROPN) preceding the verb (head);
- Verb as an adverbial clause modifier preceding another verb (head);
- Punctuation position (preceding of following) possible heads;

As seen in the features concerning lang2vec syntactic vectors, with exception of Arabic, all PUD languages have a value of 1.0 regarding the feature “S_SUBJECT_BEFORE_VERB” which is equivalent to the head directionality features “NOUN_nsubj_precedes_VERB” and “PROPN_nsubj_precedes_VERB”. What is possible to see is that the values of these two features are much lower for Arabic (i.e.: much lower frequency in the corpus) in comparison to other PUD languages. Thus, the described phenomena can also happen in Arabic but this information is not considered in the lang2vec description, showing that quantitative methods like this one is more realistic when describing languages than literature-based ones.

The number of attested features in each PUD corpus is not the same. It varies from 243 for Japanese to 637 for Icelandic. The table 4.26 presents the number of word order patterns concerning heads and dependents for each language.

This distribution of features among PUD languages does not correspond directly to the number of DEPREL tags used to describe each corpus (Table 4.8). For instance, regarding Icelandic, its tag-set is composed of 36 labels, less than most of PUD languages, however, it has the largest number of head directionality features. Furthermore, Polish has the largest DEPREL tag-set but not the largest number of attested features. However, it is possible to find some correspondences, as it is the case of Japanese and Korean.

Before conducting the clustering analysis of the PUD languages, it is possible to check the general tendencies concerning right or left-branching phenomena. For that, the total frequencies of all right and left-branching patterns are calculated, excluding the dependency relations that are always right-branching according to the UD guidelines: “conj”, “appos”, “flat”, and “fixed”. The results are presented in the Table 4.27 and in the Figure 4.28.

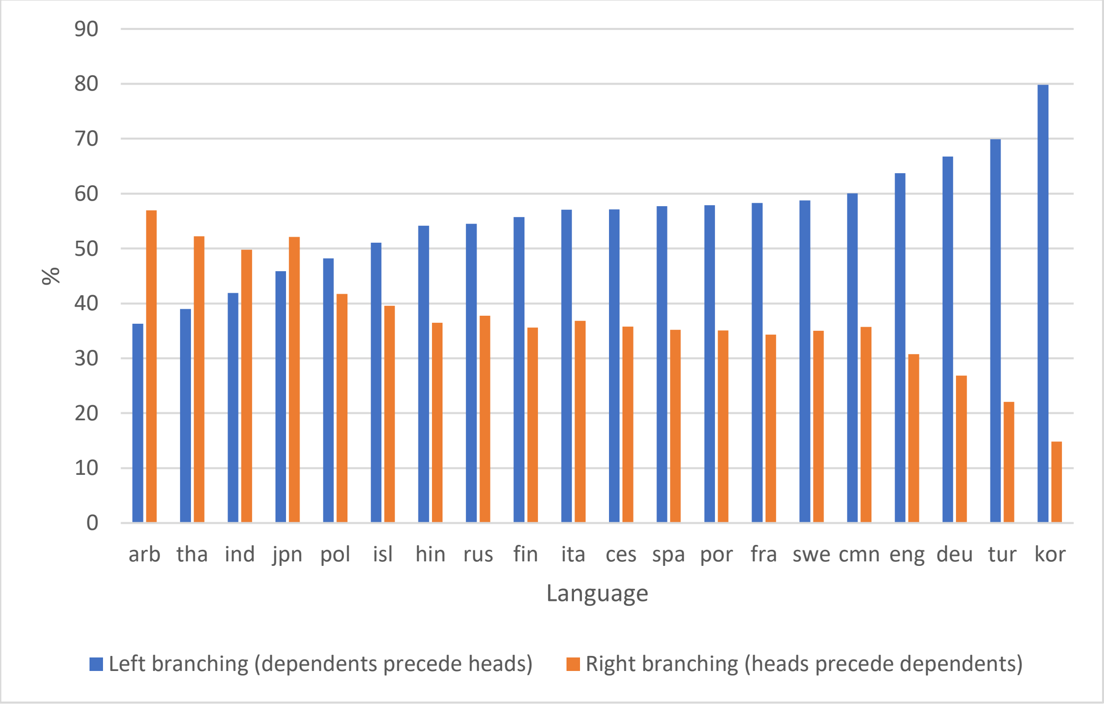
	Number of features
arb	530
cmn	491
ces	593
eng	607
fin	490
fra	462
deu	478
hin	571
isl	637
ind	431
ita	504
jpn	243
kor	309
pol	588
por	496
rus	572
spa	479
swe	590
tha	543
tur	481

Table 4.26. Number of head and dependent features, for each PUD language, with frequency value different from 0.

	Left-branching features	**Right-branching features**
arb	36.32	56.98
cmn	60.06	35.70
ces	57.13	35.80
eng	63.71	30.73
fin	55.76	35.62
fra	58.28	34.32
deu	66.76	26.83
hin	54.15	36.46
isl	51.08	39.58
ind	41.88	49.78
ita	57.09	36.80
jpn	45.85	52.12
kor	79.85	14.80
pol	48.22	41.74
por	57.90	35.05
rus	54.50	37.78
spa	57.74	35.20
swe	58.75	35.01
tha	38.96	52.20
tur	69.91	22.05

Table 4.27. Frequency of left and right-branching features for each PUD language.

Figure 4.28. Overall distribution (in terms of percentage) of right-branching (the head precedes the dependent) and left-branching (the dependent precedes the head) in PUD languages.



It is possible to notice two main tendencies in terms of head and dependent positions: a) Arabic, Thai, Indonesian and Japanese are more right-branching, b) in the other PUD languages, the heads tend to appear after the dependent, especially for English, German, Turkish and Korean (i.e.: percentage of left-branching ordering higher than 60%).

In WALS database, Japanese, Hindi, Turkish, and Korean are classified as OV languages, while the others are considered VO (with exception of German for which no dominant order is attested). Although being an OV language, if all the head and dependent pairs are considered, Japanese is mostly right-branching. This is probably because the team that created the Japanese corpus decided to segment words so that agglutinative suffixes are treated as auxiliaries, and because in the UD framework auxiliaries are dependents rather than heads. Meanwhile, Hindi presents a percentage of left and right-branching similar to other Indo-European languages. The clear tendency of being left-branching which would have been expected for OV languages is only observed in Turkish and Korean. The OV languages are clustered together in the dendrograms obtained via the analysis of lang2vec language vectors (Figures 4.12 and 4.13) and with MarsaGram linear vectors compared with cosine distance metrics (Figure 4.23). On the other hand, Arabic, Thai, Indonesian, and Japanese appear as part of the same cluster in the dendrogram built with the vectors composed by all MarsaGram properties, however, in the same cluster are also present clearly left-branching languages (e.g.: Korean and Turkish).

Beside the analysis of the general right and left-branching tendencies, a more precise comparison of PUD languages was conducted via the dissimilarity matrices (Annexes 21 and 22) obtained when the language vectors composed with the head directionality features were compared. The figures 4.29 and 4.30 present the dendrograms obtained for the Euclidean and cosine distance matrices respectively.

As it was the case for the clustering analysis with MarsaGram vectors, when the ordering of heads and dependents are scrutinized, the dendrograms obtained with the different distance metrics provide dissimilar language clusters. In both dendrograms, the Romance languages form one single cluster positioned on the left side of the figures. Nevertheless, when cosine distances are considered, Germanic languages (with the exception of Icelandic) are grouped together with the Romance ones. In the Euclidean dendrogram, the Germanic sub-group is closer to the Slavic one (which in both cases include Icelandic). In both figures, it is noticeable that Thai, Arabic and Indonesian (i.e.: the most right-branching languages) are grouped closer to the Germanic and Slavic languages.

Figure 4.29. Cluster dendrogram obtained from the Euclidean dissimilarity matrix calculated with the comparison of the PUD head directionality language vectors.

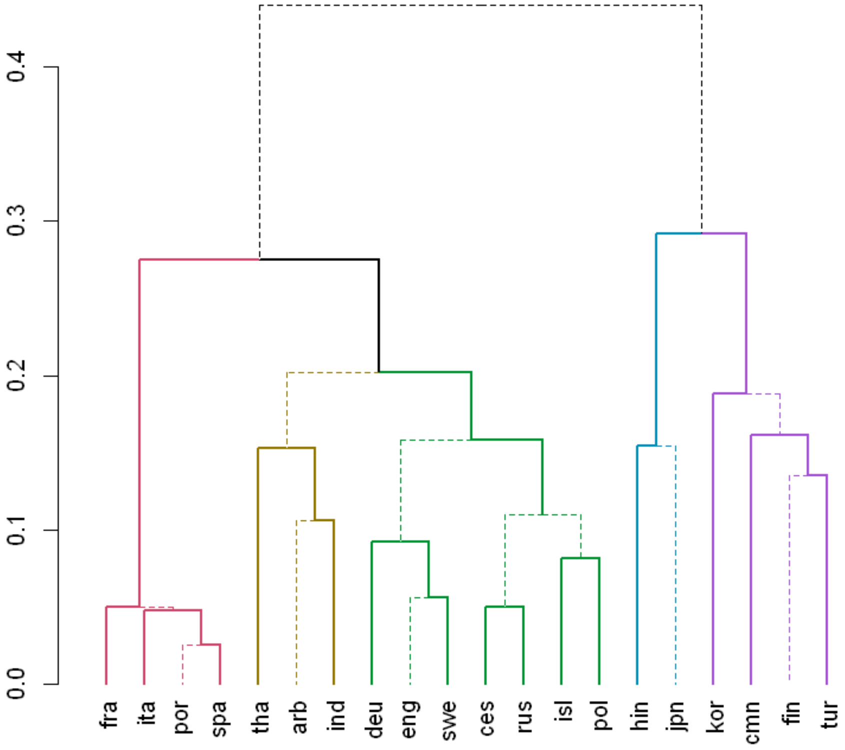
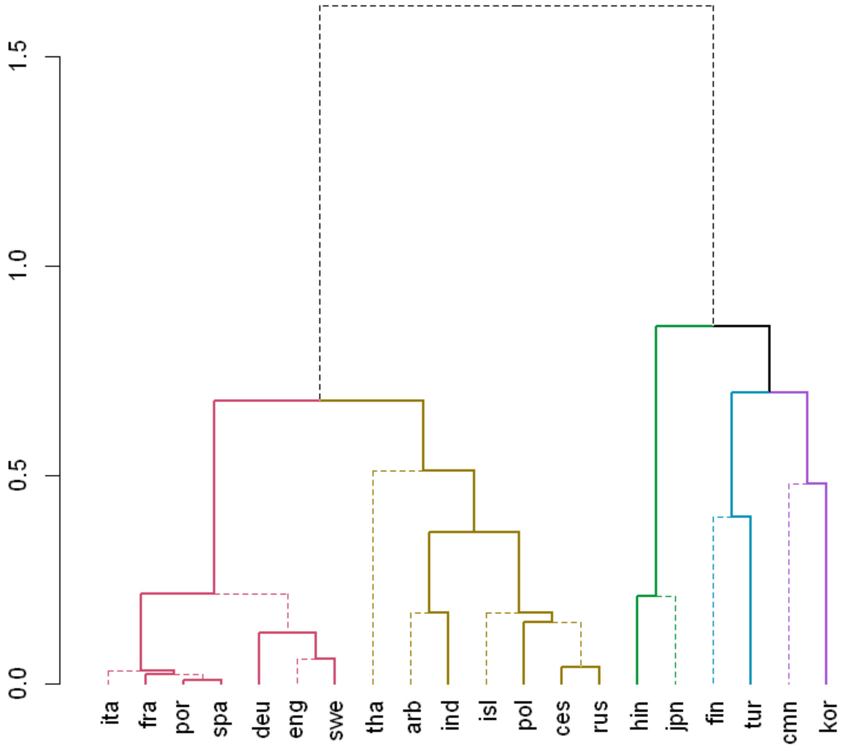


Figure 4.30. Cluster dendrogram obtained from the cosine dissimilarity matrix calculated with the comparison of the PUD head directionality language vectors.



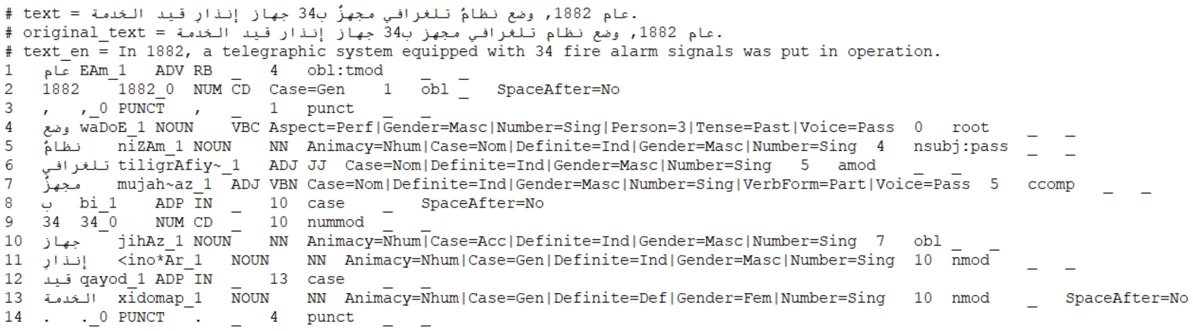
While in the general analysis of left and right-branching tendencies, Japanese was closer to Thai, Arabic, and Indonesian, in these dendrograms, it is clustered with Hindi and closer to other OV languages (i.e.: Turkish and Korean). The large cluster containing all OV languages from PUD also includes Finnish and Chinese.

When compared to the genealogical classification 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 grouped together, they are not classed among Romance ones. Moreover, although not being genealogically related, the proximity between Finnish and Turkish (also verified in other dendrograms) is similarly attested with the head directionality analysis.

As previously mentioned, both dendrograms present similarities. When we analyse in details the features concerning the cases where the classifications do not follow the genealogical one, it is possible to notice that:

1) For Thai, Arabic, and Indonesian, these three languages have only one pattern with higher frequencies than the other PUD languages (“PART_advmod_precedes_ADJ”) and also one pattern with lower frequencies (“ADJ_amod_precedes_NOUN”). No pattern where the dependent follows the head was identified as being more specific of these three languages. Arabic and Indonesian are closer in the dendrograms, and these two languages have 3 common features with higher frequency (“NOUN_nsubj:pass_follows_NOUN” being exclusive). Moreover, the pattern “ADJ_amod_precedes_NOUN” is less frequent for Arabic and Indonesian when compared to the values in the other PUD languages. An example of the exclusive pattern “NOUN_nsubj:pass_follows_NOUN” from the Arabic corpus is presented in Figure 4.31.

Figure 4.31. Sentence from the Arabic PUD corpus where the word order feature “NOUN_nsubj:pass_follows_NOUN” is attested. The dependent (NOUN) with the dependency label “nsubj:pass” corresponds to token 5 (“نظامٌ”) and the head (NOUN), to token 4 (“وضع”).



```
# text = عام 1882, وضع نظامٌ تلغرافي مجهزٌ ب34 جهاز إنذارٍ قيد الخدمة.
# original_text = عام 1882, وضع نظام تلغرافي مجهز ب34 جهاز إنذار قيد الخدمة.
# text_en = In 1882, a telegraphic system equipped with 34 fire alarm signals was put in operation.
1   عام EAm_1   ADV RB  _   4   obl:tmod    _   _
2   1882    1882_0  NUM CD  Case=Gen    1   obl _   SpaceAfter=No
3   ,   ,_0 PUNCT   ,   _   1   punct   _   _
4   وضع waDoE_1 NOUN    VBC Aspect=Perf|Gender=Masc|Number=Sing|Person=3|Tense=Past|Voice=Pass  0   root    _   _
5   نظامٌ   niZAm_1 NOUN    NN  Animacy=Nhum|Case=Nom|Definite=Ind|Gender=Masc|Number=Sing  4   nsubj:pass  _   _
6   تلغرافي tiligrAfiy~_1   ADJ JJ  Case=Nom|Definite=Ind|Gender=Masc|Number=Sing   5   amod    _   _
7   مجهزٌ   mujah~az_1  ADJ VBN Case=Nom|Definite=Ind|Gender=Masc|Number=Sing|VerbForm=Part|Voice=Pass  5   ccomp   _   _
8   ب   bi_1    ADP IN  _   10  case    _   SpaceAfter=No
9   34  34_0    NUM CD  _   10  nummod  _   _
10  جهاز    jihAz_1 NOUN    NN  Animacy=Nhum|Case=Acc|Definite=Ind|Gender=Masc|Number=Sing  7   obl _   _
11  إنذارٍ  <ino*Ar_1   NOUN    NN  Animacy=Nhum|Case=Gen|Definite=Ind|Gender=Masc|Number=Sing  10  nmod    _   _
12  قيد qayod_1 ADP IN  _   13  case    _   _
13  الخدمة  xidomap_1   NOUN    NN  Animacy=Nhum|Case=Gen|Definite=Def|Gender=Fem|Number=Sing   10  nmod    _   SpaceAfter=No
14  .   ._0 PUNCT   .   _   4   punct   _   _
```

2) In both dendrograms, Icelandic and Polish are positioned close together, these two languages have 15 patterns with higher frequencies than the rest of the PUD collection (5 “precedes” and 10 “follows”). Of these 15 features, 6 are exclusive of these two languages (with a very low frequency):

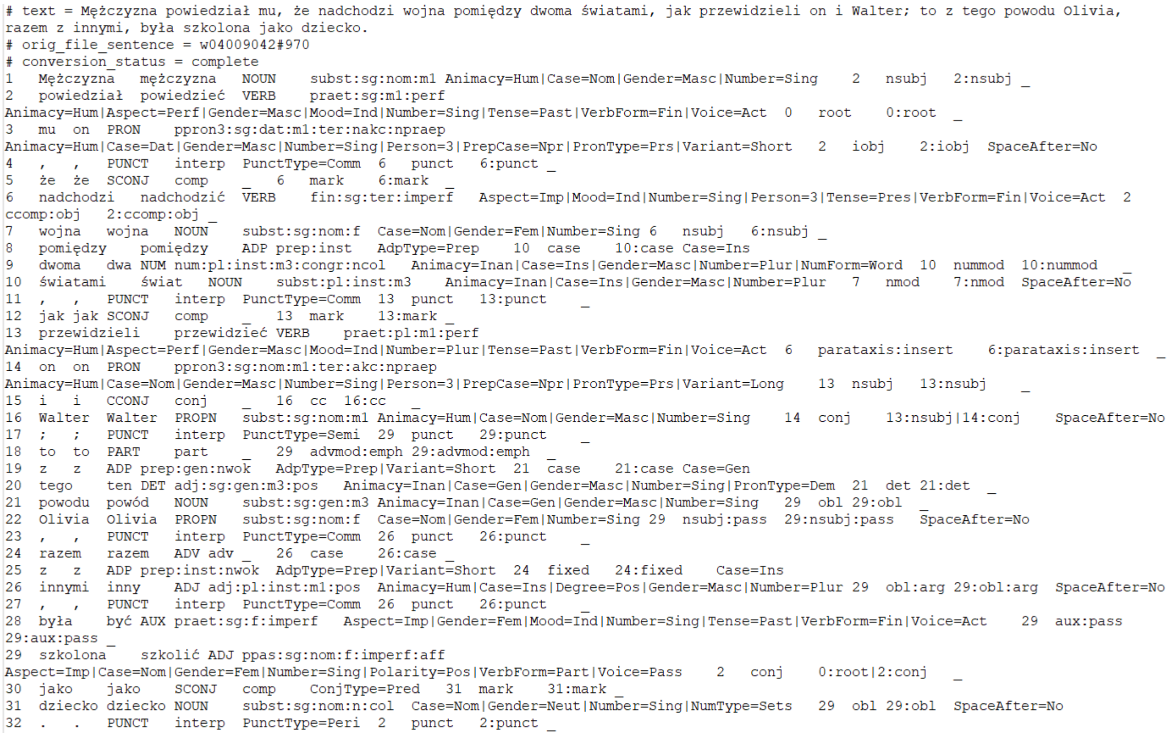
   a. Right-branching: “PROPN_obl:arg_follows_NOUN”, “PRON_obl:arg_follows_NOUN”, “PRON_obl:arg_follows_ADJ”, and “NOUN_obl:arg_follows_NOUN”.

   b. Left-branching: “ADV_case_precedes_ADJ” and “SYM_nmod_precedes_PROPN”.

   An example of the left-branching pattern “ADV_case_precedes_ADJ” from the Polish PUD corpus is presented in the Figure 4.32.

3) Hindi and Japanese form specific clusters in both dendrograms, these two languages have 17 common features with higher frequencies than the other PUD languages (10 “follows” and 7 “precedes”). Moreover, Hindi and Japanese have 4 features in common with lower frequencies: “PROPN_appos_follows_NOUN”, “NOUN_conj_follows_VERB”, “VERB_conj_follows_NOUN”, and “NOUN_conj_follows_PROPN”. The exclusive patterns of these two languages correspond to left-branching relations: “VERB_acl_precedes_NUM” and “ADV_obj_precedes_VERB”.

Figure 4.32. Sentence from the Polish PUD corpus where the word order feature “ADV_case_precedes_ADJ” is attested. The dependent (ADV) with the dependency label “case” corresponds to token 24 (“razem”) and the head (ADJ), to token 26 (“innymi”).



```
# text = Mężczyzna powiedział mu, że nadchodzi wojna pomiędzy dwoma światami, jak przewidzieli on i Walter; to z tego powodu Olivia,
razem z innymi, była szkolona jako dziecko.
# orig_file_sentence = w04009042#970
# conversion_status = complete
1   Mężczyzna   mężczyzna   NOUN    subst:sg:nom:m1 Animacy=Hum|Case=Nom|Gender=Masc|Number=Sing    2   nsubj   2:nsubj _
2   powiedział  powiedzieć  VERB    praet:sg:m1:perf
Animacy=Hum|Aspect=Perf|Gender=Masc|Mood=Ind|Number=Sing|Tense=Past|VerbForm=Fin|Voice=Act  0   root    0:root  _
3   mu  on  PRON    ppron3:sg:dat:m1:ter:nakc:npraep
Animacy=Hum|Case=Dat|Gender=Masc|Number=Sing|Person=3|PrepCase=Npr|PronType=Prs|Variant=Short   2   iobj    2:iobj  SpaceAfter=No
4   ,   ,   PUNCT   interp  PunctType=Comm  6   punct   6:punct _
5   że  że  SCONJ   comp    _   6   mark    6:mark  _
6   nadchodzi   nadchodzić  VERB    fin:sg:ter:imperf   Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act  2
ccomp:obj   2:ccomp:obj _
7   wojna   wojna   NOUN    subst:sg:nom:f  Case=Nom|Gender=Fem|Number=Sing 6   nsubj   6:nsubj _
8   pomiędzy    pomiędzy    ADP prep:inst   AdpType=Prep    10  case    10:case Case=Ins
9   dwoma   dwa NUM num:pl:inst:m3:congr:ncol   Animacy=Inan|Case=Ins|Gender=Masc|Number=Plur|NumForm=Word  10  nummod  10:nummod   _
10  światami    świat   NOUN    subst:pl:inst:m3    Animacy=Inan|Case=Ins|Gender=Masc|Number=Plur   7   nmod    7:nmod  SpaceAfter=No
11  ,   ,   PUNCT   interp  PunctType=Comm  13  punct   13:punct    _
12  jak jak SCONJ   comp    _   13  mark    13:mark _
13  przewidzieli    przewidzieć VERB    praet:pl:m1:perf
Animacy=Hum|Aspect=Perf|Gender=Masc|Mood=Ind|Number=Plur|Tense=Past|VerbForm=Fin|Voice=Act  6   parataxis:insert    6:parataxis:insert  _
14  on  on  PRON    ppron3:sg:nom:m1:ter:akc:npraep
Animacy=Hum|Case=Nom|Gender=Masc|Number=Sing|Person=3|PrepCase=Npr|PronType=Prs|Variant=Long    13  nsubj   13:nsubj    _
15  i   i   CCONJ   conj    _   16  cc  16:cc   _
16  Walter  Walter  PROPN   subst:sg:nom:m1 Animacy=Hum|Case=Nom|Gender=Masc|Number=Sing    14  conj    13:nsubj|14:conj    SpaceAfter=No
17  ;   ;   PUNCT   interp  PunctType=Semi  29  punct   29:punct    _
18  to  to  PART    part    _   29  advmod:emph 29:advmod:emph  _
19  z   z   ADP prep:gen:nwok   AdpType=Prep|Variant=Short  21  case    21:case Case=Gen
20  tego    ten DET adj:sg:gen:m3:pos   Animacy=Inan|Case=Gen|Gender=Masc|Number=Sing|PronType=Dem  21  det 21:det  _
21  powodu  powód   NOUN    subst:sg:gen:m3 Animacy=Inan|Case=Gen|Gender=Masc|Number=Sing   29  obl 29:obl  _
22  Olivia  Olivia  PROPN   subst:sg:nom:f  Case=Nom|Gender=Fem|Number=Sing 29  nsubj:pass  29:nsubj:pass   SpaceAfter=No
23  ,   ,   PUNCT   interp  PunctType=Comm  26  punct   26:punct    _
24  razem   razem   ADV adv _   26  case    26:case _
25  z   z   ADP prep:inst:nwok  AdpType=Prep|Variant=Short  24  fixed   24:fixed    Case=Ins
26  innymi  inny    ADJ adj:pl:inst:m1:pos  Animacy=Hum|Case=Ins|Degree=Pos|Gender=Masc|Number=Plur 29  obl:arg 29:obl:arg  SpaceAfter=No
27  ,   ,   PUNCT   interp  PunctType=Comm  26  punct   26:punct    _
28  była    być AUX praet:sg:f:imperf   Aspect=Imp|Gender=Fem|Mood=Ind|Number=Sing|Tense=Past|VerbForm=Fin|Voice=Act    29  aux:pass
29:aux:pass _
29  szkolona    szkolić ADJ ppas:sg:nom:f:imperf:aff
Aspect=Imp|Case=Nom|Gender=Fem|Number=Sing|Polarity=Pos|VerbForm=Part|Voice=Pass    2   conj    0:root|2:conj   _
30  jako    jako    SCONJ   comp    ConjType=Pred   31  mark    31:mark _
31  dziecko dziecko NOUN    subst:sg:nom:n:col  Case=Nom|Gender=Neut|Number=Sing|NumType=Sets   29  obl 29:obl  SpaceAfter=No
32  .   .   PUNCT   interp  PunctType=Peri  2   punct   2:punct _
```

4) On the extreme right side of both dendrograms we identify two sub-clusters, one composed of Finnish and Turkish, and the other one of Chinese and Korean.
    a. Finnish and Turkish have 25 common features with higher frequencies (22 left-branching). Of them, 7 are exclusive (all left-branching).
    b. Chinese and Korean have 30 common features with higher frequencies (26 left-branching). 18 out of these 30 patterns are exclusive (15 left-branching). Moreover, Chinese and Korean have lower frequencies for 3 patterns (1 right-branching): “CCONJ_cc_precedes_VERB”, NOUN_appos_follows_PROPN”, and “ADJ_amod_follows_NOUN”. The last one, being exclusive of these two languages.

The obtained dendrograms show that a more fine-grained classification is obtained when compared to the analysis provided by the examination of the general tendencies, as it is the case of Japanese which is closer to Hindi, Korean and Turkish in the figures, similar to the classification obtained using lang2vec vectors.

4.7 Quantitative Typological Classification Using Verb and Object Ordering

The previous method considered all existing head and dependent orderings attested in PUD corpora to generate language vectors, being aligned with Hawkins idea (1983) that a language classification should not be based only on the verb and object positions in the sentences. However, the importance of these components is observed in Greenberg's universals (1963), and also in Dryer's analysis of correlations (1992).

Thus, it seems relevant to conduct a quantitative typological analysis of the observed word ordering phenomena in PUD corpora concerning these two elements. The idea is to extract, from the obtained patterns in the previous section, the features where the dependent has the dependency label "obj" (object), and the head has the part-of-speech label "VERB". After that, PUD language vectors were generated with these features associated to the frequency of occurrence of each word order phenomenon. In total, 13 OV and 12 VO features were attested, as presented in Tables 4.28 and 4.29 respectively. The difference between them concerns the part-of-speech of the dependents.

OV Features	Number of PUD corpora
CCONJ_obj_precedes_VERB	1 (swe)
SYM_obj_precedes_VERB	1 (deu)
PRON_obj_precedes_VERB	18 (exception: cmn, ind)
DET_obj_precedes_VERB	6 (ces, eng, deu, hin, pol, tha)
PROPN_obj_precedes_VERB	13 (exception: arb, eng, fra, ind, ita, spa, swe)
SCONJ_obj_precedes_VERB	2 (isl, rus)
ADJ_obj_precedes_VERB	7 (ces, eng, fin, hin, jpn, swe, tur)
ADV_obj_precedes_VERB	2 (jpn, hin)
VERB_obj_precedes_VERB	2 (tha, tur)
NOUN_obj_precedes_VERB	19 (exception: tha)
NUM_obj_precedes_VERB	6 (deu, hin, jpn, kor, swe, tur)
ADP_obj_precedes_VERB	2 (ita, por)
X_obj_precedes_VERB	2 (deu, tur)

Table 4.28. Ensemble and overall distribution of OV features extracted from PUD corpora.

Regarding OV features, the most common ones inside PUD corpora (i.e.: present in 15 or more languages) concern dependents which are pronouns and nouns. Moreover, only Indonesian does not have any occurrence of the listed OV patterns. Meanwhile, in the case of VO features, the most common ones are linked to dependents which are either proper nouns, nouns, numerals, or pronouns. Korean and Turkish are the only PUD languages without any occurrences of VO patterns in their corpus.

VO Features	Number of PUD corpora
PART_obj_follows_VERB	1 (cmn)
PROPN_obj_follows_VERB	16 (exception: hin, jpn, kor, tur)
VERB_obj_follows_VERB	7 (cmn, eng, fin, hin, ita, rus, tha)
ADP_obj_follows_VERB	2 (isl, por)
ADJ_obj_follows_VERB	13 (exception: deu, hin, ind, ita, jpn, kor, tur)
ADV_obj_follows_VERB	10 (ces, eng, fin, fra, isl, ita, rus, spa, swe, tha)
DET_obj_follows_VERB	6 (cmn, ces, eng, deu, ita, pol)
NOUN_obj_follows_VERB	18 (exception: kor, tur)
SYM_obj_follows_VERB	13 (exception: arb, cmn, hin, jpn, kor, swe, tur)
X_obj_follows_VERB	3 (cmn, ind, ita)
NUM_obj_follows_VERB	15 (exception: deu, hin, jpn, kor, tur)
PRON_obj_follows_VERB	17 (exception: jpn, kor, tur)

Table 4.29. Ensemble and overall distribution of VO features extracted from PUD corpora.

When analysing the frequency of occurrences of the different features, it is possible to notice that for OV patterns, the values are much higher for attested OV languages (i.e.: Japanese, Hindi, Turkish, and Korean). The inverse is observed for VO features, as expected. Moreover, in both cases, the frequency of occurrences is much higher in features concerning dependents which are nouns, pronouns and proper nouns. When the phenomenon is present in few corpora (i.e.: less than 5), the frequency is usually very low.

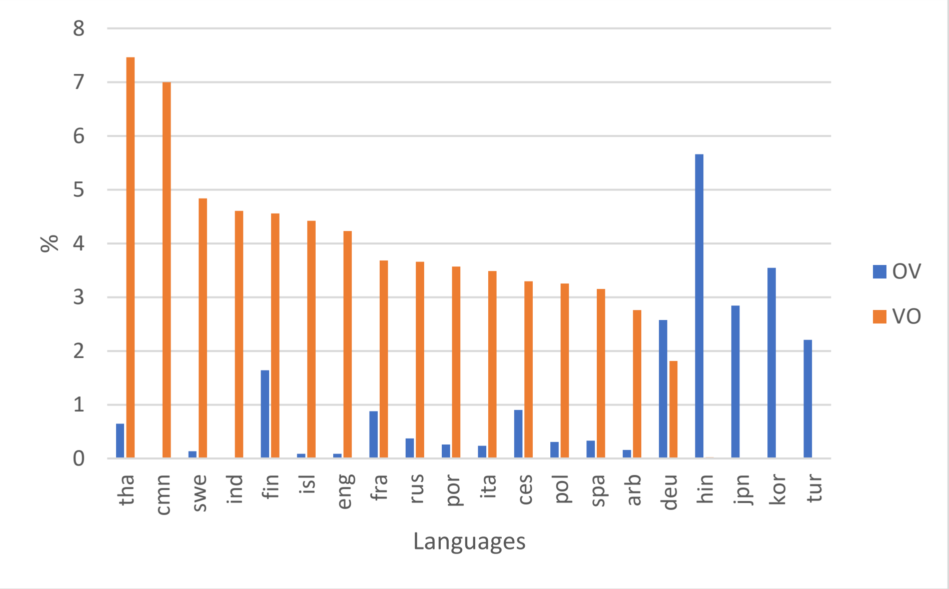
In the theoretical typological studies previously described, when verb and object are examined, the object concerns only nouns. In this thesis, all possible part-of-speech regarding the dependent are retained, thus, allowing a more fine-grained language classification. If only nominal objects were considered, languages would follow WALS classification (VO or OV). Thus, we decided to add all possible objects so that a more detailed classification would be obtained.

As it was conducted for all the head directionality features, it is possible to analyse the overall tendencies of PUD languages in terms of verb and object position as presented in Table 4.30 and in Figure 4.31.

	%OV	%VO
arb	0.16	2.76
cmn	0.01	7.00
ces	0.90	3.29
eng	0.08	4.23
fin	1.64	4.56
fra	0.88	3.68
deu	2.57	1.81
hin	5.66	0.02
isl	0.08	4.42
ind	0.00	4.60
ita	0.23	3.48
jpn	2.84	0.00
kor	3.55	0.00
pol	0.30	3.25
por	0.26	3.57
rus	0.37	3.66
spa	0.33	3.15
swe	0.13	4.84
tha	0.65	7.46
tur	2.20	0.00

Table 4.30. Percentage of occurrences of OV and VO features inside each PUD corpus.

Figure 4.31. General distribution in terms of frequency of OV and VO features for each PUD language.



It is noticeable that the expected distinction between VO and OV languages is also observed in the graph presented in the Figure 4.31. Moreover, the WALS classification concerning German language (i.e.: no dominant order) is also coherent with what is noted in the graph. The percentage of OV features in VO languages are mostly due to the position of the pronoun objects, while for Hindi, the slight amount of VO features concerns noun and verb objects, and for Japanese, only noun type. Moreover, Finnish and German have higher frequencies of proper nouns objects being positioned before the verb, when compared to other VO languages.

Another interesting aspect that is noticeable in the Figure 4.31 concerns the total frequency of direct objects in each corpus. As all the corpora are parallel, it is clear that some languages favour this type of dependency relation to express meaning: it is the case of VO languages Thai and Chinese, and OV language Hindi.

Following the general analysis of the OV and VO characteristics of each PUD language, the dissimilarity matrices regarding Euclidean and cosine distance measures where generated (Annexes 23 and 24 respectively). Moreover, the Figures 4.32 and 4.33 present the obtained dendrograms.

Figure 4.32. Cluster dendrogram obtained from the Euclidean dissimilarity matrix calculated with the comparison of the PUD OV/VO language vectors.

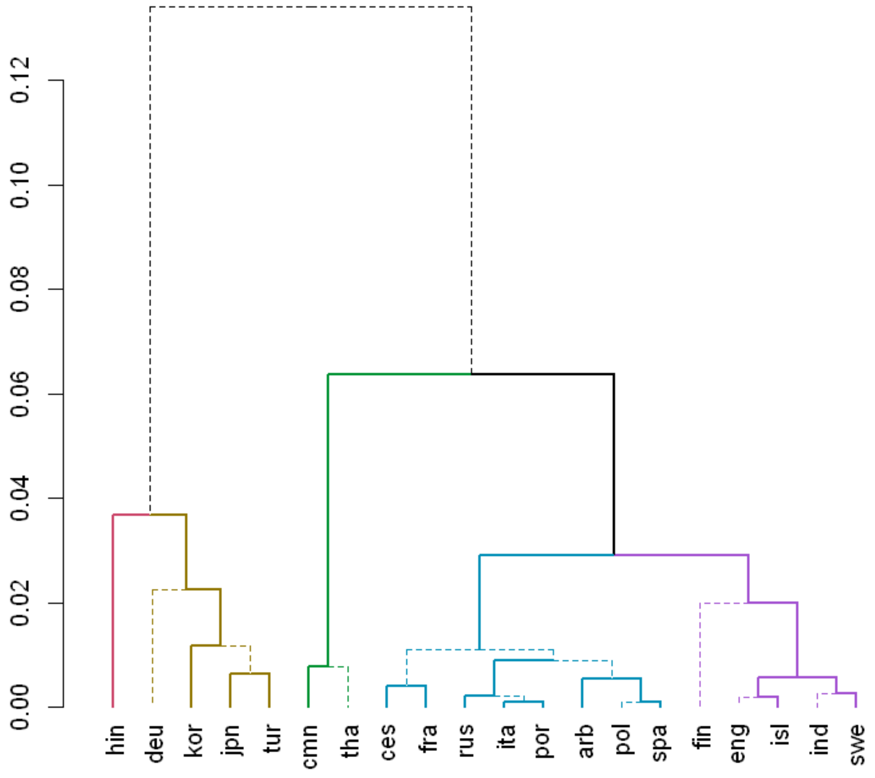
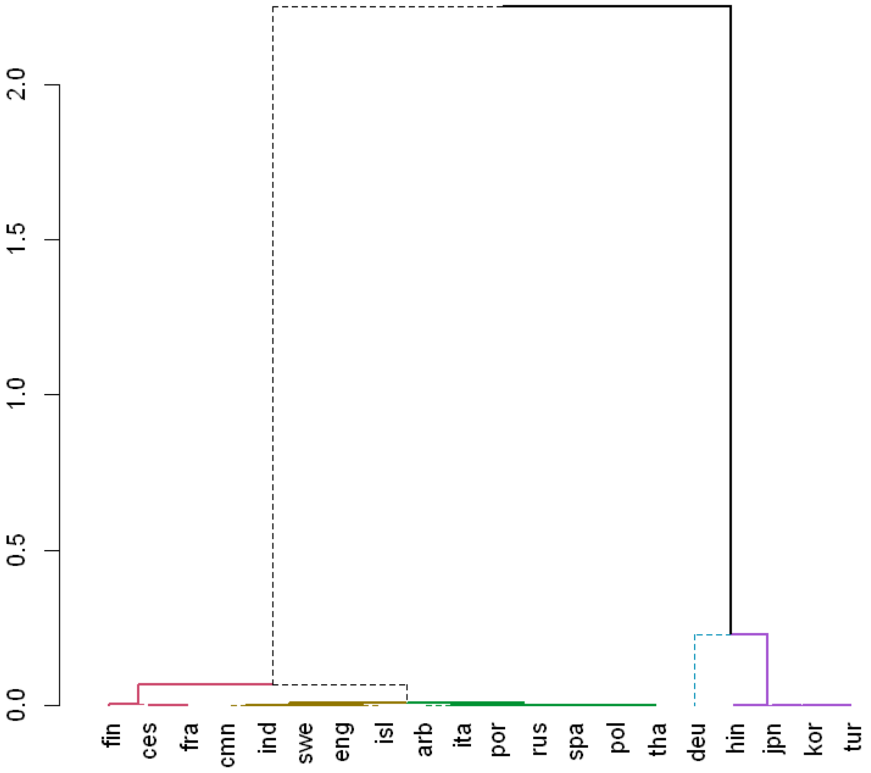


Figure 4.33. Cluster dendrogram obtained from the cosine dissimilarity matrix calculated with the comparison of the PUD OV/VO language vectors.



The different distance metrics used to generate the dendrograms generate different language clusters. One similitude concerns the division in two parts of each figure: one contains all OV languages and German (with no dominant order), and the other, the VO languages. In the cosine dendrogram, German forms an isolated sub-group, while in the Euclidean one it is placed closer to Korean, Japanese and Turkish (in this case, Hindi is in an isolated sub-cluster).

Another visible difference concerns how PUD Romance, Germanic and Slavic languages are classed. French and Czech are closer in both dendrograms, however, they form a cluster with Finnish only in the cosine dendrogram. On the other hand, Finnish is placed together with Germanic languages (except for German) and Indonesian in the Euclidean clustering graph. This proximity between Indonesian and Swedish, English, and Icelandic is also observed in the cosine dendrogram. Slavic languages (except for Czech) are clustered with Romance languages (except for French) in a group which also contains Arabic. Thai language is placed together with this eclectic group in the cosine dendrogram but, for the Euclidean one, it forms a small sub-cluster with Chinese (which is coherent with the similarity of these two languages regarding the general OV/OV analysis).

The obtained dendrograms are coherent with the general classification of languages in typological databases, but they present more fine-grained analysis that allow languages to be classified in a more precise way concerning the attested word ordering between objects and verbs.

4.8 General Discussion

In the previous sub-sections, several methods to compare languages were presented: from the genealogical classification to methods involving syntactic features extracted from typological databases, and other quantitative strategies regarding the extraction of syntactic patterns in PUD corpora.

It is noticeable, when examining all obtained dendrograms, that each different method offers a specific view on the syntactic phenomena occurring in the different languages. Nevertheless, some similarities can be found between the different strategies. Also, it is possible to see that some methods are more coherent with well-known typological classifications (e.g.: Hawkins' language types, 1983).

The calculated dissimilarity matrices allow the identification of the most similar language-pairs, and these results that vary according to each specific method. On the other hand, when the dendrograms are generated, as what is calculated is the variance between clusters, the languages with the smallest distance between them are not always the closest ones in the figures. Thus, when checking the possible correlation between the obtained language classifications and the results when corpora are combined to train dependency parsing tools, it is important to see whether the language-pairs identified via dissimilarity matrices are a better choice than the closest languages inside the dendrograms.

As it was explained in the section “Objective and Hypothesis of Research”, our aim is to find the best corpus-based way of classifying languages typologically in terms of improvement of dependency parsing results. It is the evaluation of the correlation between the language classifications presented in this part with the parsing results obtained via corpora association that will determine which method could be more relevant for this specific task of Natural Language Processing.

In this section, all 20 PUD languages were analysed using the different proposed typological strategies. Once the best typological strategy is defined for the improvement of dependency parsing metrics, the EU languages which are not part of PUD will be examined in details in comparison to the PUD ones (Section 6).

5. Dependency Parsing Improvement with Typological Strategies

This section addresses the second hypothesis of this thesis: “The typological classification using the quantitative syntactic typological distance between languages is an efficient way to identify related languages whose corpora can be combined to optimize the performance of deep learning tools in terms of automatic syntactic annotation”.

The aim is to understand how the different typological approaches presented in the previous section correlate with the parsing results obtained when different languages are combined. The dependency parsing software selected to perform the automatic annotation in terms of dependency relations is the UDify tool which uses state-of-the-art technologies regarding this specific task.

First, the material used for the experiments will be presented, followed by the methodology built for comparing the different typological strategies with regard to the LAS and MLAS parsing results. Then, the obtained results will be displayed with a detailed analysis and the identification of the best typological approaches.

As previously explained, the experiments were conducted, in this step, with the PUD corpora which concern 10 EU languages. The other 14 EU languages will be analysed using the best typological strategies in the next section.

5.1 Tools

In this sub-section, the deep-learning tool that has been selected for the ensemble of parsing experiments is detailed, together with the multilingual language model which is an important component of this software.

5.1.1 UDify

UDify is an NLP tool developed by Dan Kondratyuk and Milan Straka (2019) as a deep-learning tool capable of predicting lemmas, part-of-speech, morphological features, and dependencies relations using multilingual BERT (Devlin et al., 2018) in its embedding, encoder and projection layers.

For each NLP task, this system has specific attention and prediction layers. For part-of-speech tagging, it uses a softmax layer along each word input, calculating a distribution of probabilities over the possible labels in the vocabulary to identify the most appropriate UPOS label.

Concerning morphological features (FEATS), the tool uses a similar architecture as the one built for UPOS, however, each annotation (i.e.: the combination of a feature and a value) is treated as one single token, this way, it eliminates invalid combinations of features.

Lemmatization is done by predicting a class representing an edit script which is responsible for the character operations that are required to transform the word form into the lemma using the Wagner-Fischer algorithm (Wagner and Fischer, 1974).

And, finally, dependencies relations are determined by a graph-based biaffine attention parser developed by Dozat et al. (2016) where the bidirectional LSTM layers have been replaced by the multilingual BERT (mBERT) language model. The final embeddings are projected through arc-head and arc-dep feedforward layers that are combined using biaffine attention, inferring, this way, a probability distribution of arc heads for each word. The parsing tree is then decoded from the obtained distribution by using the “Chu-Liu/Edmonds algorithm” (Chu, 1965; Edmonds, 1967).

As previously mentioned, UDify relies on the fine-tuning of mBERT for predictions. Instead of using just the last layer of this language model, or restricting it to a specific set of layers, UDify creates a simple layer-wise dot-product attention, producing a weighted sum of all intermediate outputs of all twelve mBERT layers (each token with the same weights). To avoid overfitting, a layer dropout is defined, redistributing probability mass to all other layers, thus, forcing the system to consider all of them. For each task, a different layer attention is computed. UDify follows the strategy developed by Howard and Ruder (2018) called Universal language model fine-tuning (ULMFiT) with some modifications in terms of parameters choices (e.g.: dividing the network into two parameters groups) and regarding the learning rate decay (i.e.: inverse square root instead of linear one). During the fine-tuning step, UDify proceeds with the strategy of masking words randomly as it reduces the tendency of overfitting.

Kondratyuk and Straka (2019) showed that a UDify model trained using all corpora from the Universal Dependencies dataset version 2.3 (one hundred twenty-four languages) performs well compared to UDPipe 2.0 (Straka, 2018), a state-of-the-art NLP tool: dependency parsing results are comparable for well-resourced languages and better for under-resourced languages.

Although using language combination as a strategy to improve results of languages with few resources, the authors did not apply any typological strategy to identify similar languages which could have provided even better parsing results.

5.1.2 Multilingual BERT (mBERT)

BERT is a language representation model (Devlin, 2018) developed by Google which stands for “Bidirectional Encoder Representations from Transformers” and that was designed to pre-train deep bidirectional representations using unannotated text considering both left and right contexts in all layers. It can be fine-tuned with additional output layers in many different NLP tasks (as is the case in UDify tool). The first BERT version concerned only the English language.

In 2019, a multilingual version (mBERT[34]) was created with data from 104 languages. It contains 12 layers, 768 hidden states, 12 heads and 110 million parameters. These languages were chosen as they were the ones with the highest amount of textual data available in Wikipedia at the moment when the texts were extracted. For each language, its entire dump was mined, excluding only users and talk pages.

As the size of Wikipedia dump varies significantly for each language, an exponential smoothed weighting of the data was performed during the pre-training data creation to avoid under-representation of the under-resourced ones, and overfitting of the languages with bigger amount of data.

All PUD languages are present in mBERT, and the range[35] of the size of the training data used for each language is presented in Table 5.1.

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

Table 5.1. List of PUD languages and the respective size range of the training corpus used to generate the mBERT language model (Wu and Dredze, 2020).

---

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

[35] In the work presented by Wu and Dredze (2020), languages are grouped in terms of the size of the data used to train mBERT. The exact value of each training corpus was not available in their study. This source was selected as it provided at least an approximative estimation of the size of the language representation for all PUD languages.

It is possible to notice that "there is a huge discrepancy regarding the amount of data from different languages used to generate the mBERT language model. As expected, English is the language that has the largest pre-training corpus size, followed by German, French, Spanish and Russian. It is possible to observe that the largest mBERT pre-training corpora come from Indo-European languages, only Chinese and Japanese languages are also quite well represented. Icelandic is the one with the smaller pre-training corpus, therefore, not as well represented in this language model as the other PUD languages" (Alves et al., 2022).

Concerning the 24 EU languages, only Maltese is not represented in the mBERT language model. Table 5.2 presents the size range of this set of languages.

Language	**Size Range (GB)**
eng	[11.314, 22.627]
deu, fra, spa	[2.828, 5.657]
ita, pol, por	[1.414, 2.828]
ces, swe, nld, hun	[0.707, 1.414]
fin, ron	[0.354, 0.707]
bul, hrv, dan, est, ell, slk, slv	[0.177, 0.354]
lav, lit	[0.088, 0.177]
gle	[0.022, 0.044]

Table 5.2. List of EU languages and the respective size range of the training corpus used to generate the mBERT language model (Wu and Dredze, 2020).

The same divergence in terms of the size of the language representation in mBERT that was observed in the PUD set is attested when EU languages are considered. Germanic (except for Danish) and Romance (except for Romanian) languages are the ones with the largest training corpora. The ones with the smallest representation in this language model are Latvian, Lithuanian and Irish. Slavic languages are not the smallest ones; however, their training set sizes are closer to the low-resourced ones.

Even though mBERT applies a weighting procedure to overcome the differences in terms of training corpora sizes, in the end, the system is not homogeneous, thus, affecting multilingual tools using this language model (i.e.: languages with bigger mBERT training corpora tend to have better scores).

Due to this fact, many studies presenting typological strategies for improving dependency parsing results avoid using systems based on language models (e.g. Glavaš & Vulić, 2021;

Litschko et al., 2020). However, it is undeniable that most of the state-of-the-art dependency parsing tools use pre-trained language models as the base for this task as presented by Otter et al. (2019) in their overview of state-of-the-art tools for NLP tasks.

5.2 Methodology

The methodology applied for the dependency parsing analysis can be divided into three main steps:

1) Definition of the baseline in terms of parsing results;
2) Language combination experiments using UDify;
3) Typological strategies evaluation in relation to parsing results:
    a. In terms of overall correlation measures;
    b. In terms of choice of best language pairs.

Each step is detailed further in the sub-sections below.

5.2.1 Definition of the baseline in terms of parsing results

The baseline that is considered for the dependency parsing experiments concerns LAS and MLAS results obtained with UDify trained by each one of the PUD languages alone. With these reference points, it is possible to verify the impact, in terms of these metrics, when languages are combined.

As described beforehand, PUD corpora have a limited size in terms of the number of sentences (i.e.: 1,000 per language), thus, all languages are considered, in this part of the thesis, as low-resourced ones. The aim is to find out which typological strategy is the most optimized one in the low-resourced scenario to be applied to some EU low-resourced languages in the following section.

It is important to mention that even though all the PUD corpora have the same size (i.e.: all languages have the same number of sentences for training, development, and test set), as UDify uses mBERT as a crucial component of its architecture, languages are not equal in terms of overall linguistics resources. The influence of this bias is described in the sub-section that presents the baseline results.

The PUD collection was originally conceived as an ensemble of test sets, however, in this thesis, each corpus was divided into training, development, and test data. For the ensemble of the experiments regarding the establishment of the baseline, the following distribution was adopted:

1) Training set: 600 sentences (first 600 sentences from the original corpus);
2) Development set: 200 sentences (next 200 sentences from the original corpus);
3) Test set: 200 sentences (final 200 sentences from the original corpus).

Thus, the experiments were conducted with a 60/20/20 distribution of ratios in regard to training, development, and test sets.

When comparing dependency parsing metrics, it is important to conduct statistical analysis to check whether the obtained values are significantly different from each other before drawing any conclusion. Thereby, for each language, we decided to vary the random seed value to obtain a set of results with which it is possible to calculate the mean and the standard deviation of each experiment, and the p-value when two scores are compared.

In deep-learning models, the random seed defines how the instances from the corpus (in our case, sentences), that are used for training phase of the model, are divided into smaller sub-sets. Using the same random seed for a defined dataset ensures that the final results are the same, enabling the experiments to be reproducible. However, the variation of this parameter for the same dataset permits the obtention of different models and different results which enables the statistical analysis of the obtained metrics to be conducted (Colas et al., 2018).

The UDify tool can be used without the definition of a random seed. In this case, it establishes a standard value for this parameter. Besides, in the configuration file which defines the ensemble of parameters of the training steps, there is a suggestion for the random seed value: 13370. Thus, the set of the chosen random values for the experiments of this thesis are the standard, 13370, 10, 100, 1000, and 100000. We decided to keep both the standard and the proposed values, and to add four others that correspond to powers of 10. These values were determined as such to cover the perimeter from 10 to 100,000 (the 10,000 is represented by the proposed one, 13370).

Therefore, for each language, 6 different results of LAS and MLAS were obtained. With these values, it was possible to calculate:

1) The arithmetic mean of LAS and MLAS metrics with the following formula:

$$(5.1)\ Mean_{metric} = \frac{1}{n}\left(\sum_{i=1}^{n} metric_i\right)$$

   Where, $metric_i$ is the LAS or MLAS value obtained for each random seed value, and n is the total number of experiments conducted for each language (i.e.: 6).

2) The standard deviation of LAS and MLAS metrics with the subsequent equation:

$$(5.2)\ std\_dev = \sqrt{\frac{\sum_{i=1}^{n}(metric_i - Mean_{metric})^2}{n-1}}$$

Thus, the LAS and MLAS mean and standard deviation values will be presented for each one of the PUD corpora in the results sub-section. These scores were calculated using the Python module "Statistics[36]" via the functions mean() and stdev().

As it was previously stated, the mean and the standard deviation values are necessary when different scenarios are statistically compared to establish if results are similar or different. P-value stands for probability value and can be defined, in a simple manner for two elements, as the probability of obtaining similar results in the ensemble of experiments (Nuzzo, 2014). Thus, it means that the p-value is comprised between 0 and 1 (i.e.: results are always different if p-value is 0, and results are always similar when it is equal to 1). In this thesis, the threshold defined in terms of p-value is 0.01. Thus, when two values of LAS or MLAS are compared, if the p-value is lower than 0.01, we consider that they are statistically different. On the other hand, if the p-value is higher than the threshold, the values are taken as statistically similar. The p-values were calculated using the function ttest_ind_from_stats() of the Python module Scipy (scipy_stats[37]).

The UDify tool is presented by its developers as an optimized tool that can be directly applied to train new dependency parsing models.

---

[36] https://docs.python.org/3/library/statistics.html

[37] https://docs.scipy.org/doc/scipy/reference/stats.html

Thus, all experiments were conducted with the standard values as proposed by the UDify creators (except for the random seed value for the reasons aforementioned):

- Number of epochs: 80;
- Warmup: 500.

Finally, it is important to mention that UDify models do not perform sentence split or tokenization. Thus, the obtained results do not consider possible tokenization problems which may impact differently the ensemble of PUD languages. As the focus of this study is dependency parsing, it seemed judicious to avoid this bias, thus, we decided to keep the gold tokenization of PUD corpora and not perform an automatic one with a different tool.

5.2.2 Language combination experiments using UDify

After the definition of the baseline, the PUD languages were combined in pairs to provide bilingual training sets. Every PUD language was combined with the other 19 ones, a total of 380 experiments.

The corpora association was held only at the training set level, and it was done via simple concatenation of the data:

$$(5.3)\ combined\ training\ set = training\ set_{language1} + training\ set_{language2}$$

Thus, after the combination, the final distribution in terms of sentences was:

- Training set: 1,200 sentences (language$_1$ and language$_2$);
- Development set: 200 sentences (only language$_1$);
- Test set: 200 sentences (only language$_1$).

It means that the development and test sets are the same as the ones used for the establishment of the baseline for each PUD language. Moreover, all the training parameters and the random seed variation strategy were also the same.

It is also important to mention that the added sentences correspond semantically to the sentences that compose the language1 training set. The concatenation was done without delexicalization of the concatenated training set, and without any transliteration to the same alphabet.

Thus, for each language, we obtained a table comprising the mean and the standard deviation of the LAS and MLAS metrics for each possible association with the other PUD languages.

This table was completed with a delta and p-value for each metric (language$_1$) in all tested combinations (languages$_{1and2}$) as defined below:

- Delta:

$$(5.4)\ Delta_{metric} = Mean_{metric_{language1and2}} - Mean_{metric_{language1}}$$

  Thus, a positive delta means that the combination provided an improvement for the respective metric, while a negative value indicates a decrease in the metric result when compared to the baseline.
- p-value: for each language association experiment, the calculated delta was only considered statistically significant if the p-value was lower than 0.01.

With the complete tables, it was possible to analyse, for each language, which are the associations that provide a real improvement, and the ones which correspond to a negative synergy (negative delta).

5.2.3 Typological strategies evaluation in relation to parsing results

With the results of the dependency parsing experiments concerning the language associations, it is possible to check if the typological approaches previously defined provide useful information for the definition of the best language pair to be combined to improve parsing metrics.

Two different evaluation strategies were chosen:

1) The first one concerns the verification, for each typological method, of the correlation between the deltas (LAS and MLAS) and the language distances.
2) The second evaluation determines the number of right combination choices (in terms of the best significant positive delta) in comparison to the results obtained with the empirical association experiments. The right choice means that the language pair proposed by the typological approach (i.e.: with the lowest distance between the two languages) is the same or has the same positive delta (i.e.: is statistically similar) as the best empirical pair.

Each type of evaluation is detailed in the following paragraphs.

Concerning the evaluation using correlation metrics, the idea is to determine, for each language, if there is a correlation between the language distances (as presented in the dissimilarity matrices) and the empirical delta (LAS and MLAS) obtained when languages were associated.

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 (Tutorials, S. P. S. S, 2022), while the second one determines how well the relationship between two variables can be defined as a monotonic function (Lehman, 2005).

The correlation coefficients vary from -1 to 1:

- -1 means that the variables are inversely correlated;
- 0 means that the variables do not correlate;
- 1 means that the variables are directly correlated.

In our case, the correlation coefficients are expected to be negative as a smaller distance between languages should provide a higher delta.

Thus, for each language and each typological strategy, Pearson's and Spearman's correlation coefficients were calculated for both LAS and MLAS metrics. To compare the typological strategies, we verified the number of languages for which the correlation coefficient was inferior to -0.7 (strong inverse correlation) and between -0.7 and -0.5 (moderate inverse correlation).

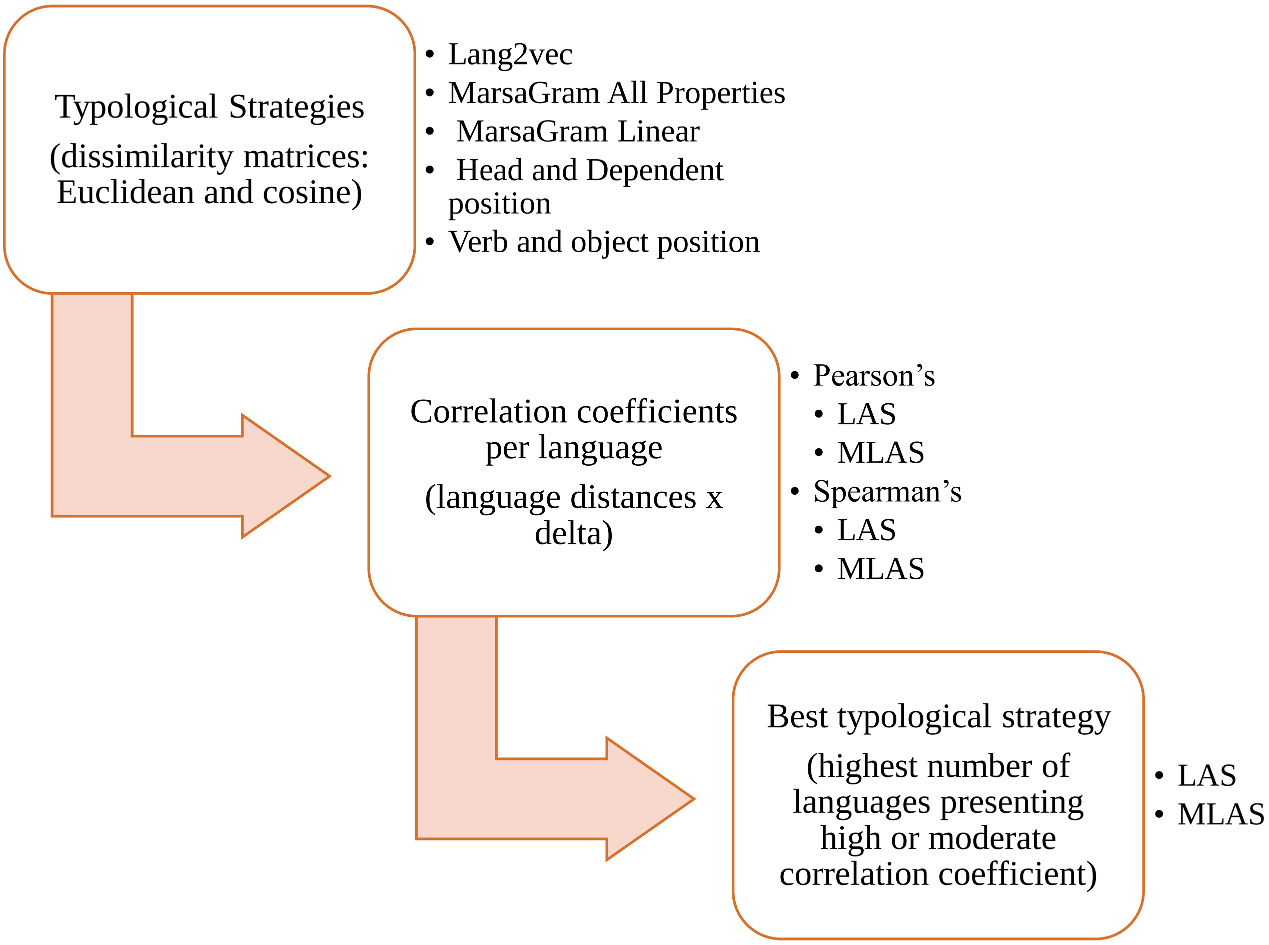
This first evaluation strategy is schematized in Figure 5.1.

This analysis enables us to define the best strategy of typological comparison which provides the most optimized classification of languages that is more probable to guarantee positive deltas (i.e.: the highest number of strong and moderate correlations). It also allows checking if the LAS and MLAS metrics correlate in a linear or monotonic way to the language distances.

The second evaluation strategy is aligned with the way typological approaches are used in dependency parsing studies. Usually, when languages are compared, the selected language pair to be associated corresponds to the one presenting the smallest distance between the two languages when compared to the other possible associations.

Thus, besides the overall correlation between the language distances and the deltas, we decided to check, for each typological strategy and each language, if the language pair with the smallest distance is the same as the one observed in the experiments involving language combination (with the highest delta), or at least if the proposed pair provides an empirical improvement statistically similar to the best delta. Alongside that, when the typological approaches did not propose the right pair, or one statistically similar, we analysed the number of times that they indicated, at least, a pair with a positive delta (statistically different from the baseline), and the number of cases that the proposed pair corresponded to a statistically negative delta (indicating a negative synergy between the languages).

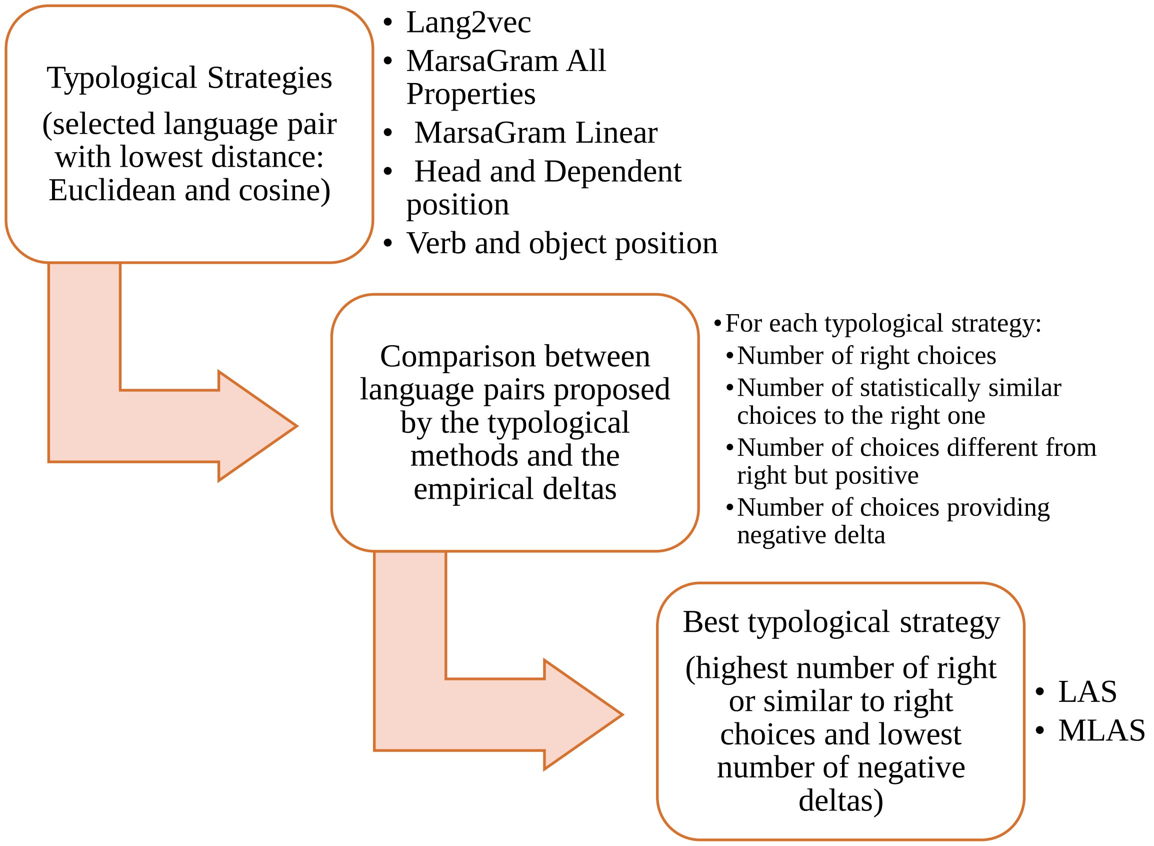
Figure 5.1. Schema representing the first evaluation strategy for the definition of the best typological approach for the improvement of LAS and MLAS by using corpora association.


Typological Strategies
(dissimilarity matrices: Euclidean and cosine)
• Lang2vec
• MarsaGram All Properties
• MarsaGram Linear
• Head and Dependent position
• Verb and object position
Correlation coefficients per language
(language distances x delta)
• Pearson's
• LAS
• MLAS
• Spearman's
• LAS
• MLAS
Best typological strategy
(highest number of languages presenting high or moderate correlation coefficient)
• LAS
• MLAS

Thus, the best typological strategy regarding this other type of evaluation is the one that predicts the highest number of right or similar language pairs with the minimum number of selected pairs conducting to negative deltas. The information concerning the number of selected pairs that provide positive deltas but that are statistically inferior to the best empirical one is only considered in cases where more than one typological strategy provides the same number of right and similar to right choices. This second strategy is schematized in Figure 5.2.

In both evaluation scenarios, the new typological strategies described in the previous chapter are compared to the language classification using lang2vec to check if the innovative quantitative approaches provide better results than the state-of-the-art one that uses typological information from databases.

Figure 5.2. Schema representing the second evaluation strategy for the definition of the best typological approach for the improvement of LAS and MLAS by using corpora association.


Typological Strategies
(selected language pair with lowest distance: Euclidean and cosine)
• Lang2vec
• MarsaGram All Properties
• MarsaGram Linear
• Head and Dependent position
• Verb and object position
Comparison between language pairs proposed by the typological methods and the empirical deltas
• For each typological strategy:
• Number of right choices
• Number of statistically similar choices to the right one
• Number of choices different from right but positive
• Number of choices providing negative delta
Best typological strategy
(highest number of right or similar to right choices and lowest number of negative deltas)
• LAS
• MLAS

5.3 Results and Analysis

The results are presented in this sub-section in the same order as the experiments that were described previously: establishment of the dependency parsing results baseline, language association experiments, and evaluation of typological strategies with respect to the empirical parsing results.

5.3.1 Definition of the baseline in terms of parsing results

As previously described, the baseline consists of the dependency parsing results (i.e.: mean and standard deviation of LAS and MLAS metrics) regarding the UDify models which were trained and tested with each PUD language alone. Tables 5.3 and 5.4 present in ascending order the obtained results (for LAS and MLAS respectively).

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 5.3. LAS results of the dependency parsing results obtained using UDify tool and PUD corpora. The colour scale indicates the differences in the order of magnitude of the results.

LAS results vary from 74.68 (for the Thai language) to 91.57 (for Japanese), a difference of around 17 points. On the other hand, the variation in terms of MLAS is much larger, around 34 points (the highest MLAS result was obtained for Japanese, and lowest for Icelandic).

As expected, the MLAS score is always lower than the LAS one. In most cases, the decrease is around 10 to 15 points, however, for Arabic, Czech, German, Hindi, Icelandic, Polish, and Turkish, the decrease is even higher than 20 (the maximum decrease was observed for Icelandic: 30.11).

When analysing the LAS table (5.3), it is noticeable that, 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 have quite similar results, however, Icelandic is positioned with the languages with the lowest scores (below 80) which are: Thai, Chinese, Turkish, and Hindi.

Language	MLAS	Std. Dev.
isl	48.79	0.52
hin	54.00	0.25
tur	56.02	0.38
arb	57.66	0.72
ces	57.83	0.48
pol	61.31	0.51
cmn	62.73	0.73
tha	63.85	0.00
deu	67.00	0.29
fin	68.26	0.17
swe	69.89	0.00
rus	70.47	0.20
eng	74.99	0.26
ita	76.27	0.35
ind	77.04	0.34
por	78.05	0.40
kor	78.23	0.13
fra	79.83	0.34
spa	79.84	0.19
jpn	82.90	0.36

Table 5.4. MLAS results of the dependency parsing results obtained using UDify tool and PUD corpora. The colour scale indicates the differences in the order of magnitude of the results.

When the MLAS metric was described, it was mentioned that this metric is more adequate when doing multilingual comparative analysis of parsing results. What is possible to observe in the MLAS table (5.4) is that Japanese and Romance languages are still the ones with higher scores, however, Korean and Indonesian also appear in the top positions (while concerning LAS, they had average scores). Concerning Slavic languages, Czech and Polish are the ones that have the highest decrease when comparing LAS and MLAS metrics, Russian is also impacted but in a lesser way. With respect to Germanic languages, Icelandic is the one with the lowest MLAS score of all PUD languages. German, which had a quite high value of LAS, has a much lower MLAS when compared to English. Arabic, Hindi, and Turkish present quite low scores of MLAS, as was the case for the LAS metric too. Moreover, the position of the Finnish language among PUD languages is better when MLAS scores are compared in contrast to when LAS metric was considered.

It is possible to compare these results, in terms of LAS and MLAS, with the ones published in the article which describes UDify tool (Kondratyuk and Straka, 2019). The authors tested their multilingual model trained with all Universal Dependencies v.2.3 corpora (i.e.: 124 languages) with a large collection of test corpora, including 18 PUD ones (the exceptions being Icelandic and Polish, for which there is no result available in the mentioned article). Thus, it is possible to see how well the monolingual UDify PUD models test against the state-of-the-art UDify multilingual one. The deltas between the LAS and MLAS values obtained via our monolingual models and the ones from the literature are displayed in Annex 25. In general, the monolingual models perform much better than the multilingual one: the increase in terms of LAS varies from 5.09 for German to 48.62 for Thai. It is also the case for MLAS, with even higher delta values (i.e.: from 12.77 to 69.63). For this specific metric, UDify multilingual model fails to provide satisfactory results in many cases (i.e.: below 20 for 9 PUD languages), while the monolingual models guarantee at least values higher than 48.

Nevertheless, there are a few cases where the multilingual UDify model provides better scores than the monolingual ones. For LAS, the standard model works slightly better for Czech, English, Finnish, Italian, Japanese, and Swedish (with a delta varying from 1.15 to 4.12 in favour of the multilingual model). On the other hand, regarding MLAS, the magnitude of the deltas is more important for a few languages: the multilingual model presents better results for Czech (19.56), English (0.62), Finnish (9.57), and Japanese (1.96). This considerable difference observed for Czech is probably due to the fact that in the multilingual model, there is considerably more morphological information in the training-set which is crucial for MLAS evaluation.

Thus, it is possible to observe that even though the 20 languages selected for this part of the thesis are tested in a low-resourced scenario (i.e.: 1,000 sentences per language), results are comparable and even better than the state-of-the-art parsing model. This shows how well deep-learning models based on language models work, and also the advantage that can be observed when the test set has the same tag-sets and is from the same genre as the training one. The diversity in terms of morphosyntactic annotations can be one of the reasons why MLAS metrics are quite low in many cases for the standard multilingual UDify model.

As detailed in the section regarding the methodological aspects, all models were trained with the same amount of data and with the same semantic content. However, the final scores do not rely exclusively on the training set size as it can be attested by the discrepancy in terms of LAS and MLAS results. Moreover, the obtained results cannot be explained by simple linguistic features such as VO versus OV, or agglutinative versus synthetic languages.

One possible factor that influences the overall LAS and MLAS results is the difference in terms of the size of the language representation in the language model which is part of the UDify architecture. It means that PUD languages are not equally treated when models are trained by this tool even though using the training sets with the same size. To analyse the extent of this influence, we calculated the Pearson and Spearman's correlation coefficients between the LAS and MLAS scores and the size of the training corpus used to compose mBERT (i.e.: mean value between the size range limits presented in Table 5.2). The obtained correlation coefficients are displayed in table 5.5.

	Pearson's	**Spearman's**
LAS – mBERT	0.33	0.72
MLAS - mBERT	0.33	0.53

Table 5.5. Correlation coefficients calculated between the dependency parsing scores and the mean size of the mBERT training corpora concerning PUD languages.

The correlation coefficients indicate that there is a strong correlation between the size of the language representation and the LAS scores, and a moderate one regarding the MLAS metric, meaning that the largest the language representation is inside the language model, it is more probable that the scores for this language are higher. The correlation is not linear (i.e.: Pearson's coefficients are lower than Spearman's ones), and is not total (e.g.: English has, by far, the largest mBERT training corpus, but does not have the highest LAS and MLAS scores). Nevertheless, it is possible to verify that Romance languages and Japanese have quite large mBERT representations and also good scores. Also, in terms of Germanic languages, Icelandic has the smallest mBERT training set and has the lowest scores when compared to the other ones from this linguistic genus.

To better understand what other factors may influence the LAS and MLAS metrics, we calculated the correlation coefficient between the scores and:

1) Number of UPOS tags;
2) Number of Dependency Relation tags;
3) Percentage of relations where the head follows the dependent;
4) Percentage of occurrences of objects before verbs.

The obtained correlation coefficients are displayed in table 5.6.

From the obtained correlation coefficients, it is possible to notice that all the listed variables do not seem to correlate with LAS scores (i.e.: coefficients close to 0). However, for MLAS, we observe that the number of dependency labels (with sub-types included) seem to have a low inverse correlation, almost moderate (i.e.: coefficients close to -0.5). Moreover, MLAS scores also present a low correlation with the left-branching character of the PUD languages (i.e.: it indicates that if a language has a stronger left-branching tendency, MLAS results tend to be better).

	Pearson's	**Spearman's**
LAS – UPOS	0.11	0.13
MLAS – UPOS	-0.21	-0.16
LAS – DepRel	-0.03	-0.01
MLAS – DepRel	-0.46	-0.43
LAS – % head after dependent	0.10	0.09
MLAS – % head after dependent	0.30	0.23
LAS – % O before V	-0.13	0.15
MLAS – O before V	0.03	0.25

Table 5.6. Correlation coefficients calculated between the dependency parsing scores and different features concerning PUD languages.

From these analyses, it seems that LAS scores are strongly positively impacted by the size of the language representations in mBERT. MLAS also is influenced by mBERT composition, however in more weakly way, thus, other features may impact, in different ways, the PUD ranking in terms of this metric, such as the left or the right-branching overall tendency of the languages and other linguistic characteristics that were not analysed here.

The establishment of the baseline results (with models trained with languages alone) was conducted to allow the evaluation of the results when different languages are combined which is the object of the following section.

5.3.2 Language combination experiments using UDify

As explained in the methodological sub-section, all 20 PUD languages were combined in pairs to compose bilingual training sets which were then used to train UDify models. In total, 380 experiments were conducted (19 per PUD language).

For each language, the obtained results regarding the combination experiments are presented in a table containing:

- LAS and MLAS of each language association experiment (i.e.: mean value of the 6 experiments varying random seed values);
- LAS and MLAS Standard Deviation of each language association experiment;

- LAS and MLAS Deltas (i.e.: the difference between the baseline LAS and MLAS scores concerning the language alone and the scores of the language association experiments);
- LAS and MLAS p-values calculated between the baseline results and the scores of each language combination experiment.

The whole ensemble of these detailed results is displayed in the Annex section (from Annex 26 to 45). With these tables, it is possible to identify for each language, the number of combinations for which the delta is statistically higher than 0 (positive synergy), and the number of language associations for which a negative synergy is observed (i.e.: delta is statistically lower than 0). Thus, the PUD languages can be analysed in relation to the amount statistically valid positive and negative deltas (i.e.: p-value lower than 0.01) as presented in the tables 5.7 and 5.8 for LAS and MLAS respectively.

	Positive LAS Deltas (p<0.01)	**Negative LAS Deltas (p<0.01)**
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 5.7. Number of positive (delta higher than 0) and negative (delta lower than 0) synergies concerning the LAS scores of the language combination experiments with the UDify tool. Languages are organized in ascendant order regarding the number of positive synergies. The colour scale indicates the order of magnitude of the number of positive deltas: red for the lowest values, and green for the highest numbers.

In terms of the LAS metric, it is possible to observe that the group of languages with more than 10 cases of language combination which conduct to 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, Portuguese, and French, all of them with less than 10 positive deltas. The Finnish language is the most favoured 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 of 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.

	Positive MLAS Deltas (p<0.01)	**Negative MLAS Deltas (p<0.01)**
jpn	0	12
kor	0	17
tha	0	7
cmn	1	0
ind	2	0
hin	4	0
por	4	0
fra	5	4
ces	7	1
eng	7	0
isl	7	0
swe	7	1
deu	8	2
ita	10	0
spa	10	0
pol	11	0
fin	16	0
rus	18	0
arb	19	0
tur	19	0

Table 5.8. Number of positive (delta higher than 0) and negative (delta lower than 0) synergies concerning the MLAS scores of the language combination experiments with the UDify tool. Languages are organized in ascendant order regarding the number of positive synergies. The colour scale indicates the order of magnitude of the number of positive deltas: red for the lowest values, and green for the highest numbers.

In terms of MLAS, Turkish and Arabic languages present a statistically valid positive delta when combined with every other PUD language. These two languages did not present the same tendency for positive synergy in terms of LAS. Russian also presents a high number of cases in which the delta is positive (18 out of 19), and so does Finnish, however, these two languages presented also good tendencies with LAS. Regarding Germanic languages, all of them present a number of language associations with positive deltas lower than 10. In the Romance language genus, as was the case for LAS, when MLAS is considered, French and Portuguese have a smaller number of positive delta when compared to Italian and Spanish. The French language has even a quite high number of cases for which the obtained MLAS delta is negative.

The language group with the highest number of negative MLAS delta is consistent with the one observed when LAS deltas were analysed: Japanese, Korean, and Thai. Moreover, Hindi, Chinese and Indonesian are the languages with lesser significant impact in terms of dependency parsing metrics when combined to other languages.

What is possible to notice is that for both LAS and MLAS metrics, Indo-European languages (except for Hindi) tend to be favoured when combined with other PUD languages. This is most probably due to the fact that the number of languages from this linguistic family is quite high (13 out of 20) with more than one language per genus regarding Romance, Slavic and Germanic ones. For non-Indo-European languages, Finnish is the one with the highest number of positive synergies when considering both LAS and MLAS. However, when only MLAS is analysed, Arabic and Turkish surpass Finnish in terms of positive deltas, the score being improved in every single experiment of language association for these two languages.

Besides this overall analysis in terms of the number of positive and negative deltas, it is possible to check, for each language, how large is the increase provided by the best language combination and the number of language pairs with a statistically similar result to the best LAS and MLAS values. This information is displayed in Tables 5.9 and 5.10 (concerning LAS and MLAS respectively).

From the results provided in table 5.9, it is possible to notice that the maximum increase in terms of LAS varies a lot when all PUD languages are compared. As it was observed before, LAS results cannot be improved for Hindi, Japanese and Korean when only PUD languages are combined. LAS for the Thai language is slightly improved (0.29), but this delta is much smaller than the ones obtained for Chinese (1.18), Czech (1.58), Finnish (1.95), Icelandic (1.57), Polish (1.66), Portuguese (1.35), Russian (1.19), Swedish (1.30) and Turkish (1.08).

Regarding Romance languages, the best delta is always obtained with a language from the same genus. However, not all Romance languages provide the same improvement magnitude when combined with each other (e.g.: Portuguese has the highest LAS increase with Spanish and French, but not with Italian). Italian is a particular case of Romance language as the LAS delta is maximum when it is combined with French but also with Russian, English, and Indonesian. The Indonesian language also forms optimized pairs with Chinese, Finnish, German, and Icelandic, although it does not belong to the same linguistic family of these languages. However, the Indonesian score is only improved with Hindi. Concerning Germanic languages, English is the best pair for Swedish and German, however, for English and Icelandic, French is the best associated language (from the same linguistic family but from a different genus). These inter-genera combinations are also observed in the Slavic languages: for Czech, Russian is the best pair, but for Russian, Icelandic (Germanic) is the best candidate, and for Polish, French (Romance).

	Best combination	**Delta LAS**	**Other statistically similar combinations**
arb	ita	0.57	spa
cmn	fra	1.18	ind, ita, rus
ces	rus	1.58	-
eng	fra	0.93	-
fin	rus	1.95	ind
fra	ita	0.68	por
deu	eng	0.39	ind, ita, kor, pol, por
hin	-	-	-
isl	fra	1.57	ind, por
ind	hin	0.41	-
ita	fra	0.78	eng, ind, rus
jpn	-	-	-
kor	-	-	-
pol	fra	1.66	-
por	spa	1.35	fra
rus	isl	1.19	-
spa	por	0.86	pol
swe	eng	1.30	fra, rus
tha	fra	0.29	-
tur	kor	1.08	-

Table 5.9. Language pairs that provide the best delta in terms of LAS and other languages that deliver statistically similar LAS when compared to the best one.

Korean LAS cannot be improved via the association of its training set to other PUD languages, however, this language provides the best combination for Turkish, and a statistically similar to the best combination for German. Moreover, Arabic seems more prone to be improved when associated with Romance languages (best results obtained when combined to Italian and Spanish).

Overall, it is possible to see that in 7 cases, the best pair is formed with languages belonging to the same genus. However, this criterion solely does not guarantee that the combination provides an optimized LAS result. It is interesting to see how distant languages in terms of phylogenetic classification can provide good synergy in terms of parsing results in some scenarios. It is the case of Indonesian combined with Chinese, German, Finnish, Icelandic, and Italian; German combined with Korean; and Arabic combined with Spanish and Italian. These results confirm the interest in going beyond the classic genealogical classification to find the most optimized language pairs for dependency parsing improvement.

Moreover, in the lang2vec clustering analysis, Arabic and Thai were classed with Romance languages which could explain the improvement they show when combined with some languages of this genus. Indonesian was also part of the same cluster; however, it provides considerable LAS improvement for languages belonging to a different language group (as presented in table 5.9). Thus, lang2vec analysis does not seem to explain all the LAS results concerning the language association experiments.

If we consider Hawkins' (1983) language types together with the lang2vec classification, Turkish, Hindi, Japanese and Korean belong to the same group. However, the only improvement observed was when Turkish was combined with Korean. Furthermore, Arabic is not from the same language type of Romance languages, according to Hawkins, but it can be improved when associated to languages from this genus. Hawkins' classification does not explain either other possible combinations such as the inter-genera ones (e.g.: Icelandic and French).

It is possible to observe in table 5.10, that the best language pairs concerning MLAS do not exactly match the ones for LAS. In terms of the best combination, the choice is the same for Arabic, Czech, English, French, German, Indonesian, Polish, Portuguese, Spanish, and Turkish. However, the number and the possible similar other choices vary considerably. Indonesian is no longer a good language to be combined with some PUD languages as it was the case for LAS metric (only Swedish presents an improvement when associated with Indonesian).

It is also noticeable that the MLAS improvements are higher than the LAS ones (i.e.: minimum of 0.23 for Italian, and maximum of 4.11 for Russian).

	Best combination	**Delta MLAS**	**Other statistically similar combinations**
arb	ita	2.91	deu, eng, fra, jpn, pol, por, rus, spa
cmn	-	-	-
ces	rus	2.03	pol
eng	fra	1.73	-
fin	ces	2.47	pol
fra	ita	1.39	deu, por, spa
deu	eng	1.41	fra, ita, por
hin	swe	0.96	-
isl	ita	1.92	deu, pol, rus
ind	hin	0.70	ita
ita	pol	0.23	ces, eng, fra, por
jpn	-	-	-
kor	-	-	-
pol	fra	2.24	-
por	spa	2.29	-
rus	fra	4.11	ita, pol, por, spa
spa	por	1.40	fra, pol
swe	fra	1.61	eng, ind, rus
tha	-	-	-
tur	kor	2.65	deu, fin, pol, rus, swe

Table 5.10. Language pairs that provide the best delta in terms of MLAS and other languages providing statistically similar scores when compared to the best one.

Moreover, as was the case for LAS metric, Romance languages, in general, have their best scores when combined with other languages from this genus. However, for French, a similar score is obtained when it is combined to German. Considering Italian, its MLAS improvement is quite low and is obtained when it is combined with Polish, however, a statistically similar score is obtained with the other Romance languages (Spanish being the exception), and with Czech and English.

In terms of MLAS, Arabic has also a strong affinity with Romance languages, however some Germanic and Slavic languages also improve its score considerably, as well as Japanese (this being the only case where this isolated language can bring significant improvement in our experiments).

Again, Japanese, Korean, and Thai did not show any improvement when combined with other PUD languages. However, while Hindi LAS value was not optimized, its MLAS is slightly improved with Swedish. On the other hand, Chinese LAS was increased in several combination pairs, but its MLAS score was never enhanced considerably.

As it was the case for LAS, Russian is the best language to be associated with Czech, however, in terms of MLAS, the Czech score is also increased when it is associated with Polish (another Slavic language). The Russian LAS score was better improved with Icelandic, but when MLAS is considered, this language shows a higher positive synergy with Romance languages and Polish. The MLAS of this latter is better improved when associated with French.

When Germanic languages are analysed, the best improvement for English is only obtained with French, while for German, other Romance languages, as well as English, also provide statistically high increases. On the other hand, Icelandic shows good improvement with German but also with Italian and two Slavic languages (Russian and Polish). Swedish best MLAS concerns its association with French, but also with English, Indonesian, and Russian.

Turkish showed an improvement in terms of LAS when combined with Korean, and it is also the case for MLAS. However, this score can also be significantly improved with languages which are quite diverse concerning their word order strategies such as: German, Finnish, Polish, Russian, and Swedish.

Again, we observe some interesting positive synergy obtained when languages from different genera or even families are associated. Moreover, simple characteristics such as verb and object order cannot explain the complexity of the observed language pairs providing best parsing improvements. Furthermore, it is noticeable that the languages which provide the best improvements in terms of LAS and MLAS have a mBERT training corpus with a size higher than 0.354 GB. The only exception is the LAS increase provided by Hindi when associated with Indonesian.

Moreover, it is possible to notice that when languages are combined the synergy observed do not correspond to a bi-directional property. In many cases, one language can provide a positive delta to a second one, but the inverse is not observed (e.g.: a positive LAS and MLAS delta is observed when Indonesian is combined with Hindi, however, Hindi does not present a significant improvement when combined with Indonesian).

In relation to other typological studies regarding dependency parsing improvement, it is not possible to conduct a direct comparison between the deltas obtained in this study and the ones published in the literature due to the fact that these experiments usually consider different sets of languages, and use different parsing methods trained with delexicalized corpora. Moreover, most studies present strategies to parse unknown languages (without any annotated data), thus, with much lower parsing scores. In these works, the typological approach is used to determine the most similar language to parse an unknown one, and the results are, in most cases expressed in terms of Unlabelled Attachment Score (UAS) with quite low baseline values, which allows the obtention of high deltas (e.g.: UAS delta of 13.58 in the study developed by Agić, 2018).

However, it is possible to notice that the LAS deltas obtained in this thesis are coherent with the ones from the literature regarding experiments slightly similar to ours. For example, in the study conducted by de Lhoneux et al. (2018), the LAS deltas varied from 0.16 to 0.86 (with p-values inferior to 0.01), while in our study, delta LAS fluctuates from 0.23 to 1.95.

The next sub-section is dedicated to a detailed analysis in terms of correlations and prediction of best language pairs to identify which typological strategy is the most suitable for dependency parsing improvement. It also aims to identify which type of word order patterns play a major role when languages are associated to train deep-learning systems.

5.3.3 Typological strategies evaluation in relation to parsing results

In the previous section, a brief analysis was conducted regarding the results of language combination experiments for dependency parsing improvement. It was possible to confirm that this type of approach is capable of increasing evaluation metrics, especially in terms of MLAS. However, it was also noticeable that the best choice, in many cases, is not linked to phylogenetic factors, as was already seen in the literature such as in the study conducted by Lynn et al. (2014).

Thus, it is pertinent to provide a more precise evaluation of the typological approaches proposed as a response to the first hypothesis of this thesis in relation with the empirical results of the language association experiments.

As detailed in the methodological sub-section, the idea is to provide two types of evaluation:

a) calculation of the overall correlation between the language distances provided by each method and the obtained deltas in terms of LAS and MLAS.
b) identification of the best typological approach which provides the highest number of right choices in terms of positive delta.

The first evaluation strategy concerns the overall tendency of each method of providing consistent information in terms of language distance with respect to the improvement of dependency parsing metrics. Thus, the types of word order patterns, which were used to establish the typological approaches with the highest values of correlation, are the most relevant ones for deep learning systems regarding language association strategies.

The second evaluation approach is intended to assess the results more pragmatically. It is based on how different typological strategies are evaluated in state-of-art studies regarding parsing improvement: the best methods provide the highest number of right choices in terms of increase of LAS or MLAS.

Each evaluation procedure is described separately in the next sub-sections, followed by a comparative analysis of the selected strategies.

a) Evaluation in terms of overall correlation

For each evaluation metric (i.e.: LAS and MLAS), we calculated the correlation coefficients (both Pearson's and Spearman's) for each PUD language (i.e.: target language, monolingual test-set) regarding the distances obtained via each typological strategy and the empirical deltas. Then, we checked, for each approach, the total number of strong and moderate inverse correlations (i.e.: lower than -0.70, and between -0.70 and -0.50).

Concerning LAS, the ensemble of results regarding Pearson's and Spearman's coefficients are presented in the tables 5.11 and 5.12 respectively.

It is possible to notice, in both tables, that some languages do not present strong or moderate correlation coefficients for any typological approach. It is the case of Chinese, Hindi, Indonesian, Italian, Japanese, Korean, and Russian. For German, when Pearson's coefficients are considered, no typological approach can be identified providing moderate or strong correlation, however, with Spearman's coefficients, one single approach presents a moderate correlation: MarsaGram all properties with Euclidean distances. For Russian, only the strategy concerning the relative position of verb and object (Euclidean) presents at least a low correlation with coefficient values lower than -0.40 for both Pearson's and Spearman's.

	MarsaGram All		MarsaGram Linear		Head/Dependent		VO_OV		Lang2vec	
	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 5.11. Pearson's correlation coefficient for each PUD language and each typological strategy calculated between the language distances and the LAS deltas. Values in green indicate a strong correlation, and in yellow, a moderate one. Correlation coefficients between -0.50 and -0.40 are presented in red.

By comparing these coefficients and the LAS delta results, it is possible to notice that the 3 languages for which no statistically significant improvement was achieved (i.e.: Hindi, Japanese, and Korean) do not present strong or moderate correlation metrics. From them, Japanese shows, at least, a Pearson's correlation value of -0.40 for both MarsaGram all properties and MarsaGram linear (with Euclidean distances), however, for the other methods, the coefficients are closer to 0.00, or even positive. Korean is the most extreme language as all coefficients are positive which indicates that, in these cases, a higher distance between Korean and another language provides, most probably, a better LAS value regarding their combination. Hindi presents a negative Pearson's coefficient only for MarsaGram all properties (Euclidean and cosine), and negative Spearman's values also for the same typological approaches, and a slight negative result for the head and dependent (Euclidean) strategy, but closer to 0 when compared to the values of MarsaGram.

	MarsaGram All		MarsaGram Linear		Head/Dependent		VO_OV		Lang2vec	
	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 5.12. Spearman’s correlation coefficient for each PUD language and each typological strategy calculated between the language distances and the LAS deltas. Values in green indicate a strong correlation, and in yellow, a moderate one. Correlation coefficients between -0.50 and -0.40 are presented in red.

On the other hand, Indonesian and Italian show a larger number of negative correlation coefficients, although always higher than -0.50, which could explain why these languages present at least some LAS improvement in the experiments regarding language association.

Table 5.13 presents, for each typological strategy, the compilation of the results presented in tables 5.11 and 5.12 in terms of the number of languages that present strong or, at least, moderate correlations.

From the results displayed in table 5.13, 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 means that the linear order of components inside a same subtree seems a relevant factor which affects deep-learning systems.

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 to other new methods such as Verb and Object (cosine) and MarsaGram linear (Euclidean).

		MarsaGram All		**MarsaGram Linear**		**Head Dependent**		**VO_OV**		**Lang2vec**	
		Euc.	**cos**	**Euc.**	**cos**	**Euc.**	**cos**	**Euc.**	**cos**	**Euc.**	**cos**
Pearson	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
Spearman	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 5.13. Analysis of the number of PUD languages for each typological method presenting moderate or strong correlation coefficients (for both Pearson's and Spearman's) for LAS. In green is highlighted the highest total value.

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 inferior to -0.2, which indicates, at least, a low correlation.
- For Italian, the Pearson's correlation coefficient is lower than -0.10 but do not reach -0.20.
- 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.

Thus, the most problematic languages in this scenario are Chinese, Japanese, German, Russian, Korean, and Hindi.

From the 10 languages for which the MarsaGram linear (cosine) strategy does not present a moderate or strong correlation, only Korean has positive correlation values for all of the other typological strategies. For Japanese, MarsaGram linear (cosine) has the lowest Spearman's coefficient (-0.26), better than all other strategies. For the other languages, some other methods seem more relevant than the MarsaGram linear (cosine) one:

- German: MarsaGram all properties (Euclidean), with a Spearman's coefficient of -0.52.
- Russian: Verb and Object relative position strategy, with the same Pearson's and Spearman's correlation coefficient of -0.46.

- Chinese: MarsaGram linear (Euclidean), Pearson's and Spearman's coefficients of -0.41.
- Icelandic: Verb and Object relative position, with Spearman's coefficient of -0.50 (and Pearson's of -0.49).
- Indonesian: MarsaGram all properties (cosine), with Pearson's correlation of -0.30 (and -0.27 for Spearman's).
- Turkish: Head and dependent position strategy (cosine) with a Spearman's correlation of -0.69.
- Italian: MarsaGram all properties (cosine), with a Spearman's coefficient of -0.31, as well as Verb and Object position strategy (Euclidean) with a Pearson's coefficient of -0.30.
- Hindi: MarsaGram all properties (Euclidean and cosine) with Pearson and Spearman's coefficients from -0.36 to -0.23.

From these results, it seems that, for this group of 10 problematic languages, MarsaGram all properties strategy provides some valuable information for 4 of them. Thus, for these languages, the linear order inside the subtrees is not enough to explain the synergy observed when languages are combined. Instead, the other information provided by MarsaGram also helps (i.e.: exclude, unicity, and require).

For Russian and Icelandic, it seems that the relative order of verbs and objects is the most relevant factor playing a role when these languages are combined with others. More precisely, for Icelandic, the same correlation can be found with lang2vec classification.

Regarding MLAS metric, the correlation analyses are displayed in tables 5.14 and 5.15 (regarding Pearson's and Spearman's coefficients respectively).

Even though some similitudes can be observed between the MLAS correlation values and the LAS ones, there are some important differences, specially concerning the languages which do not present any moderate or strong correlation for LAS. The list of these languages regarding MLAS is smaller, especially regarding Spearman's coefficient.

In terms of problematic languages, Chinese, Indonesian, Japanese, and Korean were identified as such when LAS correlations were analysed, and the same situation occurs for MLAS (i.e.: no case of moderate or strong correlation). However, it is not the case for Hindi, Italian, and Russian. These languages present at least one case (i.e.: typological strategy) for which the correlation is at least moderate. Moreover, Icelandic is an exception in the PUD collection as it shows a moderate correlation regarding the verb and object relative position strategy for LAS deltas, but no typological strategy provides enough correlation regarding MLAS.

	MarsaGram All		**MarsaGram Linear**		**Head/Dependent**		**VO_OV**		**Lang2vec**	
	Euc.	**cos**	**Euc.**	**cos**	**Euc.**	**cos**	**Euc.**	**cos**	**Euc.**	**cos**
arb	0.09	-0.53	-0.18	-0.56	-0.23	-0.43	-0.63	-0.34	-0.45	-0.42
cmn	0.29	-0.19	-0.26	-0.10	0.55	0.38	0.19	0.03	-0.08	-0.09
ces	-0.18	-0.47	-0.27	-0.49	-0.29	-0.30	-0.30	-0.06	-0.40	-0.37
eng	-0.22	-0.50	-0.39	-0.66	-0.41	-0.50	-0.28	-0.15	-0.39	-0.44
fin	0.16	-0.35	-0.17	-0.37	-0.37	-0.42	-0.41	-0.32	-0.43	-0.38
fra	-0.35	-0.59	-0.52	-0.64	-0.62	-0.58	-0.33	-0.19	-0.40	-0.41
deu	-0.62	-0.53	-0.55	-0.61	-0.60	-0.64	-0.20	0.26	-0.20	-0.20
hin	-0.47	-0.54	-0.06	0.28	-0.06	-0.04	0.16	0.20	0.19	0.13
isl	0.30	-0.14	-0.09	-0.21	-0.13	-0.18	-0.01	-0.03	-0.25	-0.24
ind	0.31	-0.32	-0.06	-0.34	0.20	0.02	0.05	-0.02	-0.27	-0.20
ita	-0.26	-0.39	-0.37	-0.42	-0.40	-0.36	-0.32	-0.01	-0.39	-0.39
jpn	-0.07	0.11	-0.28	-0.31	0.13	0.06	-0.15	-0.15	0.06	0.17
kor	0.30	0.29	0.08	0.37	0.43	0.49	0.40	0.47	0.43	0.37
pol	-0.14	-0.55	-0.52	-0.67	-0.01	-0.30	-0.52	-0.21	-0.33	-0.34
por	-0.81	-0.77	-0.58	-0.82	-0.82	-0.74	-0.48	-0.48	-0.73	-0.66
rus	0.12	-0.05	0.08	-0.02	-0.04	-0.08	-0.36	-0.24	0.05	0.05
spa	-0.47	-0.51	-0.52	-0.53	-0.52	-0.41	-0.48	-0.21	-0.54	-0.49
swe	-0.40	-0.33	-0.30	-0.50	-0.46	-0.46	-0.53	-0.44	-0.48	-0.48
tha	0.37	-0.60	-0.21	-0.72	-0.58	-0.82	-0.45	-0.72	-0.65	-0.70
tur	0.06	-0.16	-0.32	-0.04	-0.37	-0.64	-0.40	-0.35	-0.52	-0.58

Table 5.14. Pearson's correlation coefficient for each PUD language and each typological strategy calculated between the language distances and the MLAS deltas. Values in green indicate a strong correlation, and in yellow, a moderate one. Correlation coefficients between -0.50 and -0.40 are presented in red.

Finnish, Italian, and Russian only present correlation values lower than -0.5 when Spearman's coefficient is considered. Moreover, for Hindi, while the LAS delta results could not be explained with any typological approach, regarding MLAS, it is moderately correlated with MarsaGram all properties (cosine) for both type of correlation coefficients. MLAS metric is also much better correlated to language distances when German results are considered, while for LAS only one strategy provided moderate correlation, for MLAS, it is the case for 3 different methods (with both Euclidean and cosine distance metrics).

	MarsaGram All		MarsaGram Linear		Head/Dependent		VO_OV		Lang2vec	
	Euc.	cos	Euc.	cos	Euc.	cos	Euc.	Cos	Euc.	Cos
arb	0,01	-0,56	-0,29	-0,52	-0,15	-0,25	-0,57	-0,31	-0,50	-0,51
cmn	0,46	-0,26	0,14	-0,08	0,62	0,36	0,22	0,01	-0,13	-0,12
ces	-0,06	-0,66	-0,47	-0,57	-0,25	-0,30	-0,47	-0,35	-0,31	-0,29
eng	-0,28	-0,67	-0,51	-0,69	-0,61	-0,66	-0,31	0,06	-0,55	-0,56
fin	0,23	-0,29	0,00	-0,31	-0,34	-0,49	-0,54	-0,53	-0,41	-0,40
fra	-0,31	-0,56	-0,57	-0,62	-0,55	-0,57	-0,46	-0,28	-0,37	-0,40
deu	-0,63	-0,60	-0,57	-0,61	-0,71	-0,69	-0,14	0,17	-0,29	-0,30
hin	-0,32	-0,56	0,10	0,21	-0,08	-0,11	0,19	0,28	0,16	0,16
isl	0,39	-0,23	-0,11	-0,21	-0,16	-0,32	-0,05	0,00	-0,32	-0,35
ind	0,23	-0,27	-0,09	-0,31	0,16	0,06	0,01	-0,20	-0,26	-0,24
ita	-0,25	-0,48	-0,37	-0,43	-0,49	-0,40	-0,52	-0,01	-0,47	-0,49
jpn	0,00	0,20	-0,11	-0,44	0,28	0,36	-0,08	-0,08	0,33	0,34
kor	0,41	0,35	0,33	0,01	0,58	0,18	-0,11	0,03	0,10	0,14
pol	-0,07	-0,74	-0,68	-0,74	-0,05	-0,40	-0,62	-0,24	-0,41	-0,43
por	-0,77	-0,74	-0,70	-0,82	-0,82	-0,84	-0,58	-0,54	-0,65	-0,67
rus	-0,05	-0,39	-0,25	-0,42	-0,22	-0,26	-0,74	-0,63	-0,17	-0,16
spa	-0,34	-0,46	-0,44	-0,49	-0,48	-0,41	-0,56	-0,14	-0,58	-0,58
swe	-0,45	-0,39	-0,39	-0,52	-0,53	-0,46	-0,64	-0,27	-0,31	-0,25
tha	0,32	-0,41	-0,22	-0,69	-0,47	-0,77	-0,21	-0,61	-0,54	-0,57
tur	0,11	-0,30	-0,24	0,06	-0,41	-0,63	-0,23	-0,46	-0,48	-0,48

Table 5.15. Spearman's correlation coefficient for each PUD language and each typological strategy calculated between the language distances and the MLAS deltas. Values in green indicate a strong correlation, and in yellow, a moderate one. Correlation coefficients between -0.50 and -0.40 are presented in red.

From this, it is possible to conclude that there is a tendency for MLAS deltas to correlate better with the proposed typological strategies when compared to LAS. In terms of the overall number of languages presenting a relevant correlation, the results are displayed in table 5.16.

Again, the best typological strategy that can be established with this evaluation method is the MarsaGram linear with cosine distances (i.e.: moderate or strong correlation for 9 PUD languages for in terms of both coefficient). Moreover, regarding MLAS, MarsaGram all properties (cosine) is also a valid strategy, with the same score (i.e.: 9 languages, considering Pearson's coefficient). However, MarsaGram linear has the advantage of presenting strong correlation for 2 languages for both coefficients, while for MarsaGram all properties, it is only the case for Spearman's.

		MarsaGram All		**MarsaGram Linear**		**Head Dependent**		**VO_OV**		**Lang2vec**	
		Euc.	**cos**	**Euc.**	**cos**	**Euc.**	**cos**	**Euc.**	**cos**	**Euc.**	**Cos**
Pearson	Strong	1	1	0	2	1	2	0	1	1	1
	Moderate	1	8	5	7	4	4	3	0	3	2
	Total	2	9	5	9	5	6	3	1	4	3
Spearman	Strong	1	2	1	2	2	2	1	0	0	0
	Moderate	1	6	4	7	3	4	7	4	5	5
	Total	2	8	5	9	5	6	8	4	5	5

Table 5.16. Analysis of the number of PUD languages for each typological method presenting moderate or strong correlation coefficients (for both Pearson's and Spearman's) for MLAS. In green is highlighted the highest total value.

When compared the languages covered by these 2 methods mentioned above, it is possible to notice that, in terms of Pearson's correlation, MarsaGram all is better for Hindi, but does not work for Swedish (while MarsaGram linear covers this language). The situation is similar when Spearman's coefficients are considered, however, MarsaGram all properties method do not present a moderate correlation for both Swedish and Thai, while the linear strategy does.

Thus, as the same strategy was identified for both LAS and MLAS, it corroborates that the linear order inside subtrees is a relevant factor for language association regarding dependency parsing. This specific quantitative word order analysis is more pertinent than the classic one based on generic syntactic features (i.e.: lang2vec).

In the case of the LAS correlation analysis, MarsaGram linear (cosine) explains the delta results for half of the 20 PUD languages. However, when MLAS are considered, 11 languages can be identified as problematic for this typological strategy:

- For Italian, Japanese, Russian, and Spanish, the Spearman's correlation obtained for this method is lower than -0.40, showing at least some low inverse correlation, almost moderate. While for Japanese, this is the best identified strategy, for the other 3 languages, the correlation is moderate for the Verb and Object strategy (i.e.: lower than -0.5), with Russian presenting a strong Spearman's correlation when Euclidean distance is considered.

- 2 languages have at least a negative correlation with a value between -0.30 and -0.40: Finnish and Indonesian. Finnish presents a moderate correlation (Spearman's) for the Verb and Object position strategy, however, for Indonesian, no other strategy is better.
- Icelandic has correlation values of -0.21 and -0.22 (Pearson's and Spearman's respectively). This language shows slightly better results of Spearman's correlation regarding Head and Dependent position (cosine) and for lang2vec strategy (both Euclidean and cosine).
- Chinese has a Spearman's coefficient of -0.08 which is close to the Pearson's one (i.e.: -0.10). When MarsaGram linear properties are considered with Euclidean distances, the Pearson's correlation result is better (-0.26), with the same value obtained when MarsaGram all properties (cosine) strategy is considered with Spearman's correlation.
- For Turkish, both coefficients regarding this strategy are close to 0.00. On the other hand, the Head and Dependent (cosine) strategy presents moderate correlation (for both coefficients).
- For Korean, the Spearman's coefficient is close to 0, but in terms of Pearson's correlation, the value is positive and quite high (0.37). All other coefficients regarding the other strategies are positive, except for the Verb and Object (Euclidean) strategy which has a negative value but relatively close to 0 (i.e.: -0.11)
- Regarding Hindi, both coefficients are positive (above 0.20). However, when all properties extracted with MarsaGram are considered (cosine), there is a moderate correlation (for both coefficients).

Consequently, MarsaGram all properties strategies showed some relevance for LAS; however, it is only pertinent for Hindi in terms of MLAS. The Head and Dependent position strategy was not convenient for any language regarding LAS, but is pertinent for Icelandic and Turkish when MLAS is involved.

In conclusion, MarsaGram linear strategy with cosine distances can be considered the most appropriate typological method in terms of correlation with LAS and MLAS empirical results. It does not explain all the obtained results but is consistent[38] for at least 16 languages in terms of MLAS (the exceptions are: Chinese, Hindi, Korean, and Turkish), and 13 languages in terms of LAS (the exceptions are: Chinese, German, Hindi, Italian, Japanese, Korean, Russian, and Turkish).

---

[38] Being consistent means that the method provides at least negative coefficients lower than -0.2.

It is interesting to notice that the OV PUD languages (which form a large group in the lang2vec cluster analysis) are the ones with lesser results in terms of correlation regarding the syntactic aspects of the different typological approaches. From these 4 languages, Japanese and Korean are the ones which do not benefit at all from the association with other PUD languages (Hindi showing at least some improvement in terms of MLAS, and Turkish having significantly MLAS increase with all PUD languages). It is important to mention that one of the reasons for these results is the fact that the PUD language set is imbalanced. It contains several Slavic languages, several Romance and several Germanic, which increases the chance that a typologically close language will be found. On the other hand, it is not the case for the non-Indo-European PUD languages. Japanese and Korean are isolates, and Hindi might have presented better results if Punjabi or Nepali were also considered.

According to URIEL database of syntactic features, Turkish differs from the other 3 OV languages as it has a value of 0.33 for features such as S_SVO, S_VSO, and S_OBJECT_AFTER_VERB, meaning that the verb can sometimes precede the object. However, this was not observed in the PUD corpora regarding the overall VO/OV tendency (Table 4.30). For Hindi, the main syntactic difference in comparison to the other OV languages is the fact that the subordinator word precedes the clause in this language, while it follows the clause in the others. Therefore, it does not seem that these small factors have such a high influence on the overall results of the language combination experiments.

Hindi and Turkish just present positive deltas in terms of MLAS, and this metric considers not only dependency relations but also morphosyntactic labels. Thus, it is possible that the improvement in terms of this score also comes from a better annotation in terms of UPOS and FEATS.

Concerning LAS, Turkish presents a moderate correlation for the head and dependent strategy, the verb and object, and for the lang2vec one, and Japanese for MarsaGram all properties and linear. Although showing that some possible correlations can be found, it is not as tangible as for Indo-European languages. One possible explanation is that these non-Indo-European languages are tested as isolated ones in this ensemble of corpora, and they differ considerably from all the other PUD languages. Some contrary examples are Arabic, Finnish, and Thai which are not non-Indo-European, but present more syntactic similarity with some PUD languages, thus, attesting a moderate or strong correlation for many typological strategies. Thus, it seems that the MarsaGram linear (cosine) strategy is more suitable for VO languages.

However, being a VO language does not guarantee a moderate correlation in terms of LAS scores as was observed for Italian, and Russian. One particularity of the Italian language in comparison to other Romance ones is the fact that this language has a value of 0.33 for URIEL features concerning OV word order (e.g.: OBJECT_BEFORE_VERB), while the other languages have 0. However, in the analysis of the quantitative tendency of objects preceding verbs, this difference is not observed. Russian differs from other PUD Slavic languages in terms of S_OVS feature (0.00 for Russian, and 0.33 for Czech and Polish), but is similar to Czech for other syntactic features concerning subject, verb, and object positions. Polish presents more specificities in terms of syntax inside the Slavic group. Thus, it does not seem that the lack of LAS correlation for Russian is linked to a specific word order feature of this language.

Furthermore, German presented a moderate correlation for only one strategy (MarsaGram all properties). This language presents a unique distribution of verbs and objects (as seen in Figure 4.31), and is the only PUD language that is classified by WALS as “no dominant order”. Thus, languages that present some variation in terms of verb and object position (e.g.: Italian and, in a more evident case, German) should be considered with special attention as the optimized method may not guarantee the best results in terms of LAS improvement.

One question that may be raised is how well these typological methods would work in different architectures of dependency parsing tools. As has been explained previously, this thesis aims to analyse the influence of syntactic features on deep-learning tools that are based on language models. However, some experiments using typological strategies concerning unknown languages use simpler parsing tools without the usage of any language model (e.g.: the study conducted by Litschko et al. In 2020).

To answer that, we decided to conduct the same language association experiment using all PUD languages using UDPipe 1.0 tool (Straka et al., 2016). The dependency parsing module of UDPipe 1.0 is a transition-based parser capable of analysing both projective and non-projective sentences and is inspired by the work of Chen and Manning (2014). It uses a neural network classifier for prediction without another feature processing and with a dynamic oracle. The UDPipe 1.0 has been replaced by its second version which includes a language model in its architecture providing better parsing results (Straka, 2018), however, since the objective here is to test with a different system to UDify, we decided to conduct this set of experiments with its first version.

We followed the identical steps and used the same corpora as defined for the UDify experiments: a) training of monolingual models, b) training of models concerning each possible PUD language association in pairs, c) evaluation of the correlation between LAS and MLAS delta and language distances provided by each typological method.

The training of UDPipe models was conducted with the following parameters:

- parser=iterations=20;
- parser=embedding_form_file.

The other parameters were the standard ones as provided by the creators of this tool. In terms of statistic validation, as it is not possible to provide different values of random seed, we proceeded with 4 different values of batch size: 5, 10, 15, and 20.

The overall results for LAS and MLAS and the delta between UDPipe scores and UDify ones are presented in Annex 46. As expected, UDify scores are consistently better than the ones obtained with UDPipe (i.e.: over 15 points regarding LAS for all languages with exception of Japanese, and over 20 points for 16 languages in terms of MLAS, again Japanese has the smallest delta, 11.69). All the detailed results per language are presented in Annexes 47 to 66.

In terms of language association when using UDPipe, Arabic, German, Finnish, French, Italian, Japanese, Korean, Spanish, Swedish, and Turkish did not present any significant improvement for both metrics with models trained with combined corpora. This alone shows that for this specific architecture, language association does not provide the same benefit, in terms of language coverage, regarding parsing results as it does with UDify. However, when positive synergy is observed, it is possible to notice that the obtained deltas are higher than the ones from UDify. Furthermore, even when deltas are positive, the final scores are still inferior to the baseline ones of UDify.

When analysing the correlation coefficients, it is clear that there is no connection between the selected syntactic features and the obtained results. The state-of-the-art typological strategy (i.e.: language vectors with lang2vec features) did not provide a significant moderate or strong correlation either. In table 5.17, we present the Pearson's correlation coefficients for each PUD language regarding UDPipe LAS results. The other tables concerning the Spearman's coefficient and MLAS metric are displayed in Annex 67 to 69, and they follow the same pattern as the one presented here.

The only language presenting consistent moderate or strong correlation for Head and Dependent, Verb and Object position, and lang2vec strategies is Arabic. For this specific language, all the combinations provide negative LAS and MLAS delta, thus, the correlation indicates that for the most similar languages, the negative synergy is attenuated.

Therefore, it is possible to conclude that when languages are combined (with lexicalized corpora), the quantitative syntactic phenomena extracted by each typological method described in this thesis play a more important role in parsing tools with architectures similar to UDify (i.e.: tools developed with language models, which are the state-of-the-art in terms of parsing technologies). Completely different phenomena have more influence on the language association results for software such as UDPipe.

	MarsaGram All		**MarsaGram Linear**		**Head/Dependent**		**VO_OV**		**Lang2vec**	
	Euc.	**cos**	**Euc.**	**cos**	**Euc.**	**cos**	**Euc.**	**cos**	**Euc.**	**cos**
arb	0.02	-0.44	-0.29	-0.66	-0.81	-0.85	-0.73	-0.75	-0.85	-0.87
cmn	-0.14	-0.30	0.04	-0.23	0.21	0.17	0.08	0.17	-0.01	0.05
ces	-0.08	0.15	0.09	0.16	-0.16	-0.02	0.28	-0.05	0.05	0.12
eng	0.38	0.24	0.46	0.36	0.32	0.29	0.00	0.04	0.21	0.21
fin	0.20	0.32	0.08	0.05	0.07	0.14	0.21	0.14	-0.06	-0.03
fra	0.11	0.14	0.13	0.05	0.09	0.09	-0.07	-0.12	-0.04	-0.03
deu	0.73	0.61	0.29	0.33	0.58	0.57	0.06	-0.14	0.03	0.08
hin	0.49	0.16	0.23	0.34	-0.17	-0.19	-0.17	-0.04	-0.33	-0.45
isl	-0.04	-0.11	-0.38	-0.38	-0.47	-0.44	-0.21	-0.37	-0.48	-0.46
ind	0.16	0.13	-0.27	-0.11	-0.13	-0.21	-0.12	-0.29	-0.23	-0.24
ita	0.58	0.49	0.62	0.52	0.53	0.47	0.13	0.25	0.39	0.41
jpn	-0.25	-0.09	0.05	0.01	0.27	0.22	0.04	0.02	0.07	0.15
kor	0.30	0.39	0.06	0.21	0.26	0.08	-0.45	-0.12	-0.05	-0.13
pol	0.04	0.02	0.01	0.00	0.00	0.02	0.24	0.00	0.05	0.06
por	-0.05	-0.13	0.05	-0.13	-0.10	-0.04	-0.26	0.10	-0.15	-0.11
rus	-0.33	-0.05	0.07	0.00	-0.23	-0.16	-0.06	-0.31	0.15	0.18
spa	-0.41	-0.53	-0.31	-0.52	-0.50	-0.43	-0.37	-0.33	-0.49	-0.45
swe	0.04	0.09	0.16	0.18	-0.10	-0.01	0.06	-0.01	-0.03	-0.08
tha	0.02	-0.44	-0.19	-0.22	-0.34	-0.35	-0.39	-0.34	-0.53	-0.54
tur	-0.34	0.13	-0.17	-0.21	0.08	0.07	-0.13	-0.19	-0.28	-0.17

Table 5.17. UDPipe Pearson's correlation coefficient for each PUD language and each typological strategy calculated between the language distances and the LAS deltas. Values in green indicate a strong correlation, and in yellow, a moderate one. Correlation coefficients between -0.50 and -0.40 are presented in red.

b) Evaluation in terms of the number of right choices

In this part what is evaluated is the capability of each typological strategy of finding the best choice in terms of language pair for each PUD language (i.e.: for each PUD language, the selected second language has the smallest distance and provides the best positive delta).

Thus, we verify each typological method, and for LAS and MLAS metrics in terms of:

1) The number of selected language pairs (i.e.: smallest distance) which coincide with the best delta score obtained empirically;
2) The number of selected pairs that are not the same as the best but that are statistically equal in terms of delta;
3) The number of selected language pairs that provide negative deltas;
4) The number of selected languages pairs that deliver significantly lower deltas compared to the best one, but that are, at least, statistically positive;

The best strategy, then, is the one with the highest value of the sum of 1 and 2, and the lowest 3. Criteria 4 is used in case two or more typological methods present the same score in terms of 1, 2, and 3.

Regarding LAS, the results are presented in table 5.18. Considering the criteria described previously and the results from table 5.18, it is possible to conclude that when the typological methods are evaluated regarding the choice of best pairs, the best strategy concerns the classification obtained using MarsaGram all properties language vectors (Euclidean). The second-best option corresponds to the same approach but with cosine dissimilarity values.

In the experiments involving LAS, the number of PUD languages that obtained a positive delta is 17 (i.e.: no improvement was observed for Hindi, Japanese, and Korean). Thus, MarsaGram all properties (Euclidean) provides the best results for this metric for 47% of the cases (8 out of 17). Together with the 4 other statistically positive deltas, this method ensures a positive synergy for 70% of the languages (12 out of 17).

The MarsaGram linear (cosine) strategy, which was the method with the best correlation scores, is not the best one when the identification of the best pairs is considered: the number of right and equal to right choices is lower (2 less than the best method), and the number of negative delta propositions is a bit higher (1 more than the best strategy). Moreover, MarsaGram linear (Euclidean) and Head and dependent (cosine) methods also provide the second-best number of right or equal to right pairs. However, they present more cases in terms of negative deltas.

The verb and object position (cosine) strategy has the best score if all 4 criteria are considered, however, the number of right choices is very low. This strategy provides statistically valid positive results in terms of LAS delta for the selected pairs, however, in most cases, the results are statistically lower than the best empirical delta.

		Right Choice **(1)**	Equal to right **(2)**	**(1) + (2)**	Negative **(3)**	**(1) + (2) – (3)**	Lower than right but positive **(4)**	**(1) + (2) – (3) + (4)**
MarsaGram All	Euc	5	3	8	1	7	4	11
MarsaGram All	cos	5	2	7	1	6	5	11
MarsaGram Linear	Euc	6	1	7	2	5	3	8
MarsaGram Linear	cos	5	1	6	3	3	3	6
Head Dependent	Euc	5	1	6	2	4	5	9
Head Dependent	cos	6	1	7	2	5	5	10
VO_OV	Euc	0	3	3	2	1	8	9
VO_OV	cos	3	2	5	1	4	8	12
Lang2vec	Euc	4	2	6	1	5	6	11
Lang2vec	cos	2	2	4	1	3	7	10

Table 5.18. Evaluation of each typological strategy regarding the selection of best pairs in comparison with the empirical LAS delta results. The cells in green correspond to the highest values and the yellow ones to the second highest values. For the negative delta column, these colours indicate the smallest scores.

As noticeable from the values presented in table 5.18, the strategies (Euclidean and cosine) concerning the syntactic information from lang2vec do not present a number of right or equal to right choices comparable to the best-identified strategy (MarsaGram all).

For MLAS, the overall results are presented in table 5.19. For this metric, the best-identified methodology to choose the right pair is the state-of-the-art one, using lang2vec information with Euclidean distances. It provides the right choice for 10 out of the 17 PUD languages (59%) for which the highest positive delta was observed. Another advantage is that this method is the only one which does not propose any pair with negative synergy. However, as it was presented previously, to extend this method to languages other than the PUD ones, all the 41 lang2vec syntactic features must have associated values in URIEL.

If the URIEL information is missing, but the new language has at least a minimum of annotated data, it is possible to apply the second-best strategy identified to improve MLAS which is the MarsaGram linear (cosine). This methodology corresponds to the one with the best values in terms of correlation. It provides almost the same number of pairs corresponding to the best or right choice (1 less than lang2vec), and the second smallest number of negative deltas (1 more than lang2vec).

		Right Choice **(1)**	Equal to right **(2)**	**(1) + (2)**	Negative **(3)**	**(1) + (2) – (3)**	Lower than right but positive **(4)**	**(1) + (2) – (3) + (4)**
MarsaGram All	Euc	5	4	9	3	6	4	10
MarsaGram All	cos	5	4	9	2	7	5	12
MarsaGram Linear	Euc	4	5	9	2	7	2	9
MarsaGram Linear	cos	4	5	9	1	8	2	10
Head Dependent	Euc	4	4	8	2	6	6	12
Head Dependent	cos	4	4	8	3		5	10
VO_OV	Euc	0	5	5	1	4	5	9
VO_OV	cos	3	3	6	3	3	3	6
Lang2vec	Euc	4	6	10	0	10	3	13
Lang2vec	cos	2	6	8	0	8	4	12

Table 5.19. Evaluation of each typological strategy regarding the selection of best pairs in comparison with the empirical MLAS delta results. The cells in green correspond to the highest values and the yellow ones to the second highest values. For the negative delta column, these colours indicate the smallest scores.

MarsaGram all properties (cosine) is also a good candidate. It provides 2 negative deltas (1 more than MarsaGram linear described in the previous paragraph), but with the same number of right or equal to right choices, and it has a larger number of cases for which the result is lower than the best one but is, at least, positive.

Thus, for each dependency parsing metric, it is possible to conclude:

- In terms of LAS, MarsaGram all properties (Euclidean) provides the best results in terms of identification of best language pairs.
- In terms of MLAS, lang2vec (Euclidean) is the best choice, followed by MarsaGram linear (cosine) and MarsaGram all properties (cosine).

When correlations were analysed, the same typological method was identified for the 2 metrics, however, results differ when the identification of best pairs is considered. From all the possible quantitative syntactic information extracted from the corpora in this study, it seems that, overall, the patterns extracted with MarsaGram tool are the most pertinent ones as this tool appears in all the selected methods (either all properties or just linear ones depending on the evaluation strategy). Moreover, when not only linear patterns (i.e.: element A precedes element B) are considered, it is possible to obtain significantly good results for both LAS and MLAS deltas (using Euclidean distances for the first, and cosine for the second).

Up to this point, each typological strategy has been evaluated individually. Each one presents different values of Pearson's correlation, indicating that some linear correlation is observed. Therefore, it is legit to ponder whether better results could be achieved if these methods are combined.

Thus, we decided to proceed with a multivariate linear regression analysis of our data to check if an improvement can be obtained in terms of the right choice of best language pairs. As defined by Alexopoulos (2010): "Linear regression is the procedure that estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable which should be quantitative".

The aim of a linear regression model is to predict Y based on variables X:

$$(5.5)\ X_1, X_2, \dots, X_n \rightarrow Y$$

Where:

- X_i is defined as "predictor", "explanatory", or "independent" variable
- Y is the "dependent", "response", or "outcome" variable

Moreover, in a multiple linear regression model, Y is defined as:

$$(5.6)\ Y = \Theta_0 + \Theta_1 X_1 + \cdots + \Theta_n X_n$$

The ensemble of Θ_i corresponds to the regression coefficients (i.e.: weights associated with each independent variable).

The idea is to find the best regression coefficients that minimize the residual standard deviation (J) (or cost) of the estimated Y (Y_{est}), and the real expected values (Y):

$$(5.7)\ J = \sum (Y - Y_{est})^2$$

Thus, a python script has been created to proceed with the experiments regarding linear regressions using NumPy library[39].

The first step of the script concerns the normalization of the variables (both independent and response ones), also called "Feature scaling". It is a crucial phase as each typological method provides different magnitudes in terms of language distances. When the discrepancy in the scale of each variable is high, the algorithm can fail to converge (i.e.: find the best optimum in terms of weights).

The second stage consists of assigning an initial value for Θ, which allows the first estimation of Y_{est}. After, the gradient function is computed. It means the calculation of the necessary gradients of the cost function that allow the model to be iterated and optimized. The gradient is defined (for each variable X_i) as:

$$(5.8)\ \frac{\partial J}{\Theta_i} = \frac{1}{n}\,(Y_{est} - Y)X_i$$

The next step is the implementation of the gradient descent algorithm which consists of:

- Obtaining the gradients of J according to the actual values of the parameters
- Calculating the cost in each iteration
- Updating all parameters according to:

$$(5.9)\ \Theta_i^{k+1} = \Theta_i^k - \alpha\,\frac{\partial J}{\Theta_i}$$

Where α is the learning rate (a pre-defined parameter), and k refers to the current iteration. This step is repeated until the convergence of the algorithm or until the number of iterations (epochs) reaches a maximum pre-determined value.

Besides the maximum number of iterations, a tolerance value (TOL) is defined to avoid useless iterations:

$$(5.10) \sum \left(\Theta^{i+1} - \Theta^{i}\right)^2 < TOL$$

---

[39] https://numpy.org/

If the change regarding the weights is below TOL, the system stops the iteration process. Once the algorithm has stopped, it is possible to check the predictions made with the calculated weights and compare them with the expected results.

More specifically, in this study, the independent variables are the set of language distances provided by each method (considering all possible language pairs), and the prediction corresponds to the empirical delta in terms of LAS and MLAS (analysed separately). As what is observed in most cases in terms of correlation is an inverse correlation, instead of using Y as the normalized deltas, we define the predicted variable as:

$$(5.11)\ Y = 1 - Delta$$

In terms of parameters that need to be defined at the beginning of the process, we decided to keep the same TOL for all experiments, but to vary the initial Θ and the learning rate. Thus:

- TOL = 10^{-7};
- Initial Θ: 0.1, 0.4, and 0.7;
- Learning rate: 0.1 and 0.5.

In terms of variables, we decided to combine the Euclidean and the cosine methods separately. Thus, we have (in each case):

1) MarsaGram all properties;
2) MarsaGram linear;
3) Head and Dependent position;
4) Verb and Object position.

MarsaGram linear is a subset of the MarsaGram all properties, and verb and object position, a subset of Head and Dependent. Thus, the possible combinations tested are:

1) All methods combined (1+2+3+4);
2) MarsaGram all + Head and Dependent (1+3);
3) MarsaGram all + Verb and Object (1+4);
4) MarsaGram linear + Head and Dependent (2+3);
5) MarsaGram linear + Verb and Object (2+4).

The first one consists of the overall combination and the other possibilities concern association of methods that are not a subset of the others. Thus, with all these 5 combinations of methods and with the variation of the initial weights and learning rates, 30 experiments were conducted for each metric (LAS and MLAS) and each distance calculation (Euclidean and cosine), a total of 120 experiments.

The results obtained for each method association were analysed in terms of the right choices as described previously when methods were evaluated independently. Regarding LAS, the ensemble of results is presented in the Annex 70.

In table 5.20, we compare the best LAS result obtained via the combination of the typological methods with the best-identified candidate when the strategies were considered independently (i.e.: MarsaGram all Euclidean).

		Right Choice **(1)**	Equal to right **(2)**	**(1) + (2)**	Negative **(3)**	**(1) + (2) – (3)**	Lower than right but positive **(4)**	**(1) + (2) – (3) + (4)**
MarsaGram All	Euc	5	3	8	1	7	4	11
MarsaGram All + HD (learning rate = 0.5, Θ = 0.4)	Euc	6	3	9	0	9	3	12

Table 5.20. Comparison of results of the best typological method and the best combination of methods in terms of the highest number of right choices (LAS improvement). The best results are presented in green while the second-best ones are displayed in yellow.

Thus, it is possible to observe that by combining different methods, the overall results are improved. The best combination concerns the association of the previously identified best candidate (MarsaGram All Euclidean) with the head and dependent strategy. There is an increase in terms of the best or equal to the best choice and a decrease in terms of negative deltas. With the combination of methods, no selected language pair provide a decrease in LAS. In the combined scenarios, from all the 17 cases in which PUD languages present a positive synergy, this method provides the right choice in 52% of the cases, and allows us to predict at least a positive significant delta for 70% of the languages.

In terms of optimized Θ, which allows a better estimation of language distances ($\mathrm{D}_{\mathrm{opt}}$) to find the best language associations, we have:

$$(5.12)\ D_{opt} = 0.036\ D_{Marsagram\ all} + 0.397\ D_{head\ and\ dependent}$$

Thus, although alone MarsaGram all properties strategy is the best candidate, when the combination is established, this method is associated to a relatively lower weight.

Concerning LAS, Hindi, Japanese, and Korean are the languages for which no positive delta has been observed. Besides these languages, the selected combined strategy fails to provide at least a positive delta to Arabic, Chinese, Icelandic, Indonesian, and Thai. From this group of challenging languages, except for Icelandic, all the others are non-Indo-European. However, the method was successful in identifying the right pair for Turkish and Finnish.

For the abovementioned languages, some other typological strategies are able to propose the right choice or at least one with a positive delta:

- Arabic: right or statistically similar to right with verb and object position method (Euclidean and cosine);
- Chinese: the right choice is proposed by the verb and object position (cosine) strategy;
- Icelandic: all methods fail in proposing the right choice but MarsaGram linear (Euclidean and cosine), verb and object (Euclidean and cosine), and lang2vec (Euclidean and cosine) propose language pairs with positive deltas;
- Indonesian and Thai: all methods fail in proposing either the right choice or positive deltas.

Thus, it seems that if the selected combination of MarsaGram all and head and dependent position fails in providing an improvement, the verb and object position can be tested instead. It does not guarantee positive deltas for Indonesian and Thai, but at least it does not implicate negative deltas for these two languages. Thus, it seems that the verb and object position method allows a better identification of the best pairs for languages that differ the most in terms of genealogical features.

In terms of MLAS, the complete results of the association of typological methods are presented in Annex 71. The comparison between the best combined results with the best methods identified previously is displayed in table 5.21.

It is possible to notice that MarsaGram all properties combined with head and dependent position strategy provides the good results when considering the associations tested. In the specific case of MLAS, different values of Θ and learning rates for this type of association converge to the same optimized result. Moreover, in some cases, the combination of MarsaGram all properties and verb and object position method also provide similar results. However, all the combined experiments do not present any improvement when compared to the previously identified methods. The combined scores are identical to the ones obtained with MarsaGram all properties (cosine) alone.

		Right Choice **(1)**	Equal to right **(2)**	**(1) + (2)**	Negative **(3)**	**(1) + (2) – (3)**	Lower than right but positive **(4)**	**(1) + (2) – (3) + (4)**
MarsaGram All	cos	5	4	9	2	7	5	12
MarsaGram Linear	cos	4	5	9	1	8	2	10
Lang2vec	Euc	4	6	10	0	10	3	13
MarsaGram All + HD (learning rate = 0.5, Θ = 0.4)	cos	5	4	9	2	7	5	12
MarsaGram All + HD (learning rate = 0.5, Θ = 0.1)	cos	5	4	9	2	7	5	12
MarsaGram All + HD (learning rate = 0.5, Θ = 0.7)	Euc	6	3	9	2	7	5	12
MarsaGram All + HD (learning rate = 0.5, Θ = 0.7)	cos	5	4	9	2	7	5	12
MarsaGram All + VO (learning rate = 0.5, Θ = 0.7)	Euc	6	3	9	2	7	5	12
MarsaGram All + HD (learning rate = 0.1, Θ = 0.7)	Euc	6	3	9	2	7	5	12
MarsaGram All + VO (learning rate = 0.1, Θ = 0.7)	Euc	6	3	9	2	7	5	12

Table 5.21. Comparison of results of the best typological method and the best combination of methods in terms of the highest number of right choices (MLAS improvement). The best results are presented in green while the second-best ones are displayed in yellow.

Thus, regarding MLAS, we keep the previously selected methods:

- Lang2vec (Euclidean);
- MarsaGram all properties (cosine);
- MarsaGram linear properties (cosine).

No combined method permitted the reduction of the number of negative deltas and to reach the number of right or similar to right choices provided by lang2vec strategy.

As previously presented, Korean, Japanese, and Thai did not present any improvement in terms of MLAS when combined with other languages. Besides them, each one of the selected methods fails to find the best choice (or at least one providing positive delta):

- Lang2vec (Euclidean): Chinese, English, Hindi, Indonesian;
- MarsaGram all (cosine): Chinese, English, Indonesian;
- MarsaGram linear (cosine): Chinese, Czech, German, Hindi, Indonesian, Italian.

As previously seen, lang2vec is the only method without any selected pair with a negative delta. The advantage of MarsaGram linear (cosine) is that it has only one pair with a negative synergy (Japanese), while MarsaGram all (cosine) has 2 (Japanese and Korean). However, MarsaGram all (cosine) has the smallest number of languages listed above (as it predicts the correct pair for Hindi). MarsaGram linear (cosine) has the largest number of languages without any improvement but it predicts the right choice for English while the other 2 selected methods fail.

For Chinese and Indonesian, no other tested strategy can predict a pair with a significantly positive delta. Thus, the three selected ones correspond to the most optimized strategies.

5.4 Overall Discussion

In section 4, we presented the different typological approaches for language classification concerning either phylogenetic aspects or different types of syntactic features. In this section, we tested the different strategies to see which one corresponds better to what is observed when languages are combined to improve dependency parsing results.

It is clear that when different parsing evaluation metrics are considered, different results are obtained, thus different optimized strategies are selected. Moreover, it was possible to observe that results are not similar when applying different evaluation methods concerning the relation between language distances and the empirical deltas.

Concerning LAS, it was possible to identify that the MarsaGram linear (cosine) is the method with a higher correlation with the obtained deltas. However, when it comes to the selection of best language pairs, we observed that a specific combination of MarsaGram all properties (Euclidean) with head and dependent method (cosine) is the one proposing the best scores.

When analysing the cases for which this combined method fails, it was possible to notice that the verb and object method provides positive results in most of these complicated cases.

For MLAS, the same method (MarsaGram linear with cosine distances) is identified as the one with the best correlation results. However, in terms of the identification of best pairs, MLAS differs considerably. The best results were obtained with lang2vec, followed by MarsaGram all (cosine) and MarsaGram linear (cosine). Each one of these 3 strategies have specificities in terms of languages for which they fail, and no other method selects pairs with better results in these challenging cases.

Thus, when the phenomenon of dependency parsing improvement via typological strategy is analysed in its globality, the specific word order between components which belong to the same subtree but are not necessarily a pair of head and dependent (i.e.: MarsaGram linear patterns) is the factor that influences the most the empiric results regarding language combination when compared to the other syntactic phenomena examined in this thesis. And this is valid when both dependency parsing evaluation metrics are considered.

On the other hand, when a specific analysis is conducted focusing on the identification of the closest languages, the scenario is different. Regarding only the quantitative typological methods, for LAS, better identification of the pairs providing the best improvements is provided when not only the specific word order inside subtrees is considered, but also includes other information of phenomena happening inside each subtree (i.e.: exclude, require, and unicity relations), and when this information is combined with specific head and dependent word order patterns (i.e.: MarsaGram all properties combined with head and dependent strategy). However, it was observed that this combined method is most efficient in cases where the analysed languages share some phylogenetic features. When languages come from genealogical groups which are distant, the best pair can be better identified with the comparison of the patterns regarding verb and object positions.

For MLAS, the quantitative methods did not provide a real improvement in terms of the identification of the best pairs when compared to the language comparison provided by lang2vec. However, the results are quite close. It was possible to observe that the typological strategy with the best correlation results (i.e.: MarsaGram linear) also provides interesting results in terms of best deltas. However, in terms of the overall number of pairs providing at least a statistically valid improvement, it is better to use not only MarsaGram linear properties but also the other possible relations between the components inside the subtrees.

As explained previously, MLAS metric is more complex than LAS as it considers not only the identification of the heads and the dependency labels but also analyses the UPOS and FEATS. This difference is most probably the cause of the observed variances in terms of the best typological strategies for each metric. Moreover, when corpora are combined, not only information regarding the dependency relations is provided in the added corpus. Instead, we kept the second language corpora as it is presented in the PUD collection, thus, presenting lexical (word forms and lemmas), part-of-speech (UPOS), and morphosyntactic (FEATS) information. Therefore, when the UDify models are trained with the combined corpora, there are also changes in the training of the other modules of this tool, certainly with an impact on the UPOS and FEATS annotation.

Moreover, it is noteworthy to mention that the Universal Dependencies framework and its collection of corpora are in constant evolution. In each new version, some changes can be observed in terms of pre-established labels, and corpora may present some corrections to be better harmonized with the framework guidelines. In this thesis, the ensemble of experiments was conducted with v.2.7 which was released in November 2020. Since then, new versions have been published. When we compare the PUD corpora v.2.7 with the v.2.10 (released on May 2022), some modifications are noticeable. The ensemble of differences between the two versions is presented in Annex 72. For English, Icelandic, Italian, Japanese, Spanish, and Swedish, there are none or a few differences. In most cases, there are some improvements in terms of FEATS annotation, and some slight corrections in terms of DEPREL labels (e.g.: suppression of the label with type and subtype “det:predet” which is replaced by “det”). These changes could have a positive impact mostly on MLAS results, as the ensemble of FEATS annotation is more consistent. However, for 2 PUD languages some major changes are observed in terms of UPOS: for Thai, in many cases, the UPOS labels VERB and PART were changed to AUX, and for Turkish, NOUN and ADJ were changed to VERB.

With the best strategies identified in this section, it is possible to proceed with the complete classification of all European Union languages and to test how well EU low-resourced languages can be improved with these typological methods.

6. Typological Analysis and Dependency Parsing Improvement of EU Languages

The aim of this section is to provide a complete typological analysis of all 24 European Union languages using the most adapted methods determined in the previous chapter. Moreover, a series of corpora combination experiments regarding some EU low-resourced languages for dependency parsing improvement is presented. These trials rely on the typological information (i.e.: language distances) obtained with each one of the selected strategies.

First, all EU languages are described in terms of some typological features and regarding their genealogical families and genera. Then, we present an analysis regarding different aspects to determine which EU languages can be considered low-resourced ones concerning dependency parsing. Following, we detail the different corpus-based typological classifications of EU languages with the following quantitative methods:

1) MarsaGram linear (cosine): this method showed the best overall correlation between LAS and MLAS metrics and the distance between language vectors. Moreover, this method was identified as a good candidate for choosing the best combination in terms of MLAS.
2) Combination of MarsaGram all properties and Head and Dependent strategies (Euclidean): the specific association of these two methods proved to be the most optimized way to identify the best candidate for LAS improvement when two corpora are combined.
3) Verb and Object relative position method (cosine): this method has shown interesting results in terms of LAS improvement for specific PUD languages that do not have any close-related languages in the analysed ensemble in terms of phylogenetic features.
4) MarsaGram all properties (cosine): this approach was identified (together with MarsaGram linear) as one of the best strategies for MLAS improvement (i.e.: the lowest number of proposed associations with negative synergy).

Although presenting the best results in terms of MLAS improvement (i.e.: identifying the language-pairs which provide the best deltas), the lang2vec approach cannot be applied to all EU languages as it was presented in this thesis regarding the PUD collection. As explained previously, PUD languages were compared in terms of 41 syntactic features which have valid values for all of them. However, when analysing EU languages, the number of common lang2vec syntactic features is 0, thus, this method cannot be exploited for this specific set of languages.

After the detailed analysis of the typological methods, we present the results of the dependency parsing experiments regarding the identified low-resourced languages in terms of LAS and MLAS improvements.

6.1. European Union Languages Characterization

The European Union has 24 official languages: Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish, and Swedish.

These languages are presented in Table 6.1 together with their respective ISO 639-3 code and the phylogenetic information provided by WALS (Dryer et al., 2013). The geographical information is not displayed in the table as all languages belong to the Eurasia area.

Language	ISO 639-3 code	Family	Genus
Bulgarian	bul	Indo-European	Slavic
Croatian	hrv	Indo-European	Slavic
Czech	ces	Indo-European	Slavic
Danish	dan	Indo-European	Germanic
Dutch	nld	Indo-European	Germanic
English	eng	Indo-European	Germanic
Estonian	est	Uralic	Finnic
Finnish	fin	Uralic	Finnic
French	fra	Indo-European	Romance
German	deu	Indo-European	Germanic
Greek	ell	Indo-European	Greek
Hungarian	hun	Uralic	Ugric
Irish	gle	Indo-European	Celtic
Italian	ita	Indo-European	Romance
Latvian	lav	Indo-European	Baltic
Lithuanian	lit	Indo-European	Baltic
Maltese	mlt	Afro-Asiatic	Semitic
Polish	pol	Indo-European	Slavic
Portuguese	por	Indo-European	Romance
Romanian	ron	Indo-European	Romance
Slovak	slk	Indo-European	Slavic
Slovenian	slv	Indo-European	Slavic
Spanish	spa	Indo-European	Romance
Swedish	swe	Indo-European	Germanic

Table 6.1. List of EU languages with their respective ISO 639-3 three-character code, their phylogenetic, and geographical information.

As previously mentioned, 10 out of the 24 EU languages are present in the PUD collection. In the official EU language-set, the vast majority concerns Indo-European languages (i.e.: 20 out of the 24, being 6 from the Slavic genus, 5 from the Germanic, 5 from the Romance, 2 from the Baltic, 1 from the Celtic, and one from the Greek genus). There are 3 Uralic languages (2 Finnic and 1 Ugric), and one Afro-Asiatic language (Maltese) from the Semitic genus.

Thus, a language-set composed of PUD and EU languages (i.e.: 34 languages) has 9 different linguistic families and 16 different genera. The Indo-European family is the best-represented (23 out of the 34). Although presenting some linguistic variability, the PUD collection does not have any language from the Baltic and Celtic genera of the Indo-European family. This larger ensemble of language is, as expected, richer in terms of phylogenetic information but with the same extent in terms of geographical area as the PUD collection.

Another approach in terms of genealogical analysis of this set of languages concerns the usage of the phylogenetic features provided by lang2vec (as conducted for PUD languages in Section 4.3). The dendrograms regarding the cluster analysis using both Euclidean and cosine distances are presented in the Figures 6.1 and 6.2 respectively, and the obtained dissimilarities matrices, in Annexes 73 and 74.

Figure 6.1. Euclidean dendrogram concerning the lang2vec phylogenetic comparison of EU and PUD languages.

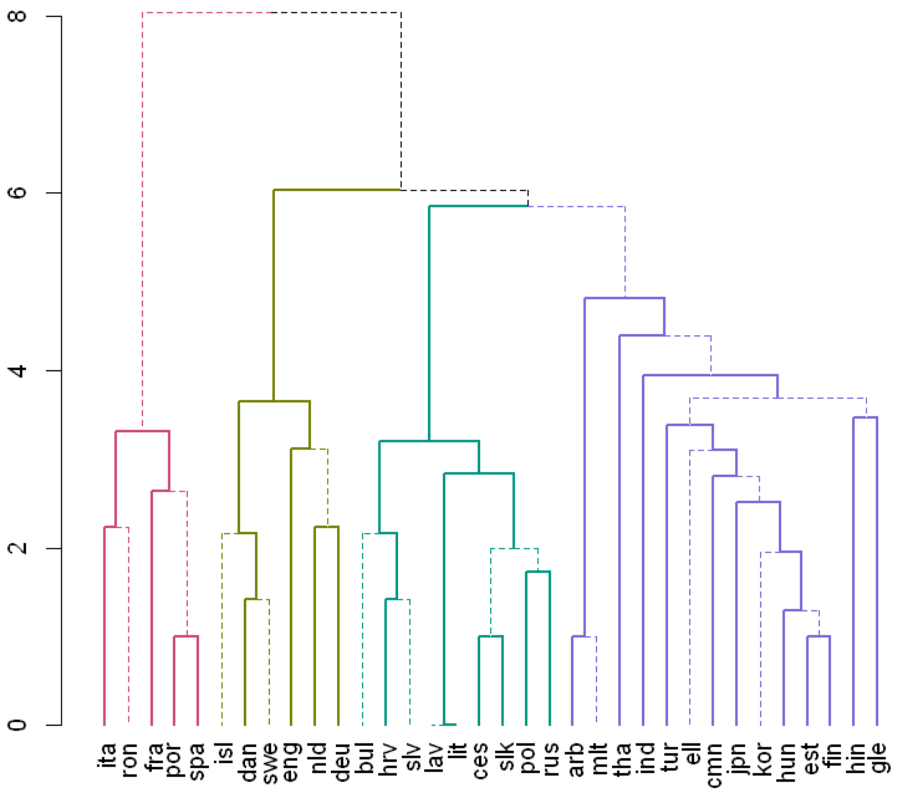
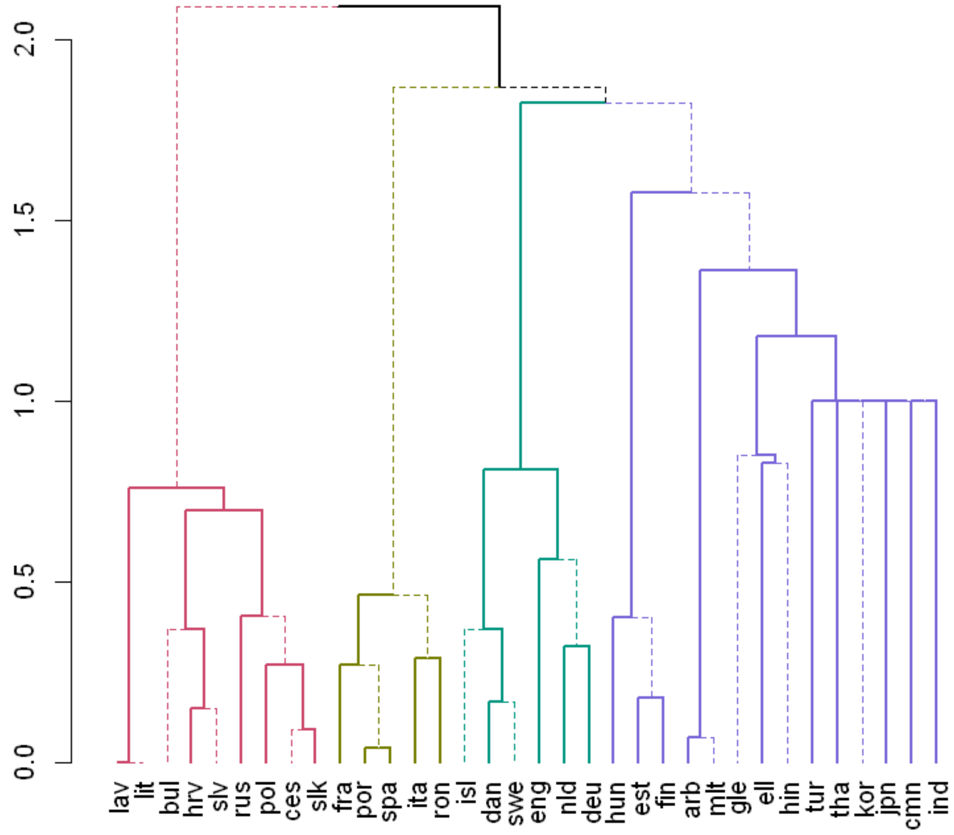


Figure 6.2. Cosine dendrogram concerning the lang2vec phylogenetic comparison of EU and PUD languages.



As was the case when only PUD collection was examined, the cosine dendrogram provides a better classification of the 34 languages of the enlarged language-set. It is possible to observe in both graphs that all 5 Romance languages are clustered together (with a sub-cluster formed by Portuguese and Spanish which is closer to French, and another one composed of Italian and Romanian). Germanic languages are also similarly clustered in both dendrograms, with the proper identification of the genealogical proximity between German and Dutch, and between Danish and Swedish.

Another identified cluster in both figures is composed of Slavic and Baltic languages. In this case, the cosine dendrogram is more accurate as it presents a clear separation between the sub-clusters of these 2 genera. Regarding the Slavic genus, it is possible to notice the proximity of Croatian and Slovenian (sub-cluster closer to Bulgarian), and the closeness of Czech and Slovak (sub-cluster closer to Polish and Russian).

Moreover, the Afro-Asiatic languages from the Semitic genus (i.e.: Arabic and Maltese) are also grouped in an isolated sub-cluster of the purple cluster in both dendrograms.

The advantage of the cosine dendrogram is more precisely verified when checking the classification of the other Indo-European languages as they form specific sub-clusters of the purple group in the figure. It is possible to identify the Uralic languages in a small group (with a clear distinction between the Finnic and Ugric ones), and another sub-class composed of Greek, Hindi, and Irish. Furthermore, in the cosine genealogical classification, it is noticeable that the languages that do not share any common phylogenetic features with the others (i.e.: Chinese, Indonesian, Japanese, Korean, Thai, and Turkish) are represented as an isolated sub-cluster with distance 1 between them).

Besides the genealogical classification of the EU languages, it is also possible to analyse them in terms of certain typological aspects. Table 6.2 displays some word-order characteristics and the associated language type provided by Hawkins (1983) for the ensemble formed by PUD and EU languages. There is no description for 7 languages in Hawkins (1983) reference: Bulgarian, Croatian, Hungarian, Latvian, Maltese, Polish, and Slovak. Chinese, English, German, and Slovenian are described but not associated with a language-type. Considering the languages which are classified into language-types, we have:

- Type 1: Arabic and Irish;
- Type 9: French, Indonesian, Italian, Portuguese, Romanian, Spanish, and Thai;
- Type 10: Czech, Dutch, Greek, Icelandic, and Russian;
- Type 11: Danish, Lithuanian, and Swedish;
- Type 15: Estonian and Finnish;
- Type 23: Hindi, Japanese, Korean, and Turkish.

As was the case when only PUD languages were considered, all Romance languages of this enlarged set are classed as the same type (9) which also includes Indonesian and Thai. Although being Indo-European, in terms of word-order, Irish is grouped with Arabic in type 1. Estonian and Finnish are from the same linguistic family and genus and were grouped as type 15. The language-type 10 presents Slavic, Germanic and Greek languages. However, not all Germanic ones are included in this type, Danish and Swedish share word-order patterns with Lithuanian (type 11). Type 23, composed of OV languages, contains Hindi, Japanese, Korean, and Turkish. The majority of languages are SVO, and the number of SOV languages in this set is the same as in the PUD collection (i.e.: there is no SOV language in the EU official languages). Four languages are exceptional:

- Chinese: SOV and SVO

- English: SVO and V-1 (less common)
- German: SOV, V-1 (less common), and V-2
- Irish: V-initial

Lang.	**Hawkin's Word Order Summary**			**Type**
arb	VSO	Pr	NumN/nnum, DN, NPoss, NA, NG, Nrel	1
bul	-	-	-	-
cmn	SOV/SVO	Pr/Po	DN, AN, GN, RelN	-
hrv	-	-	-	-
ces	SVO	Pr	NumN, DN, AN, NG, NRel	10
dan	SVO	Pr	NumN, DN, PossN, AN, GN, NRel	11
nld		Pr	NumN, DN, PossN, AN, NG, NRel	10
eng	SVO/v-1	Pr	NumN, DN, PossN, AN, GN/NG, NRel	-
est	SVO	Po	AN, GN	15
fin	SVO	Po	NumN, DN, AN, GN, reln/Nrel, AdvAdj, SMAdj/AdjMS	15
fra	SVO	Pr	NumN, DN, PossN, an/NA, NG, NRel	9
deu	SOV/v-1,V-2	po/Pr	NumN, DN, PossN, AN, GN/NG, reln/NRel	-
ell	SVO	Pr	NumN, DN, AN, NG, Nrel, AdjAdv, AdjMS	10
hin	SOV	Po	NumN, DN, AN, GN, NRel/RelNRel, AdvAdj, SMAdj	23
hun	-	-	-	-
isl	SVO	Pr	DN, AN, NG, NRel	10
ind	SVO	Pr	NumN, ND, NPoss, NA, NG, NRel	9
gle	V-initial	Pr	NumN, ND, PossN, NA, NG, Nrel	1
ita	SVO	Pr	NumN, DN, an/NA, NG, NRel	9
jpn	SOV	Po	NumN/NNum, DN, AN, GN, RelN, AdvAdj, SMAdj	23
kor	SOV	Po	NumN, DN, PossN, AN, GN, RelN	23
lav	-	-	-	-
lit	SVO	Pr	AN, GN, Nrel	11
mlt	-	-	-	-
pol	-	-	-	-
por	SVO	Pr	NumN/NNum, DN, PossN/NPoss, an/NA, NG, NRel	9
ron	SVO	Pr	NA, NG, Nrel	9
rus	SVO	Pr	NumN, DN, AN, NG, NRel	10
slk	-	-	-	-
slv	SVO	Pr	NumN, DN, AN, GN/NG, Nrel	-
spa	SVO	Pr	NumN/NNum, DN, PossN/NPoss, an/NA, NG, NRel	9
swe	SVO	Pr	NumN, DN, PossN, AN, GN, NRel	11
tha	SVO	Pr	NumN, ND, NPoss, NA, NG, Nrel, AdjAdv, AdjMS	9
tur	SOV	Po	NumN, DN, AN, GN, RelN, AdvAdj, SMAdj	23

Table 6.2. Typological characteristics and classification of EU and PUD languages according to Hawkins (1983). When components are written in lower cases, it means that the phenomenon is less frequent than the other possible word order structure involving the same elements.

The analysis of the position of verb and nominal objects as presented in the WALS database, shows that most languages are VO, the exceptions being:

- Hindi, Japanese, Korean, and Turkish (OV)
- Dutch and German (No determinant order)

Slovak and Maltese do not have data regarding this syntactic feature in WALS. Like the other Slavic languages in our sample, Slovak is also VO (Lemay, 2008), although word-order in these languages is much more flexible than in most Romance languages. Maltese is also a VO language as stated by Čéplö (2018) in a quantitative analysis of the Maltese Universal Dependency corpus MUDT v.1.

The lack of information in WALS of this essential typological feature for these 2 languages together with the fact that it is not possible to compare the languages composing the PUD and EU language-set using lang2vec syntactic vectors show the limitation of using established typological data-bases in NLP studies.

The corpus-based typological analyses proposed in this thesis rely on the existence of annotated corpora according to UD framework, thus, in some cases, this requirement can also be limiting. However, at least for the sample analysed here, all languages possess at least one UD corpus available, allowing us to conduct all the necessary experiments.

6.2. European Union Low-resourced Languages

In the NLP field, low-resourced (or under-resourced) languages are the ones for which there is a lack of linguistically annotated data and/or technological resources (i.e.: trained models, software, etc), while, on the other hand, well-resourced (or high-resourced) languages possess a relatively large amount of data and tools. In most cases, languages with a large number of speakers in developed countries are usually well-resourced for the ensemble of NLP tasks, it is the case of English, German, and French, for example. Moreover, minority languages suffer from a lack of financial resources which are essential for the development of quality NLP resources.

Regarding the official European Union languages, in 2012, the Multilingual Europe Technology Alliance (META) published a series of white-papers concerning the availability of NLP resources (Rehm, G. et al., 2012) divided into four different domains: 1) machine translation, 2) speech processing, 3) text analysis, 4) speech and text resources. Languages were classified in terms of the quality of the available support in these areas: excellent, good, moderate, fragmentary, and weak/no support.

Since the publication of this white-paper series, many other resources have been developed for EU languages, moreover, in these reports, dependency parsing is considered inside the “text analysis” domain together with other possible NLP tasks. Thus, to determine the low-resourced EU languages pertinent to this thesis, we conducted an updated analysis specifically targeting dependency parsing.

As detailed in the previous sections, the experiments to determine the best corpus-based typological approaches were conducted in a low-resourced scenario (i.e.: the limited size of the UD corpora), thus, the idea is to identify the EU languages with deficiency in terms of annotated corpora following the UD framework for which the identified optimized strategies are more relevant. In Table 6.3, we detail, for each EU language, the number of available corpora, and the total size of the available data in terms of the number of sentences and tokens (from the UD v.2.11 released in November 2022).

It is important to notice that corpora may vary in terms of genre and specific choices in terms of annotation (e.g.: split of tokens into words, number of FEATS, etc). Moreover, some languages possess corpora composed of transcriptions from spoken samples. These spoken corpora have specificities present in oral language and are usually used in specific NLP applications.

As expected, there is a large discrepancy in terms of UD corpora size inside the EU language-group. Clearly, the most low-resourced languages are Hungarian, Maltese, Greek, and Lithuanian (i.e.: with less than 100,000 tokens). Six EU languages have from 100,000 to 200,000 tokens: Danish, Slovak, Irish, Bulgarian, Finnish, and Croatian. Five languages have UD corpora with a medium size (i.e.: from 200,000 to 500,000 tokens): Swedish, Latvian, Slovenian, Dutch, and Polish. The other 9 EU languages have a large amount of annotated data (i.e.: more than 500,000 tokens), and five languages have more than 1,000,000 tokens: French, Spanish, Portuguese, Czech, and German.

As presented by Otter et al. (2019), the efficiency of most deep-learning tools regarding dependency parsing relies on the availability of a large amount of annotated data. For software that use language models as part of their architecture not only the training data is important but also the size of the representation of the languages inside them. As presented in Table 5.2 (section 5.1.2), EU languages are not equally represented inside mBERT.

		Total Size	
Language	**Number of UD corpora**	**Sentences**	**Tokens**
bul	1	11,138	156,149
hrv	1	9,010	199,409
ces	5	127,507	2,218,708
dan	1	5,512	100,733
nld	2	20,944	306,720
eng	9	45,815	755,969
est	2	38,158	528,387
fin	4	21,845	194,484
fra	8	48,297	1,175,147
deu	4	208,440	3,748,450
ell	1	2,521	61,773
hun	1	1,800	42,032
gle	3	5,926	135,187
ita	9	37,871	822,952
lav	1	16,951	285,425
lit	2	3,905	75,403
mlt	1	2,074	44,162
pol	3	40,398	496,682
por	5	69,722	1,238,113
Ron	4	40,480	937,551
Slk	1	10,604	106,043
Slv	2	16,623	296,585
Spa	3	34,675	993,369
Swe	3	12,269	206,856

Table 6.3. Detailed information regarding the number of UD corpora and total size for each EU language.

Considering this parameter, the most under-resourced EU languages are: Maltese (absent in this language-model), Irish, Latvian, and Lithuanian. Besides them, Bulgarian, Croatian, Danish, Estonian, Greek, Slovak, and Slovenian also have a relatively small representation in mBERT.

To complement the analysis to determine the most low-resourced EU languages, we decided to check the results in terms of LAS and MLAS obtained with UDify as presented by Kondratyuk and Straka (2019). The authors conducted a series of experiments using a multilingual model trained with 124 corpora (75 languages) from UD v.2.3 which was tested with monolingual test-sets. Thus, this experiment shows how well this specific deep-learning architecture works for languages with different sizes of training-sets and language representation in mBERT. The ensemble of UDify LAS and MLAS results are displayed in Annex 75 (for languages with more than one training set, we present the scores corresponding to the corpus providing the best MLAS). Figures 6.3 and 6.4 present the LAS and MLAS values for each EU language in an increasing order.

Figure 6.3. LAS values for each EU language obtained by Kondratyuk and Straka (2019).

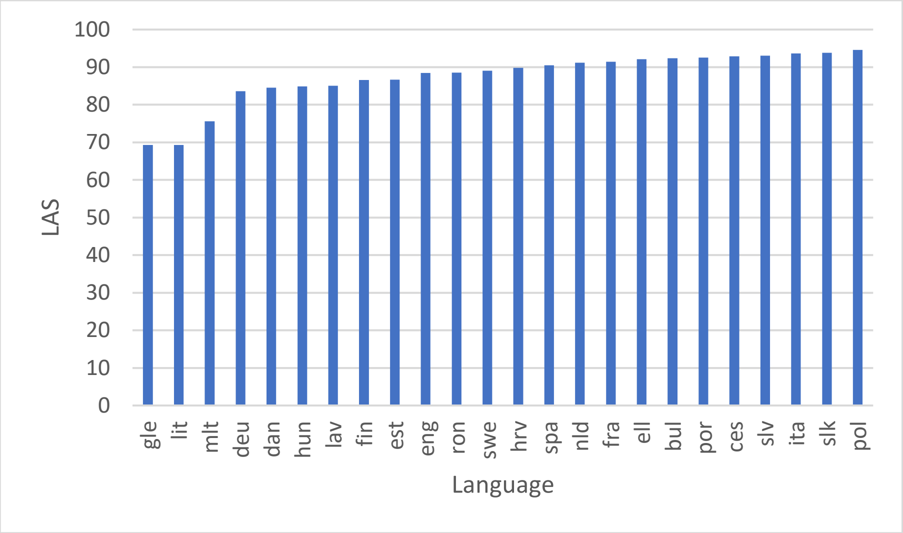
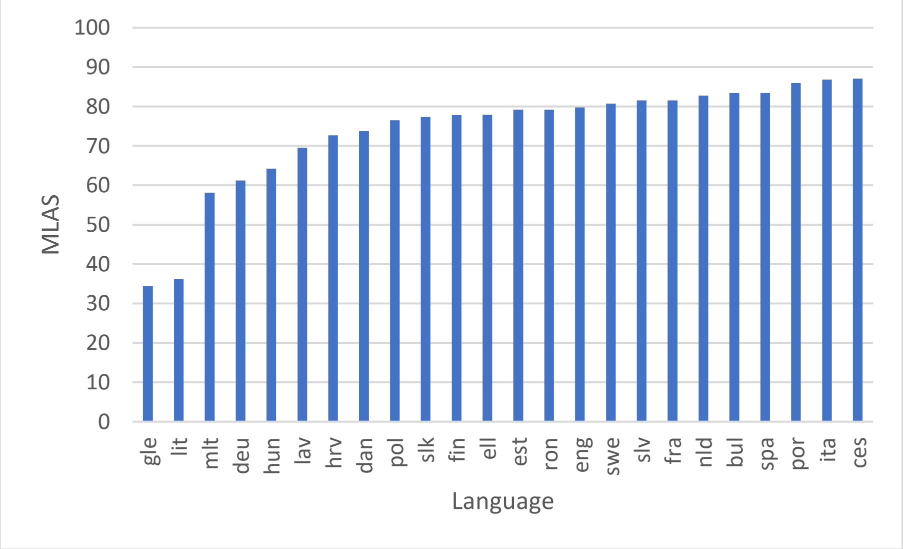


Figure 6.4. MLAS values for each EU language obtained by Kondratyuk and Straka (2019).



By analysing Figures 6.3 and 6.4, it is possible to notice that the languages with the worst results in terms of LAS and MLAS are Irish, Lithuanian, and Maltese (i.e.: LAS lower than 80 and MLAS lower than 60). Maltese and Lithuanian were also identified (together with Hungarian and Greek) as the ones with the smallest UD corpora. When the size of the language representation in mBERT was considered, the 3 languages with the lowest LAS and MLAS are also classified as low-resourced ones (especially Maltese which is not present in mBERT).

As it was discussed when UDify results regarding PUD collection were presented (Section 5.3.1), the size of the training corpus is not the only aspect with a major role in determining the efficiency of the parsing model. It is possible to see in Figures 6.3 and 6.4 that German has quite low scores even though being a well-resourced language in terms of the size of its UD corpora. Moreover, Greek has a small UD corpus, but its UDify results are comparable to languages with larger corpora.

Furthermore, even though having a small UD corpus, Hungarian has LAS and MLAS scores slightly higher than the threshold abovementioned (84.88 and 64.27 respectively) but lower than the results of the majority of EU languages. The size of this language in mBERT may explain the better scores when compared to Maltese and Lithuanian.

Therefore, considering all these elements, we decided to consider as low-resourced EU languages for the following experiments: Irish, Lithuanian, Maltese, and Hungarian. Two of them are Indo-European but from different genera. Irish does not have any similar language in the whole language-set in terms of the genus, while Lithuanian has Latvian (both from the Baltic genus). Moreover, Maltese has Arabic (both Semitic languages from the Afro-Asiatic family) but Hungarian does not have any language with the same genus (only from the same linguistic family: Finnish and Estonian). In terms of Hawkins (1983) classification, Irish is from the same language type as Arabic (type 1), and Lithuanian is classed with Danish and Swedish (type 11), but there is no data for Hungarian and Maltese.

Moreover, we decided to conduct the dependency parsing improvement with a language that is not considered a low-resourced one regarding the information described above. The aim is to check if the selected methods may provide some improvement even for languages with better resources. For this objective, we selected the Croatian language as it has a quite high LAS (89.79) that is comparable to many other EU languages, but its MLAS result is relatively small. Also, this language is from the Slavic genus which is quite well represented in our language-sample. The idea is to check whether the methods are effective in this more convenient scenario (where the language has a medium-sized UD corpus and mBERT size, and has close-related languages in the language-set).

6.3. Corpus-based Typological Classification of EU Languages

In this section, we present the obtained corpus-based typological classification of the ensemble formed by PUD and EU languages using the optimized methods previously selected:

1) MarsaGram linear (cosine);
2) Combination of MarsaGram all properties and Head and Dependent strategies;
3) Verb and Object relative position;
4) MarsaGram all properties (cosine).

For PUD languages, the syntactic features are extracted from the PUD collection as in the previous experiments. For the EU languages which are not part of PUD, the idea is to compose a typological corpus composed of 1,000 sentences (the same size as each one of the PUD corpora) from UD v.2.7 collection (same version used in the earlier experiments) to avoid bias related to corpus-size and UD annotations.

In terms of corpora choice for non-PUD EU languages, in the cases where more than 1 UD corpus is available, we selected the corpora whose genres were more similar to the ones in the PUD collection (i.e.: news and Wikipedia). For languages with only one UD corpus, no selection was required. Table 6.4 presents the description of the non-PUD EU corpora in terms of genres.

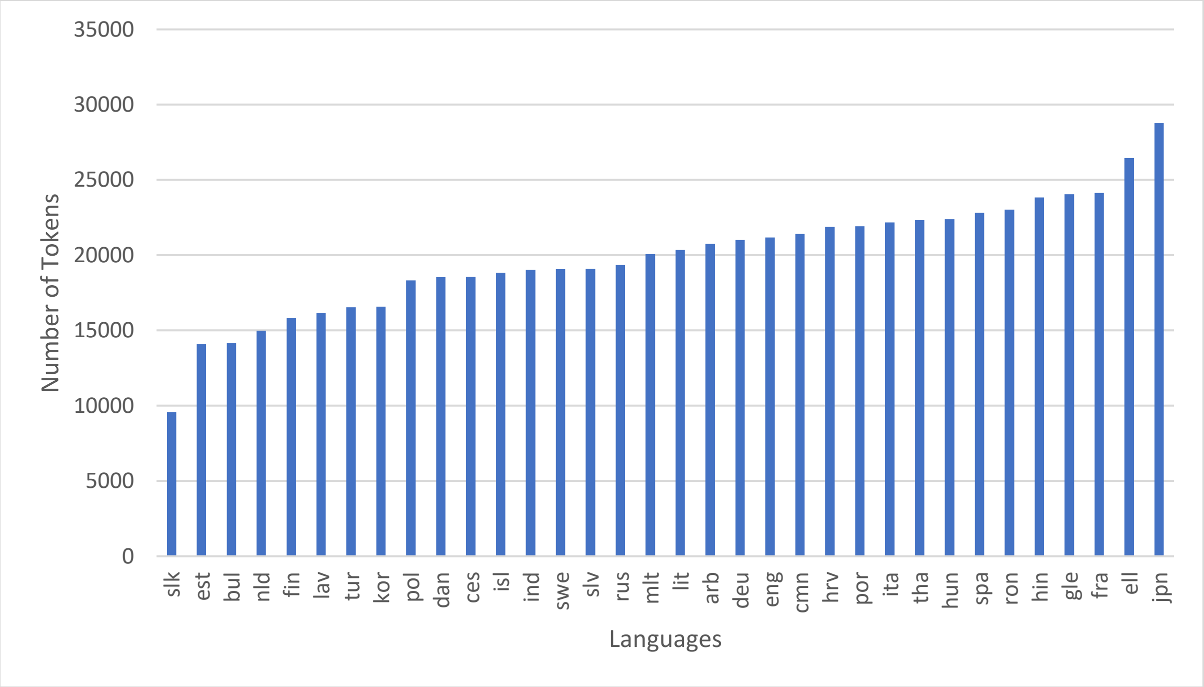
Language	**Test set**	**Genres**
bul	BTB	fiction, legal, news
hrv	SET	news, web, wiki
dan	DDT	fiction, news, nonfiction, spoken
nld	Alpino	news
est	EDT	academic, fiction, news, nonfiction
ell	GDT	news, spoken, wiki
hun	Szeged	news
gle	IDT	fiction, government, legal, news, web
lav	LVTB	academic, fiction, legal, news, spoken
lit	ALKSNIS	fiction, legal, news, and nonfiction genres
mlt	MUDT	fiction, legal, news, nonfiction, wiki
ron	RRT	academic, fiction, legal, news, nonfiction, wiki
slk	SNK	fiction, news, nonfiction
slv	SSJ	fiction, news, nonfiction

Table 6.4. UD corpora genre information for non-PUD EU languages.

It is possible to notice that genres vary a lot, however, in all of them, there is at least some Wikipedia and/or news texts. The problem is that there is no specific index in the corpora which allows separating the sentences according to genres. This is most problematic for Danish, Greek, and Latvian as these languages have some portion of spoken transcriptions as part of their corpus. Therefore, genre bias cannot be totally avoided.

To compose each typological corpus, we developed a simple Python script that randomly selects 1,000 sentences from each training-set of the corpora presented in Table 6.4. For Hungarian, as the Szeged training corpus has less than 1,000 sentences, we completed the typological corpus with 90 random sentences from the development-set. The language-collection built for the typological study of the EU languages is presented in Annex 76, and in Figure 6.5, we present the overview of the languages regarding the number of tokens of the obtained corpora.

Figure 6.5. Graph representing the EU and PUD languages in ascendant order regarding the number of tokens of their typological corpus.



Slovak has the smallest corpus (9,582 tokens), while Japanese has the largest one (28,784 tokens). Moreover, for 3 languages (i.e.: Estonian, Bulgarian, and Dutch), the correspondent typological corpus has between 10,000 and 15,000 tokens, 12 languages have a typological corpus with a size comprised between 15,000 and 20,000, 16 languages have larger typological corpus with 20,000 to 25,000 tokens, and 2 corpora have more than 25,000 tokens (Greek and Japanese).

Only 10 out of the 34 corpora are parallel, thus, it is not possible to conduct the same analysis as in Section 4.1.2. All corpora have 1,000 sentences, but the semantic content is not the same for all of them. It is possible to notice that agglutinative languages such as Estonian, Finnish, and Turkish tend to present fewer tokens, however, Hungarian has a corpus with a relatively large size (22,396 tokens). All Romance languages have more than 20,000 tokens, and that is even the case of Romanian (non-PUD) which is more synthetic in comparison to the other languages of this genus. On the other hand, Slavic and Germanic languages present a large variety in terms of corpus-size. The size of the Maltese corpus is relatively close to the Arabic one. Considering the other 2 EU languages which do not have a close-related language in the established language-set regarding in terms of the genus (i.e.: Greek and Irish), their corpus is composed of a large number of tokens, especially Greek with more than 25,000 tokens.

Therefore, it is clear that the usage of non-parallel corpora generates some bias in the typological analysis which needs to be considered in the following sub-sections. Moreover, the robustness of the developed typological methods for dependency parsing improvement will be checked with the experiments that will be conducted for the selected low-resourced EU languages and Croatian.

Besides the number of tokens, it is also possible to characterize the typological corpora in terms of the labels regarding the part-of-speech and the dependency parsing annotations[40].

Regarding the part-of-speech tags (UPOS), it is possible to notice that the ensemble of labels present in all PUD and EU corpora is the same as the ones identified when only PUD languages were scrutinized (10 in total as presented in Table 4.6). The number of UPOS labels present in the typological corpora together with the list of tags from the UD UPOS list that are not attested in them are displayed in Table 6.5.

Of all the 34 languages, 27 present all 17 possible UPOS labels or 16 of them. For 6 languages, 15 labels describe the whole corpus regarding part-of-speech, and Korean present only 13 UPOS labels.

In most cases where not all UPOS labels are used, the missing tags concern interjections (INTJ), symbols (SYM), particles (PART), and words that for some reason cannot be assigned a real part-of-speech category (X). Finnish is the only language without determiners (DET), and Korean presents more specificities as it does not contain any subordinating conjunctions (SCONJ) or adpositions (ADP).

---

[40] In the selected corpus-based approaches, these are the labels that are considered to determine the different features. The corpora can also be compared in terms of morphosyntactic features as presented in Section 4.1.2, however, this information is less pertinent in this study.

Language	Number of POS labels	Labels not present in corpus
arb	16	INTJ
bul	15	X, SYM
cmn	15	INTJ, SYM
hrv	17	-
ces	15	X, INTJ
dan	17	-
nld	16	PART
eng	17	-
est	16	PART
fin	15	PART, DET
fra	16	INTJ
deu	16	INTJ
ell	16	INTJ
hin	16	INTJ
hun	16	SYM
isl	17	-
ind	17	-
gle	17	-
ita	16	INTJ
jpn	16	X
kor	13	SYM, SCONJ, INTJ, ADP
lav	17	-
lit	17	-
mlt	17	-
pol	16	INTJ
por	16	PART
ron	16	SYM
rus	17	-
slk	16	SYM
slv	16	SYM
spa	15	PART, INTJ
swe	16	X
tha	15	X, INTJ
tur	16	PART

Table 6.5. Number of attedted UPOS labels and list of tags that are not present in each typological corpus.

In terms of DEPREL labels, in total 144 different labels are used to describe all possible syntactic relations in the language-set: 37 tags correspond to a DEPREL type and 107 are formed by a type and a sub-type. This number is higher than the one observed when only PUD languages were analysed (i.e.: 110 DEPREL tags). The complete list of dependency relation labels is presented in Annex 77.

Moreover, in Section 4.1.2, we presented the 15 DEPREL tags which are present in all PUD corpora (Table 4.9). From these labels (which are formed only by types), 13 are present in all corpora composing the PUD and EU collection. The adjectival modifier (amod) is not present in the Hungarian corpus as a label formed only by a type, however, this corpus contains DEPREL labels formed by "amod" type and sub-types. Likewise, the fixed multiword expression (fixed) is not encountered in the Hungarian and Lithuanian corpora, but in this case, no combination of this label with subtypes exists.

Table 6.6 presents the details concerning the number of DEPREL labels and the specific tags of each corpus selected for the typological analysis. When we consider the number of different tags (i.e.: type and possible subtype), Japanese has the lowest amount (25) and is followed by 3 Slavic languages: Slovenian (31), Bulgarian (33), and Croatian (34). The other 2 Slavic languages from this language-set present a larger number of DEPREL labels[41]: Russian with 39, Czech with 43, and Polish with 59 (i.e.: the largest number of all considered languages). The number of tags concerning Romance languages vary between 40 (Italian) to 48 (Romanian). The same discrepancy observed with Slavic languages also exists for Germanic ones, Danish and Dutch relations are described with 34 and 36 labels respectively, while German has 47 labels and English, 48. It is the same case regarding the Uralic family, Estonian has 36 DEPREL tags, while Finnish has a larger set (44), and Hungarian the second largest one of this language-set (54). Latvian and Lithuanian (both from the Baltic genus) have a similar number of labels, 37 and 36 respectively. The two languages from the Afro-Asiatic family have DEPREL tag-sets with more than 40 labels: Maltese with 47 labels and Arabic with 42.

It is interesting to notice that the major differences occur in terms of the usage of sub-types. When only types are considered, again Japanese is the language with the lowest number of tags (25). Besides Japanese, 9 other languages have DEPREL sets with less than 30 sub-types: Korean, Slovenian, Hindi, Greek, Polish, Bulgarian, Danish, Dutch, and Lithuanian. Some of these languages also present a low number of DEPREL composed of types and sub-types, however, Polish has the largest number when sub-types are considered, but has a relatively small number of types in its DEPREL tag-set (28). The majority of languages (22 out of 34) use from 30 to 35 labels, and 2 languages have a tag-set formed by 36 types: Maltese and English.

---

[41] The number of DEPREL labels depends not only on the language but also on the choice of the creators of the corpora, particularly concerning sub-types. Different corpora from different languages may present some differences regarding these labels.

Language	Number of DEPREL	Number of types	Number of sub-types	Specific DEPREL label
arb	42	34	8	-
bul	33	29	4	-
cmn	44	32	12	mark:adv, mark:relcl, obl:patient, discourse:sp, case:loc
hrv	34	32	2	-
ces	43	31	12	-
dan	34	29	5	obl:loc
nld	36	29	7	-
eng	48	36	12	nmod:npmod, obl:npmod
est	36	30	6	-
fin	44	30	14	nmod:gobj, xcomp:ds, nmod:gsubj, cop:own, compound:nn
fra	45	31	14	obj:agent, aux:tense, aux:caus, obl:mod, nsubj:caus, expl:comp, expl:subj
deu	45	33	12	-
ell	38	33	5	-
hin	38	28	10	compound:conjv
hun	54	31	23	advmod:to, advmod:locy, nmod:attlvc, amod:mode, advmod:obl, advmod:mode, nmod:att, advmod:tlocy, obj:lvc, advmod:tfrom, nmod:obl, amod:obl, advmod:tto, ccomp:pred, nsubj:lvc, amod:att, ccomp:obl, advmod:que, compound:preverb, amod:attlvc, nmod:obllvc
isl	36	31	5	-
ind	47	33	14	case:adv, compound:a, nmod:lmod
gle	40	28	12	obl:prep, case:voc, csubj:cleft
ita	40	33	7	-
jpn	25	25	0	-
kor	34	26	8	dep:prt
lav	37	32	5	-
lit	36	29	7	-
mlt	47	36	11	aux:part, case:det, aux:neg, cop:expl
pol	59	28	31	advcl:relcl, amod:flat, parataxis:obj, nmod:arg, aux:clitic, xcomp:subj, ccomp:cleft, aux:cnd, nmod:flat, obl:cmpr, advmod:arg, nmod:pred, parataxis:insert
por	42	33	9	-
ron	48	34	14	advcl:tcl, expl:poss, ccomp:pmod, advmod:tmod, nmod:agent, nmod:pmod
rus	39	31	8	nummod:entity
slk	41	32	9	-
slv	31	28	3	-
spa	41	32	9	-
swe	42	33	9	acl:cleft
tha	43	33	10	obl:poss
tur	41	34	7	aux:q

Table 6.6. Number of DEPREL labels present in each PUD corpus and specific tags used only in the respective corpus.

Following this analytical analysis of the corpora composition, we proceeded with the typological analysis which will be presented in the following sub-sections.

6.3.1. MarsaGram linear properties (cosine)

The typological classification obtained using the comparison of language vectors (with cosine distance) built with the linear properties extracted via the MarsaGram tool (i.e.: word order patterns between two components of the same subtree) was identified as the corpus-based method that provides the highest number of moderate and strong correlations between language distances and dependency parsing improvements when languages are combined (in terms of LAS and MLAS).

Although this strategy does not explain the ensemble of the observed phenomena in terms of dependency parsing synergy when languages are combined in pairs, it provides the best insight specially for languages that have close-related pairs in terms of genealogical features. For many problematic cases, it was possible to observe that when the MarsaGram linear properties did not provide a moderate or strong correlation, better results were obtained either considering all MarsaGram patterns or the Verb and Object position strategy.

As the statistical data regarding the linear phenomena extracted with MarsaGram provides valuable typological information for dependency parsing improvement experiments, we analysed the EU languages using this method.

The methodology applied is analogous to the one presented in Section 4.5 and the analysed corpora correspond to the typological ones described previously. As it was observed when only PUD languages were studied, the number of extracted patterns varies considerably as it can be observed in Figure 6.6 and in Annex 78.

The number of linear patterns varies from 1,049 for Slovak to 2,850 for Irish. Moreover:

- 5 languages have from 1,000 to 1,500 patterns;
- 9 languages have from 1,500 to 2,000 patterns;
- 10 languages have from 2,000 to 2,500 patterns;
- 10 languages have from 2,000 to 3,000 patterns.

Thus, for the majority of languages (59%), at least 2,000 linear properties were extracted. In some cases, it is possible to notice that languages with smaller corpora tend to provide fewer patterns, but the size of the corpora is not linearly correlated with the number of identified properties as it can be seen in Figure 6.7.

Figure 6.6. Graph representing the number of MarsaGram linear patterns extracted from the typological corpora (EU and PUD).

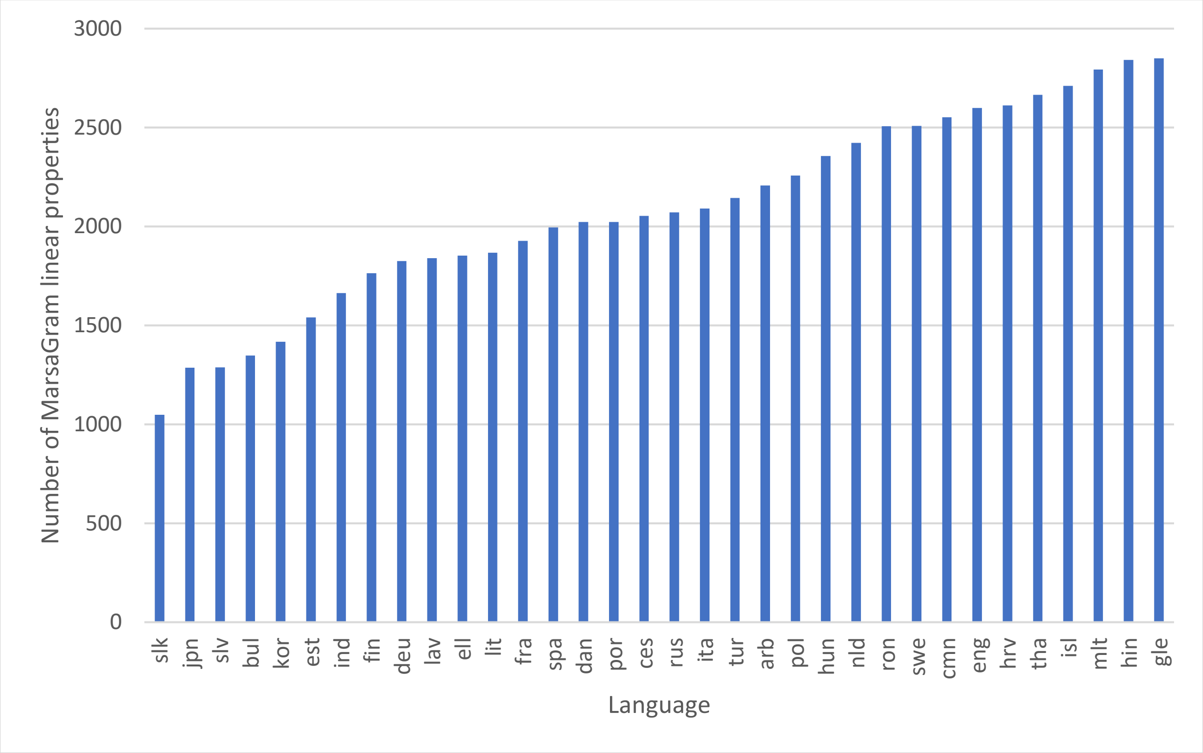
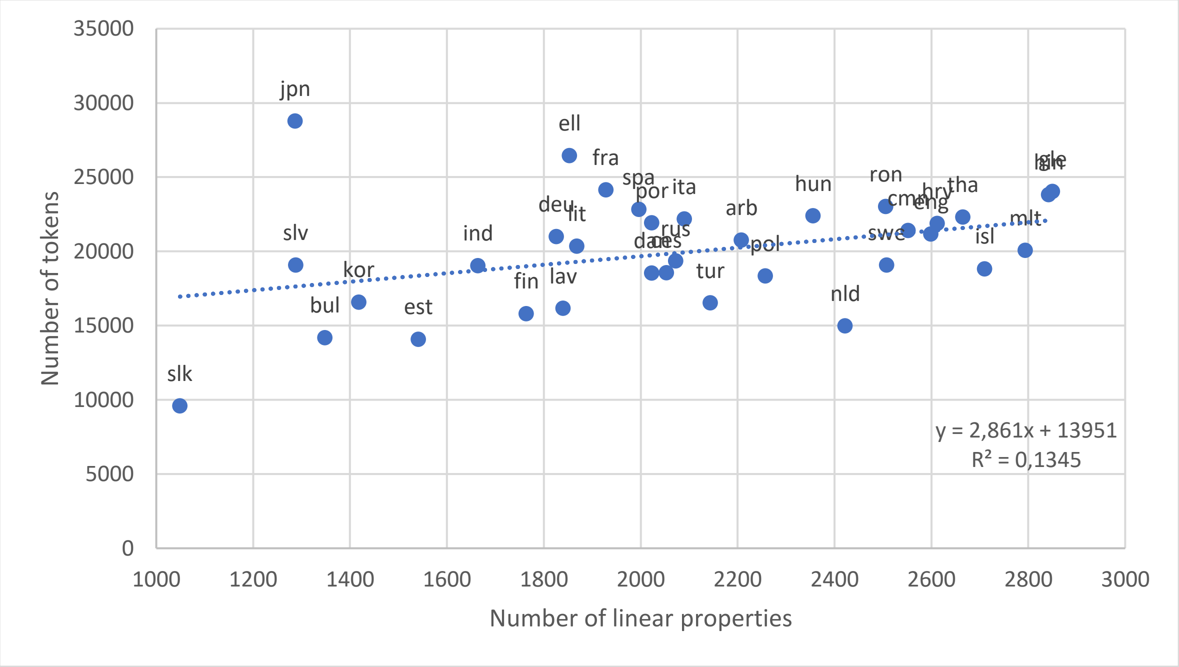


Figure 6.7. Graph representing the number of extracted linear patterns in relation to the corpora size.



Slovak has the smallest corpus and the smallest set of linear properties; however, the Japanese corpus has the highest number of tokens but it provides a relatively small number of patterns.

It is also the case for Greek and French. Moreover, if there was a linear correlation, Dutch should provide fewer patterns than it does in our experiment. The coefficient of determination (R^2) for the linear regression provided by Excel software is quite low (0.1345), thus showing that these two variables are not linearly correlated.

The total number of different MarsaGram linear patterns obtained from all PUD and EU corpora is 31,339. From these patterns, only two are present in all corpora:

- VERB-+_precede_CCONJ-cc_* - this means that in a subtree whose head is a verb, the coordinating conjunction (CCONJ) precedes the head (*) and is linked to it via the coordination dependency relation (cc).
- NOUN-+_precede_*_NOUN-appos – in this case, the head of the subtree is a NOUN and another NOUN is positioned after the head (*) and is an appositional modifier[42] (appos).

These two properties were identified previously as common features when only PUD languages were considered together with other common patterns. However, with this larger language-set, some properties regarding the word-order of subjects and objects in a subtree ruled by a verb are no longer common to all languages as they were for all PUD ones (e.g.: VERB-+_precede_NOUN-nsubj_NOUN-obj, as presented in Table 4.24).

Table 6.7 presents the distribution of linear properties inside the language-set. Again, the large majority of properties (67.67%) appear in just 1 corpus of the language-set and only a few are present in more than half of the corpora (1.25%).

	Number of patterns	**%**
Occurring in only one corpus	21,207	67.67
Occurring in more than 17 corpora	392	1.25
Occurring in all corpora	2	0.01

Table 6.7. Distribution of linear patterns inside EU and PUD corpora and respective % of the total extracted patterns.

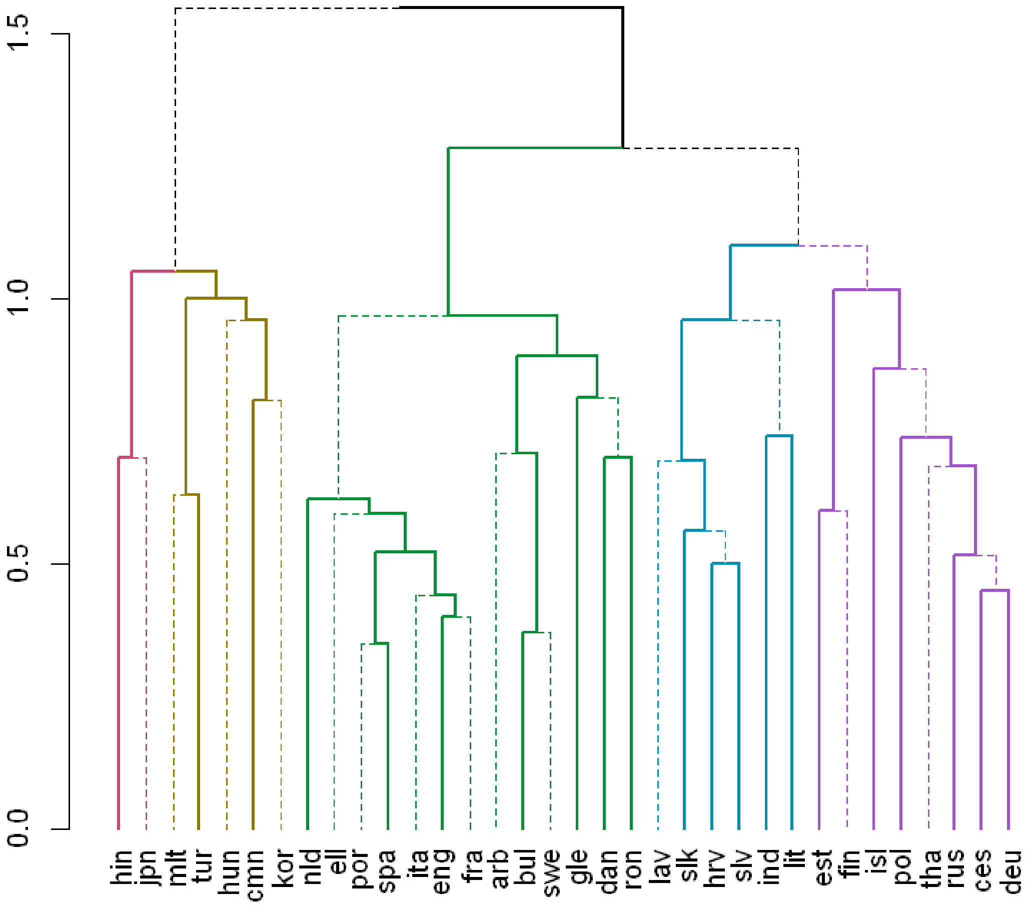
---

[42] The fact of this specific pattern being attested in all corpora shows that it is coherent with the definition of appositional modifier (appos) provided by the Universal Dependencies framework: “An appositional modifier of a noun is a nominal immediately following the first noun that serves to define, modify, name, or describe that noun”.

Hence, with the ensemble of MarsaGram linear properties, it was possible to generate the language vectors which were compared to build the cosine dissimilarity matrix (Annex 79) that was used to create the dendrogram displayed in Figure 6.8.

In Figure 4.23 of Section 4.5, we presented the dendrogram composed via the comparison of the PUD language vectors built with the MarsaGram linear properties (cosine distance). As expected, PUD languages are overall classed in the same way in the dendrogram displayed in Figure 6.8. One major difference is the position of Indonesian, while it was positioned close to Arabic, Swedish, and Romance languages when only PUD languages were analysed, in this new dendrogram composed of PUD and EU languages, Indonesian is presented close to Lithuanian and some Slavic languages (Slovenian, Croatian, and Slovak). Another visible change concerns Arabic which now forms a sub-cluster with Bulgarian, and Swedish.

Figure 6.8. Cluster dendrogram obtained from the cosine dissimilarity matrix calculated with the comparison of the PUD and EU MarsaGram language vectors (linear patterns).



When we consider each phylogenetic family and genus in the language-set composed of EU languages, it is possible to notice that:

1) Regarding the Indo-European family:
    a. Baltic genus: Lithuanian and Latvian do not form a specific sub-cluster although belonging to the same genus. As previously mentioned, Lithuanian is grouped with Indonesian. On the other hand, Latvian is an isolated branch of a cluster composed of some Slavic languages (Slovak, Croatian, and Slovenian). However, all the languages listed above form a whole group in the dendrogram (in blue). Hawkins (1983) did not provide information regarding Latvian, while Lithuanian is classed as type 11 (together with Danish and Swedish), however, in this dendrogram, these languages are not all clustered in the same group. When lang2vec features are analysed, Latvian and Lithuanian have 69 common syntactic features, and for 55 of them (78.8%) the correspondent values are exactly the same. The main word order differences concern the existence of SOV, VSO, OVS orders (0.5) in Lithuanian, but they are not attested in Latvian. Moreover, while in Latvian the adpositions and possessors are always before nouns, in Lithuanian they are also observed after. When Lithuanian is compared with Indonesian, from the 73 common lang2vec features, 44 have the same value (60.3%). Moreover, it is possible to notice that the Indonesian has the same values as Latvian for the features listed above (for which Latvian and Lithuanian differ). Slovak does not have lang2vec syntactic description, thus, the comparison between Latvian and this language cannot be done.
    b. Celtic genus: the only language from this genus in our language-set is Irish. Hawkins (1983) classified it as type 1 which is the same group as Arabic. Although these 2 languages do not form an isolated sub-cluster in the dendrogram, they are part of the same sub-group of the large green cluster. This group also contains Bulgarian, Swedish, Danish, and Romanian. Irish is classified closer to sub-cluster composed of the last 2 languages listed above. Irish and Danish have 42 lang2vec syntactic features with the same value, and 19 with differences. While Irish is a VSO, Danish is SVO. Also, in Irish, possessors and demonstrative words are positioned after the noun (in Danish they come before it). Adjectives in Irish are usually positioned after the nouns

but can also appear before, while in Danish, only the order “adjective before nouns” is attested.

c. Germanic genus: the languages of this specific genus do not form any exclusive sub-cluster. Instead, they are positioned in the dendrogram with other VO languages. Danish form a sub-cluster with Romanian, and Dutch is positioned in the extreme left of the green group in an isolated branch close to a cluster formed by Greek, some Romance languages, and English. This latter composes a sub-group with French and is relatively close to Italian. Swedish is also part of the large green group, but it is positioned in another sub-group, forming a specific sub-cluster with Bulgarian. German and Icelandic are the 2 Germanic languages that are not classified in the green group, being part of the purple cluster on the right side of the dendrogram which also contains some Slavic, Thai, and the 2 Finnic languages (Uralic family) of the language-set. German is positioned in the extreme right, forming a sub-cluster with Czech, while Icelandic is a single branch of the purple cluster closer to the Finnic languages and Polish. Regarding lang2vec syntactic features, from all the 60 common ones, 42 present the same value for all Germanic languages (70%).

d. Greek genus: Greek is the only language from this genus in the selected language-set. In the dendrogram, it is part of the green group as an isolated branch on the left side of a sub-cluster formed by Romance languages (except for Romanian), English, and Dutch. In Hawkins (1983) classification, Greek is considered as type 10 (same as Czech, Dutch, Icelandic, and Russian). Of these languages, only Dutch is relatively close to Greek in the dendrogram.

e. Romance genus: Regarding the Romance languages, the classification of the 4 PUD languages (i.e.: French, Italian, Portuguese, and Spanish) is similar to what was observed in Figure 4.23 (Section 4.5). These languages form a specific sub-cluster of the green group which also includes English. Romanian is not part of PUD collection, and although being part of the green group, it is not close to the other Romance languages. Instead, it forms a specific sub-group with Danish and is also relatively close to Irish. The analysed Romance languages have 63 common lang2vec syntactic features of which 44 have the same value (69.8%).

    Romanian differs from all the other Romance languages in terms of the following features:

        i. S_VSO : 0.333;
        ii. S_DEFINITE_AFFIX: 1.0;
        iii. S_DEFINITE_WORD: 0.0;
        iv. S_DEMONSTRATIVE_WORD_AFTER_NOUN: 1.0;
        v. S_CASE_SUFFIX: 1.0.

    f. Slavic genus: Regarding the 7 Slavic languages of our language-set, it is possible to identify in the dendrogram a specific sub-cluster formed by Slovak, Croatian, and Slovenian (these two last ones forming a specific sub-group on their own). Bulgarian is part of a different group, forming a sub-cluster with Swedish. The other 3 Slavic languages (i.e.: Polish, Russian, and Czech) are part of the purple group. Czech and Russian show more similarity, as they form a specific sub-group (together with German). On the other hand, Polish is positioned as a single branch in the middle of the purple cluster between Thai and Icelandic. It is not possible to compare all the Slavic languages regarding the lang2vec features as there is no information for Slovak in this database. If Slovak is excluded, the other Slavic languages have only 12 common features (with the same value for 8 of them which concern subject, verb, and object positions). Of all the 7 Slavic languages, only 2 are described by Hawkins (1983): Czech and Russian, both considered as type 10.

2) Concerning non-Indo-European languages:

    a. Uralic family: From this family, there are 2 languages in our language-set from the Finnic genus (i.e.: Estonian and Finnish), and 1 from the Ugric genus (i.e.: Hungarian). While the Finnic languages are part of the same sub-cluster of the purple group, Hungarian is positioned on the left side of the dendrogram, closer to Chinese and Korean. Finnish and Estonian are also part of the same language-type (15) proposed by Hawkins (1983), however, there is no information regarding Hungarian. In terms of lang2vec syntactic features, the Uralic languages have 68 common features of which 56 have the same value (82.3%).

Hungarian differs from the Finnic languages in the following features:

   i. S_SOV: 0.6667;
   ii. S_VSO: 0.3333;
   iii. S_OBJECT_BEFORE_VERB: 0.6667;
   iv. S_DEFINITE_WORD: 1.0;
   v. S_INDEFINITE_WORD: 1.0;
   vi. S_ADPOSITION_BEFORE_NOUN: 0.0;
   vii. S_RELATIVE_BEFORE_NOUN: 1.0;
   viii. S_ANY_AGREEMENT_ON_ADJECTIVES: 0.0;
   ix. S_COMPLEMENTIZER_WORD_AFTER_CLAUSE: 1.0.

b. Afro-Asiatic family: In the official European Union language-set, the only Afro-Asiatic language is Maltese (Semitic genus). If we consider the extended language-set (EU and PUD), another language from the same family and genus is present: Arabic. Although being part of the same phylogenetic group, Maltese is not close to Arabic in the dendrogram. It is positioned in a sub-cluster on the left-side of the Figure 6.8 with Turkish and is closer to many other OV languages (although being VO). Maltese is not part of the analysis proposed by Hawkins, and in terms of lang2vec features, it has values for only 6 features that are not common to Arabic.

Thus, it is possible to notice that the dendrogram composed of the clustering analysis of the data provided by MarsaGram regarding linear patterns generates a specific classification of EU languages which present some variance to what would be expected in terms of the phylogenetic characteristics (e.g.: the distribution of Germanic and Slavic languages, the distance between Arabic and Maltese, and between Hungarian and the other Uralic languages). Although lang2vec can be used to check some of the main syntactic features which are shared from languages from the same genealogical family, it is not possible to conduct an overall analysis of all 34 languages in some cases there are only a few features (or no feature at all) which are described.

Regarding the identification of the closest languages to be combined with the low-resourced ones (and Croatian) for the experiments of dependency parsing improvement, by analysing the dissimilarity matrix (Annex 79), it is possible to determine the best language pairs. The idea is to consider for each low-resourced language (and Croatian), the language with the lowest distance and other possible candidates with a distance similar to the lowest one (i.e.: with a delta regarding the distance value compared to the best choice not higher than 10% of the lowest distance). Thus, the following pairs were identified:

- Croatian: Slovenian (best), Russian, English, and German;
- Hungarian: Greek (best) and French;
- Irish: French (best);
- Lithuanian: Croatian (best), and Portuguese;
- Maltese: Croatian (best), Slovenian, and English.

By comparing the selection of the language pairs with the dendrogram (Figure 6.8), it is possible to notice that the closest languages are not necessarily adjacent in the clustering graph. The main reason for that is that the clustering algorithm establishes the clusters analysing all the provided data from the dissimilarity matrix to calculate the distance between the clusters and sub-clusters. In this precise analysis for the selection of the most optimized pairs, the distance values are considered individually.

If the dendrogram is considered to find the best language association, it is possible to determine the best pair in cases where the language forms a specific sub-cluster with another one. It is the case for Croatian, Lithuanian, and Maltese. Croatian is clustered with Slovenian, which is already considered the best possible language combination. Lithuanian forms a sub-cluster with Indonesian, and Maltese with Turkish. Thus, for these two languages, we will also consider these possibilities for the dependency parsing experiments.

6.3.2. Combination of MarsaGram all properties and Head and Dependent (Euclidean)

As presented in Section 5.3.3, another optimized typological method to identify language pairs to improve dependency parsing results concerns the association of the language distances provided by the analysis of all MarsaGram properties with the ones generated via the head and dependent method (both with Euclidean distances) with a specific formula obtained with the linear regression experiments:

$$(6.1)\ D_{opt} = 0.036\ D_{Marsagram\ all} + 0.397\ D_{head\ and\ dependent}$$

Thus, in this sub-section, we will present an overall analysis of both methods separately followed by a detailed study of the dendrogram concerning their association and the list of identified language-pairs that will be used in the dependency parsing experiments.

When the MarsaGram tool is used to extract all possible patterns (i.e.: linear, exclude, require, and unicity), the total number of features extracted in all 34 PUD and EU corpora is 240,882 (i.e.: 82,127 more patterns than the ones extracted from PUD languages only). The language with the largest number of attested patterns is Irish (25,989), while Japanese has the lowest number (5,226) as presented in Figure 6.9 and Annex 80.

The discrepancy in terms of the number of extracted patterns in PUD and EU languages is much higher when all patterns are analysed when compared to the linear ones presented in the previous sub-section. For MarsaGram all properties:

- 5 languages have from 5,000 to 10,000 patterns;
- 11 languages have from 10,000 to 15,000 patterns;
- 9 languages have from 15,000 to 20,000 patterns;
- 8 languages have from 20,000 to 25,000 patterns;
- 1 language (Irish) has more than 25,000 patterns.

It is possible to notice that although the ascendant order is not the same when we compare the graphs obtained for linear and for all patterns, in general, languages that have lower numbers of linear patterns also tend to have lower numbers when all properties are considered. In both cases, the 5 languages with the lowest number of patterns are Japanese, Slovak, Slovenian, Korean, and Bulgarian. On the other hand, it is not necessarily the case for the languages with the highest number of extracted patterns. Irish, Maltese, Hindi, and Icelandic are in the top of the list regarding the languages with the largest number of patterns in both scenarios. However, while Thai has 4th largest number of linear patterns when all properties are considered, it is in the 11th position of the list. The inverse phenomenon can be observed for Hungarian, Romanian, and Dutch which are better positioned in the list regarding all patterns.

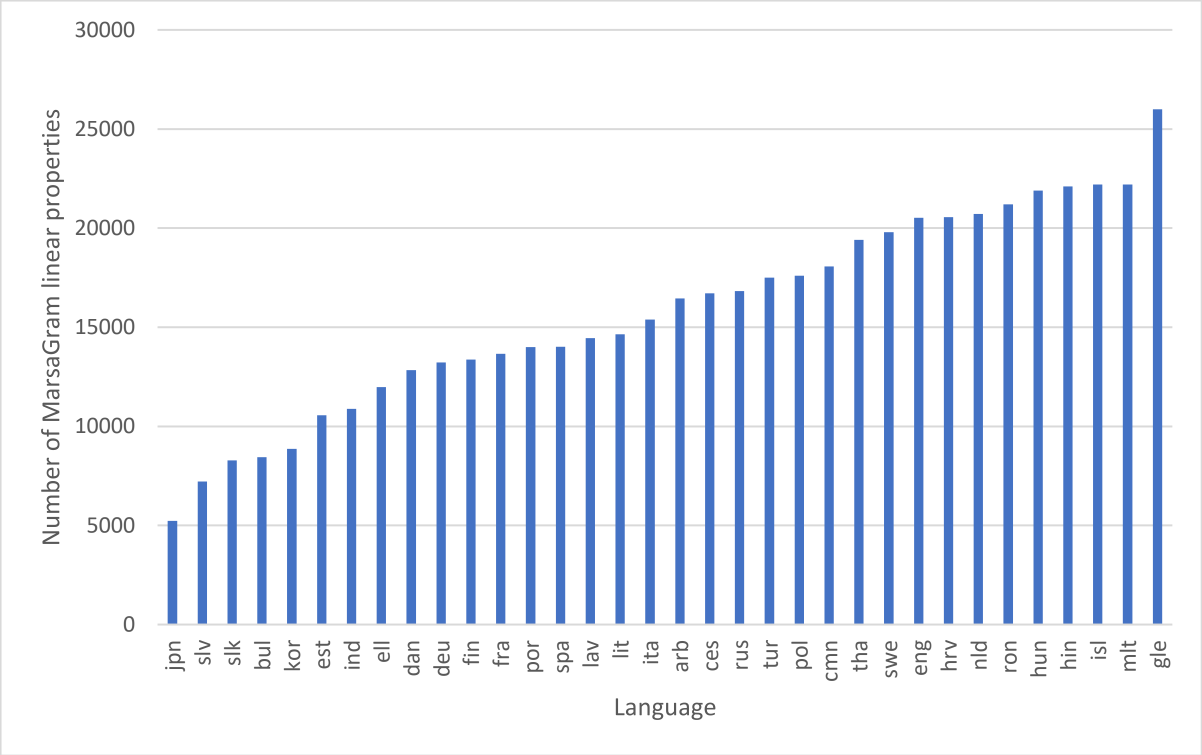
Moreover, as seen in the analysis of the linear patterns, the graphic representing the relation between the number of patterns (all properties) and the number of tokens shows that there is no linear correlation between these two variables (Annex 81).

As previously explained, when all MarsaGram patterns are extracted, 4 different properties are considered. The overall distribution of the attested patterns in terms of these patterns is presented in Table 6.8.

Type of property	Number of patterns	%
Precede (linear)	31,339	13.01
Exclude	196,761	81.68
Unicity	2,962	1.23
Require	9,820	4.08

Table 6.8. Distribution of the final set of MarsaGram properties in terms of property types.

Figure 6.9. Graph representing the number of MarsaGram patterns (all properties) extracted from the typological corpora (EU and PUD).



It is possible to observe that the distribution is similar to the one obtained when only PUD languages were analysed (as presented in Table 4.20, Section 4.5). The vast majority of patterns concern the “exclude” property, while the “linear” one, which corresponds precisely to the patterns related to the word-order inside subtrees, represents only 13% of the extracted features.

The Table 6.9 presents the distribution of the properties inside the 34 corpora. As was the case when PUD languages were analysed, most patterns occur in only one corpus. Around 1% occurs in more than half of the languages, and the number of patterns present in all corpora is very low.

	Number of properties	**%**
Occurring in only one corpus	163,226	67.76
Occurring in more than 17 corpora	2,643	1.10
Occurring in all corpora	29	0.01

Table 6.9. Distribution of MarsaGram all properties inside PUD and EU corpora and the respective % of the total selected properties.

Regarding the 29 patterns identified in all 34 languages, the complete list is presented in Annex 82. From this list, the two linear properties were described in the previous subsection, 13 patterns describe “exclude” properties, 14 concern “unicity” ones, and no common “require” pattern is identified. Most of these common patterns concern subtrees ruled by a verb (20), while 7 are governed by a noun, and 2 by a proper noun.

With the 240,882 identified patterns, we generated the language vectors that were used to calculate the language distances (Euclidean) which compose the dissimilarity matrix (Annex 83). This matrix is, then, combined with the one obtained with the head and dependent strategy (using the formula described previously).

When the 34 PUD and EU languages are analysed in terms of head and dependent relative order in the sentences (following the same methodology as presented in Section 4.6), it is possible to extract a total of 4,062 features: in 1,859 cases (45.8%) the pattern corresponds to phenomena where dependent precedes the head (left-branching), and in 2,202 patterns (54.2%), the head comes before the dependent in the sentence (right-branching).

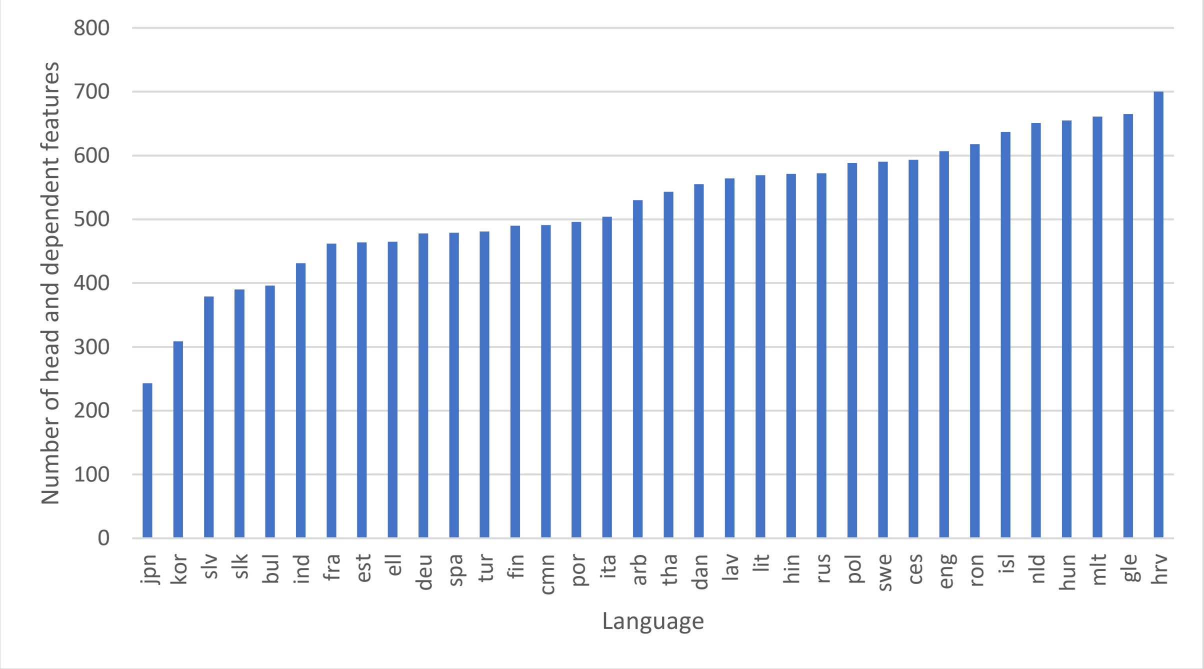
As expected, the languages present some variation in terms of the number of head and dependent attested phenomena as it presented in Figure 6.10 and in Annex 85. It is possible to notice that some languages presenting a large number of MarsaGram patterns also present a considerable set of head directionality features (e.g.: Irish, Maltese, and Hungarian). Croatian is the one with the largest number of different head and dependent attested phenomena (700). Moreover, languages presenting a low number of MarsaGram patterns, also present small sets of head directionality features (e.g.: Japanese, Slovak, Korean, and Slovenian). These facts show that, as was the case for the MarsaGram patterns, there is no linear correlation between the number of head directionality features and the number of tokens composing each corpus.

The overall distribution of the head and dependent features in the 34 corpora of our language-set is presented in Table 6.10. Half of the attested phenomena happen in only one language, and only around 6% occur in more than half of the selected languages.

	Number of patterns	**%**
Occurring in only one corpus	2,046	50.37
Occurring in more than 17 corpora	251	6.18
Occurring in all corpora	21	0.52

Table 6.10. Distribution of head and dependent word-order patterns inside PUD and EU corpora and the respective % of the total (4,062).

Figure 6.10. Graph representing the number of head directionality features extracted from the typological corpora (EU and PUD).



The number of head directionality features happening in all corpora is relatively low. The complete list of these word-order phenomena is presented in Annex 84. From the 21 common features, 10 concern punctuation. Besides, it is possible to identify:

- Adverbs as adverbial modifiers which in all corpora occur preceding specific heads (i.e.: adjective, noun, or verb).
- Coordinating conjunctions as coordination preceding specific heads (i.e.: noun, proper noun, or verb).
- Nouns as appositional modifiers being positioned after the nominal heads.

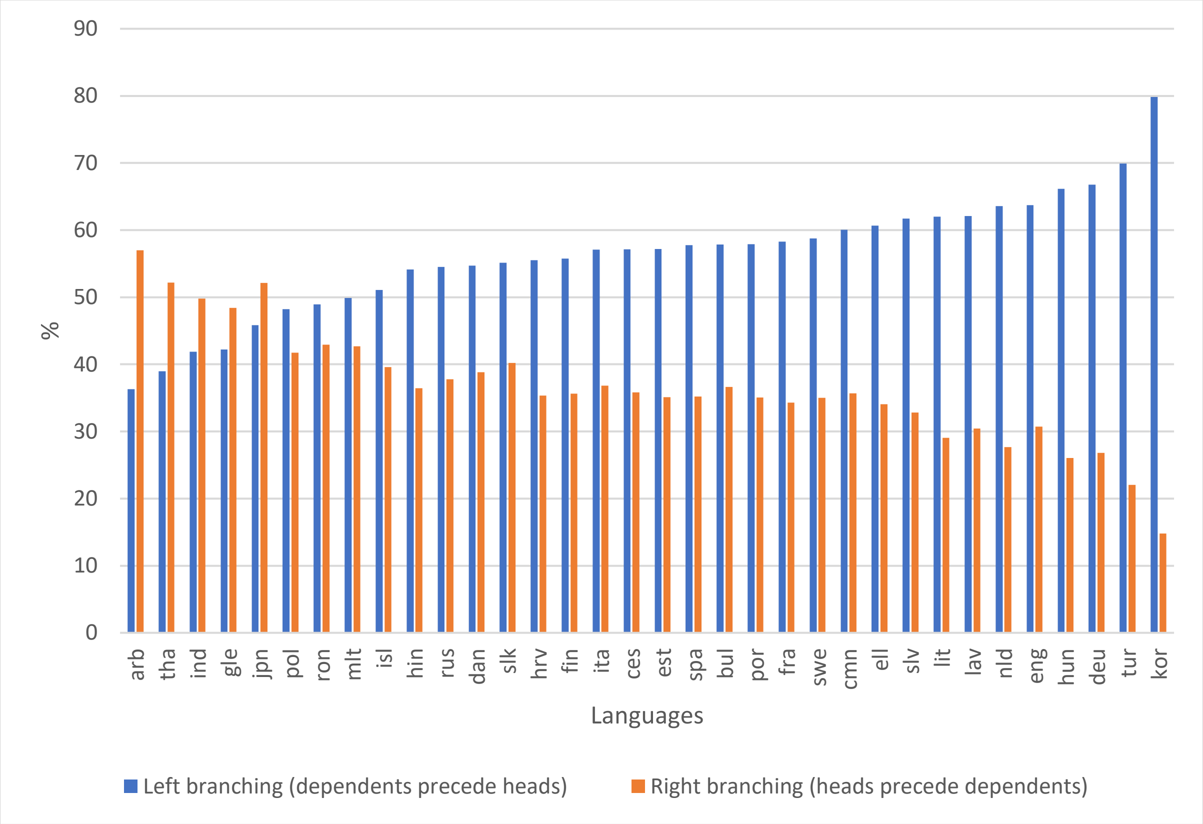
- Pronouns and proper nouns as nominal subjects (dependent) preceding verbs (head).
- Verbs as adverbial clause modifiers preceding verbs (head).

With the extracted features regarding the head and dependent positions, it is possible to analyse the overall tendency of each language concerning the head directionality (i.e.: total percentage of frequency concerning left-branching and right-branching attested phenomena in each language). Figure 6.11 displays the graph regarding this analysis (with values provided by the data presented in Annex 86). As it was the case when only PUD languages were analysed, relations that are always right-branching according to the UD guidelines were excluded (i.e.: “conj”, “appos”, “flat”, and “fixed”).

From the graph in Figure 6.11, it is possible to notice that 5 languages tend to have more right-branching relations (i.e.: dependent after the head): Arabic, Thai, Indonesian, Irish, and Japanese. Polish, Romanian, Maltese, and Icelandic have a more balanced number of left and right-branching relations when compared to other languages (with the percentage of left-branching around 50). A large number of languages (19) present more left-branching relations, but with a percentage of head preceding dependent between 30 and 40%. Korean is the one with the lowest number of right-branching relations, only 14.80%. As presented in Section 4.6, the OV languages do not necessarily present a much higher percentage of left-branching relations. It is the case for Turkish and Korean, but it is not observed for Hindi and Japanese.

Regarding EU languages, it is noticeable that Romance languages (except for Romanian) have quite similar values of left and right-branching. Germanic languages tend to have lower percentages of right-branching relations (being positioned towards the right side of the graph). From this genus, Danish and Icelandic (non-EU language) are the exceptions: Icelandic has a more balanced distribution of left and right-branching phenomena, while Danish is closer to Romanian with similar percentages of them. Regarding the Slavic languages, Polish is also part of the group composed of Romanian, Icelandic, Maltese, and Japanese. Russian, Slovak, and Croatian present more left-branching relations but a relatively high number of right-branching phenomena, followed by Czech, and Bulgarian. Slovenian is the one presenting the lowest number of cases where the head precedes the dependent. Greek is positioned close to Slovenian and the Baltic languages, with a relatively lower number of right-branching phenomena. Irish is part of the small group of languages with more right than left-branching relations, together with Arabic, Thai, Indonesian, and Japanese. The Finnic languages from the Uralic family are relatively close in the graph, but Hungarian (Ugric genus) differs, being positioned closer to the languages with a larger number of left-branching relations (e.g.: English, German, Turkish, and Korean).

Figure 6.11. Overall distribution (in terms of percentage) of right-branching (the head precedes the dependent) and left-branching (the dependent precedes the head) in EU and PUD languages.



The Euclidean dissimilarity matrix obtained with the language comparison regarding the head directionality features is presented in Annex 87. Thus, with these values and the ones calculated with the MarsaGram all properties data, it is possible to build the optimized dissimilarity matrix with the formula presented previously in this sub-section.

Before combining the results of both matrices, as the linear regression was conducted with normalized distance values, we transformed the dissimilarity matrices into normalized ones (values from 0 to 1) with the following formula which considers the maximum (max) and minimal (min) values of the dissimilarities matrices:

$$(6.2)\ Distance_{normalized}(i) = \frac{Distance(i) - \min(all\ distances)}{\max(all\ distances) - \min(all\ distances)}$$

With the normalized matrices, we generated the optimized one (Annex 88) following the proportions established via the linear regression experiments, which was then used for the generation of the corresponding dendrogram displayed in Figure 6.12.

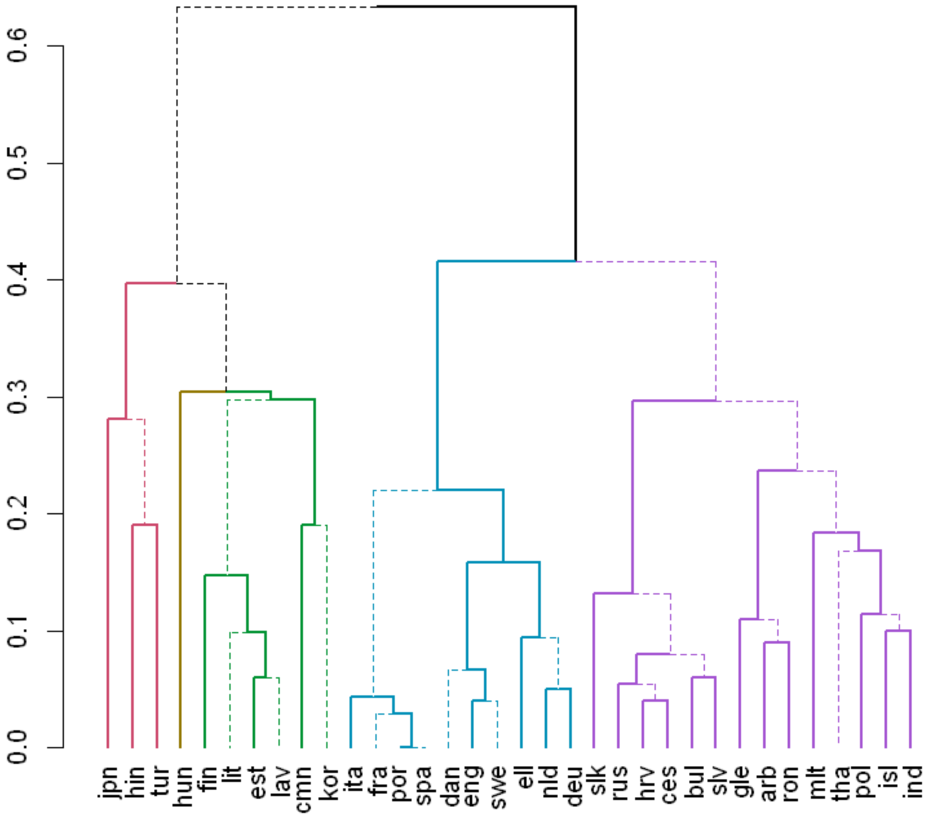
When compared to the previous dendrogram concerning MarsaGram linear properties (cosine), the one generated via the combination of MarsaGram all properties and head and dependent word order position features is much closer to the phylogenetic classification of languages. It is possible to clearly identify the Romance sub-cluster which contains all languages from this genus in the language-set with exception of Romanian which is positioned in a group with Arabic, and Irish most probably due to their proximity in terms of overall left and right-branching percentages. All 7 Slavic languages are positioned in the large purple group. While 6 of them form an exclusive Slavic sub-cluster, Polish is grouped with Icelandic and Indonesian. When the overall head directionality tendency was analysed, Polish was also identified as the one with the highest number of right-branching phenomena in the Slavic genus. It is interesting to notice that although Croatian and Slovenian are the closest Slavic languages genealogically, in this dendrogram they are part of the same cluster but not adjacent (i.e.: Croatian forms a specific sub-group with Czech, and Slovenian with Bulgarian), this also reflects the positioning of these languages in the overall analysis of the head directionality percentages.

A Germanic sub-cluster can also be identified in the large blue group, together with Greek. It is possible to identify the proximity between English, Swedish, and Danish, and between German and Dutch. From this genus, only Icelandic is part of the purple cluster, together with other languages with a balanced percentage of left and right-branching relations.

Baltic and Finnic languages form a specific mixed sub-cluster in the dendrogram which is closer to Hungarian on one side, and to Chinese and Korean on the other one. Moreover, three OV languages are grouped on the left side of the graph (i.e.: Japanese, Hindi, and Turkish). Korean is not grouped with them, forming a specific sub-group of the green cluster with Chinese (although these two languages are not close in the overall analysis of the head directionality features).

Finally, Maltese, which is from the same family and genus as Arabic, is positioned close to this language in the purple cluster but they do not form a specific sub-cluster. When both MarsaGram all properties and head and dependent positions are considered, Maltese is closer to Thai.

Figure 6.12. Cluster dendrogram obtained from the dissimilarity matrix calculated with the combination of MarsaGram all properties and head and dependent methods (Euclidean).



When the data from the optimized dissimilarity matrix is used to define the best language pairs for the low-resourced EU languages and Croatian, we have:

- Croatian: Czech (best);
- Hungarian: Dutch (best), Maltese, German, English, and Estonian;
- Irish: Romanian (best);
- Lithuanian: Latvian (best);
- Maltese: Danish (best), English, Dutch, and Swedish;

These results are quite coherent with the classification obtained via the dendrogram. Croatian forms a sub-cluster with Czech, while Irish and Lithuanian are positioned as single branches inside larger groups containing the associations described above. On the other hand, Hungarian and Maltese present results which were not expected from the analysis of the dendrogram. Thus, for these languages, we decided to include the ones which are adjacent in the dendrogram as possible pairs (i.e.: Finnish for Hungarian, and Thai for Maltese).

6.3.3. Verb and Object relative position (cosine)

Although the PUD language comparison in terms of verb and object relative position did not provide the best results in terms of correlations or of LAS and MLAS improvement when the best language pairs are selected, this method (with cosine distances) shows some interesting results in terms of LAS and MLAS correlation in cases where the MarsaGram linear (cosine) did not perform well as presented in section 5.3.3. That is why this method is also being considered in the dependency parsing improvement experiments for low-resourced EU languages and Croatian.

The verb and object method concerns the specific analysis of the head and dependent features for which the head is a verb, and the dependent DEPREL label is “obj” (object). Differently from Greenberg (1963), Vennemman (1973), Hawkins (1983), and Dryer (1982), our analysis is quantitative and includes all possible objects, not only nominal ones. In total, 17 different features where the object preceded the verb were attested, while 16 correspond to phenomena where the verb is positioned before. The distribution in the 34 corpora of these different features is detailed in Tables 6.11 and 6.12.

OV Features	Number of corpora
ADJ_obj_precedes_VERB	15 (bul, hrv, ces, dan, nld, eng, est, fin, hin, hun, jpn, lav, slv, swe, tur)
NOUN_obj:lvc_precedes_VERB	1 (hun)
SYM_obj_precedes_VERB	2 (est, deu)
ADV_obj_precedes_VERB	4 (bul, nld, hin, jpn)
CCONJ_obj_precedes_VERB	1 (swe)
SCONJ_obj_precedes_VERB	2 (isl, rus)
PRON_obj_precedes_VERB	32 (exception: cmn, ind)
PART_obj_precedes_VERB	1 (gle)
NUM_obj_precedes_VERB	9 (hrv, nld, deu, hin, hun, jpn, kor, swe, tur)
PRON_obj:agent_precedes_VERB	1 (fra)
X_obj_precedes_VERB	6 (hrv, deu, ell, hun, lit, tur)
DET_obj_precedes_VERB	13 (bul, hrv, ces, eng, deu, hin, hun, lit, mlt, pol, slk, slv, tha)
NOUN_obj_precedes_VERB	31 (exception: ind)
AUX_obj_precedes_VERB	1 (hrv)
PROPN_obj_precedes_VERB	25 (exception: arb, eng, fra, ind, gle, ita, ron, spa, swe)
ADP_obj_precedes_VERB	3 (dan, ita, por)
VERB_obj_precedes_VERB	4 (dan, nld, tha, tur)

Table 6.11. Ensemble and overall distribution of OV features extracted from PUD and EU corpora.

The analysis of the OV features (Table 6.11) shows that the most relevant ones correspond to the patterns where the dependent is either a noun, a pronoun, or a proper noun. In these cases, the number of languages attesting the phenomena described is equal or higher than 25 (73.3%). Moreover, these are the features for which OV languages have a much higher percentage than the VO ones. It is interesting to notice that some Slavic, Germanic, Baltic, and Uralic languages also attest some of the other listed OV features, while Romance languages do not (usually presenting occurrences for just 2 or 3 features). Chinese is also very limited in terms of OV phenomena (only 2 out of the 17), and Indonesian is the only language in the set without any occurrence in its corpus of an object preceding the verb.

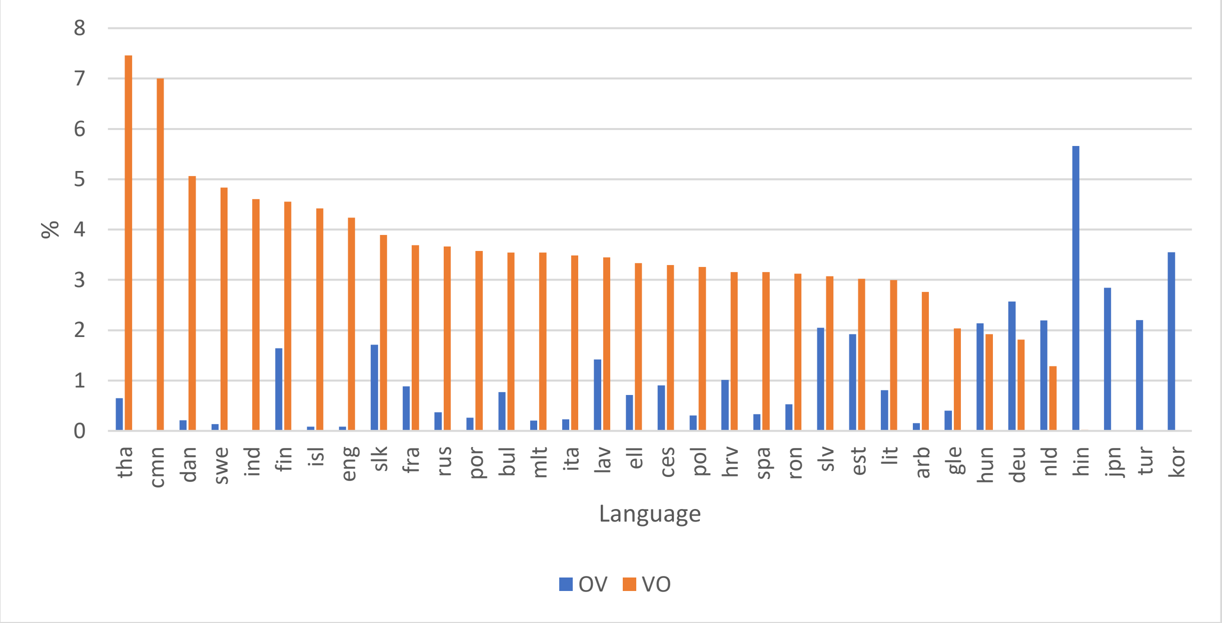
VO Features	Number of corpora
NOUN_obj:agent_follows_VERB	1 (fra)
ADV_obj_follows_VERB	21 (exception: arb, cmn, est, deu, hin, hun, jpn, kor, lav, pol, por, slv, tur)
NUM_obj_follows_VERB	26 (exception: deu, hin, hun, jpn, kor, lav, slv, tur)
DET_obj_follows_VERB	14 (bul, cmn, hrv, ces, nld, eng, deu, ita, lit, mlt, pol, ron, slk, slv)
NOUN_obj:lvc_follows_VERB	1 (hun)
ADP_obj_follows_VERB	4 (isl, gle, por, ron)
SYM_obj_follows_VERB	14 (ces, eng, fin, fra, deu, isl, ind, ita, mlt, pol, por, rus, spa, tha)
PART_obj_follows_VERB	(cmn, hrv, gle)
PRON_obj_follows_VERB	31 (exception: jpn, kor, tur)
NOUN_obj_follows_VERB	32 (exception: kor, tur)
SCONJ_obj_follows_VERB	1 (nld)
VERB_obj_follows_VERB	12 (cmn, hrv, dan, nld, eng, fin, hin, ita, lav, ron, rus, tha)
AUX_obj_follows_VERB	1 (hrv)
ADJ_obj_follows_VERB	8 (exception: deu, hin, ind, gle, ita, jpn, kor, tur)
X_obj_follows_VERB	10 (cmn, hrv, dan, nld, ell, ind, ita, lit, mlt, slk)
PROPN_obj_follows_VERB	(exception: hin, jpn, kor, tur)

Table 6.12. Ensemble and overall distribution of VO features extracted from PUD and EU corpora.

In terms of VO features, again the highest percentages of occurrences correspond to dependents which are nouns, pronouns, or proper nouns. Turkish and Korean do not present any occurrence of an object following the verb, while Japanese has a low frequency of nouns before verbs, and Hindi of nouns, pronouns, and other verbs before the verbal head.

The total percentage of VO and OV features extracted from each PUD and EU corpus is presented in Figure 6.13. It is possible to notice that the OV languages are positioned on the right side of the graph, with the lowest frequencies of VO features. Moreover, German and Dutch (which are considered as “no dominant order” in WALS) present quite similar percentages of both VO and OV. The other Germanic languages, on the other hand, present much higher percentage values regarding VO phenomena. Hungarian is characterized in WALS as VO, however, the distribution of VO and OV features for this language is closer to what is observed for German and Dutch. The other Uralic languages have more VO features but with a considerable percentage of OV. All Romance languages can be found in the middle part of the graph, with a higher frequency of VO features but with some OV phenomena (e.g.: pronominal objects). The Slavic languages present different distributions regarding the verb and object position. Slovak and Slovenian have a relatively high number of OV occurrences, followed by Croatian. Russian and Polish, however, have a very low percentage of this type of word order. Maltese and Arabic have a similar distribution of VO and OV occurrences, however, in the Maltese corpus, the overall percentages are higher. Greek is positioned in the middle of the graph, presenting a distribution quite similar to most of the other Indo-European languages. In the Irish corpus, the amount of verb and object occurrences is relatively low, however, this language presents a much higher amount of VO phenomena, as expected. Furthermore, Chinese and Thai are the VO languages with the highest percentage of attested verb and object phenomena, being positioned on the extreme left side of the figure.

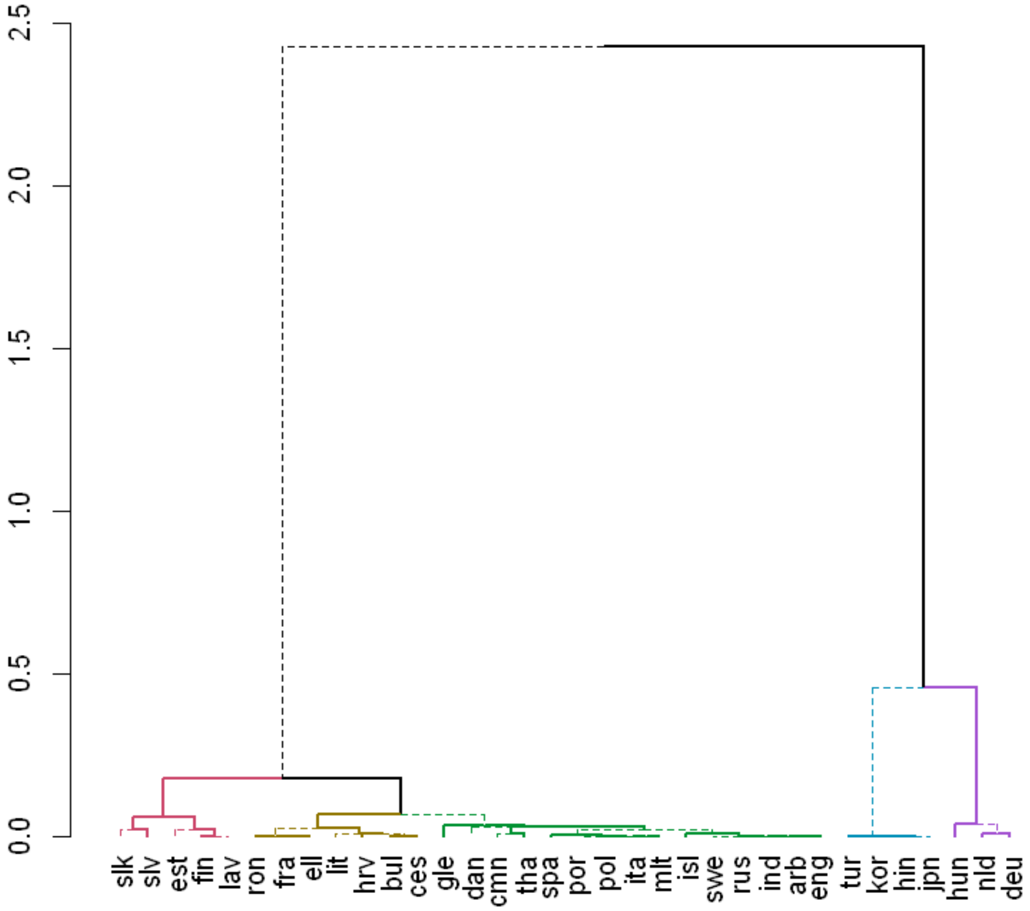
Figure 6.13. General distribution in terms of frequency of OV and VO features for each language.



In terms of the percentage of the total amount of extracted verb and object occurrences, it is noticeable that most of the languages (23 out of 34) have between 3 to 4%. Four languages have less than 3%: Turkish, Irish, Japanese, and Arabic. Six languages have between 5 to 7%: Slovenian, Danish, Slovak, Hindi, Finnish, and Chinese. Thai is the language with the highest percentage of verb and object constructions (8.1%).

With the extracted features and frequencies, we generated the cosine dissimilarity matrix (Annex 89) which was used to build the dendrogram presented in Figure 6.14.

Figure 6.14. Cluster dendrogram obtained from the cosine dissimilarity matrix calculated with the comparison of the PUD and EU language vectors built with VO and OV features.



The dendrogram presents some similar results to what was observed when the total percentage of VO and OV occurrences were analysed for each language. OV languages form a specific cluster, which is closer to the 3 languages which were identified as having a balanced amount of VO and OV phenomena (i.e.: German, Dutch, and Hungarian). On the left side of the dendrogram, the 4 VO languages which present a considerably high number of OV phenomena are clustered together, divided into two sub-groups (i.e.: Slovak and Slovenian; Estonian and Finnish). In the middle of the dendrogram, it is possible to find the languages with the lowest amount of OV occurrences. French, Romanian, and Greek form a specific sub-cluster, closer to a group formed by Lithuanian, Croatian, Bulgarian, and Czech. The large green cluster is composed of a great variety of VO languages, including many Indo-European ones together with Chinese, Thai, Maltese, and Arabic.

When the dissimilarity matrix is used to identify the possible optimized combination pairs for Croatian and the low-resourced EU languages, we have:

- Croatian: Czech (best);
- Hungarian: German (best);
- Irish: Maltese (best), Polish, English, Russian, Indonesian;
- Lithuanian: Czech (best);
- Maltese: Russian (best), Polish.

From the analysis of the dendrogram in comparison with the languages selected via the distance values from the dissimilarity matrix, we decided to add Italian as a possible combination for Maltese.

6.3.4. MarsaGram all properties (cosine)

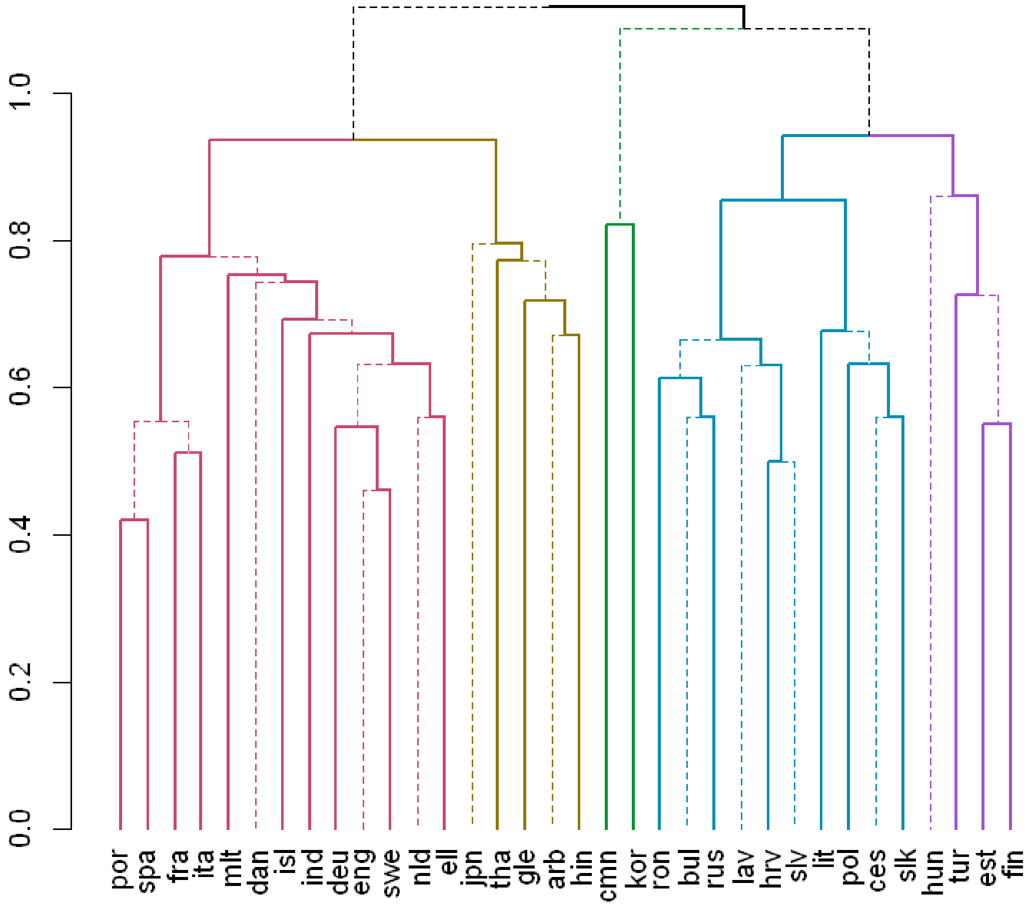
The last strategy to be considered concerns all properties extracted using MarsaGram with the cosine distance metric for the comparison of the language vectors. This typological method showed interesting results for MLAS improvement as it presented the lowest number of proposed associations with negative synergy from all the proposed corpus-based methods.

In Section 6.3.2, we detailed the obtained features (240,882) regarding MarsaGram all properties. In that case, Euclidean distance was used to compare languages and the results were combined with the head and dependent analysis.

When cosine distance is used to compare the PUD and EU languages, we obtain the dissimilarity matrix presented in Annex 90. The correspondent dendrogram is presented in Figure 6.15.

The language classification obtained from the cluster analysis of the language vectors composed of the information extracted using MarsaGram (all properties) shows some similarities when compared to the phylogenetic classification regarding some Indo-European genera, and Uralic languages.

Figure 6.15. Cluster dendrogram obtained from the cosine dissimilarity matrix calculated with the comparison of the PUD and EU MarsaGram language vectors (all patterns).



Concerning EU languages, it is possible to observe that:

1) Regarding the Indo-European family:
    a. Baltic genus: As was the case when only linear patterns were considered, Lithuanian and Latvian do not form a specific sub-cluster in the MarsaGram all properties dendrogram. However, in this case, they are classed in the same cluster (in blue) together with all the Slavic languages and Romanian. When only the linear patterns were analysed, Latvian was closer to some Slavic languages, but Lithuanian formed a sub-group with Indonesian, which in this case is clustered with Germanic languages in the pink cluster. Hawkins (1983) did not provide information for Latvian, while Lithuanian is considered as type 11 (same type as Danish and Swedish), however, in this dendrogram, these languages are not all clustered in the same group.

b. Celtic genus: Irish is the only language from this specific Indo-European genus. Hawkins (1983) classified it as type 1 which is the same group as Arabic. In this dendrogram, although these 2 languages do not form a specific sub-cluster, they are relatively close inside a cluster which also contains Japanese, Thai, and Hindi (i.e.: 2 OV languages).
c. Germanic genus: this genus can be easily identified as a sub-group of the pink cluster. It is possible to notice the proximity between English and Swedish which are positioned close to German. However, this latter is not classed adjacently to Dutch. Instead, Dutch forms a sub-cluster with Greek. Danish and Icelandic are the Germanic languages with the highest values of distance to the other ones from this genus. The classification of Germanic languages does not follow precisely the types defined by Hawkins (1983), as if it was the case, Swedish should be closer to Danish, Icelandic close to Greek, and Indonesian should be part of the Romance cluster. However, Dutch and Greek (both type 10 according to Hawkins, 1983) form a specific sub-cluster.
d. Greek genus: Greek is the only language from this genus in the selected language-set. In the dendrogram, as presented above, it is part of the pink group, forming a sub-cluster with Dutch, which follows Hawkins classification (1983).
e. Romance genus: As was the case when only MarsaGram linear patterns were analysed, Romance languages except for Romanian are all part of the same sub-cluster with a clear sub-division between one sub-group formed by Portuguese and Spanish, and another one composed of French and Italian. Romanian is placed with Slavic and Baltic languages in the blue cluster, being positioned closer to the sub-cluster formed by Bulgarian and Russian. All Romance languages are classed as type 9 by Hawkins (1983) together with Thai and Indonesian. The latter is also part of the pink cluster but is classed with Germanic languages. Thus, although Romanian presents similarities in terms of the word-order patterns analysed by Hawkins, in this specific study of MarsaGram patterns, this language shows more similarity with the Slavic ones. One possible explanation for that is the fact that among Romance languages, Romanian is more flexible in terms of word order structures due to its richer morphological characteristics when compared to the others, in this way being more similar to Slavic ones.

    f. Slavic genus: Regarding the 7 selected Slavic languages, it is possible to notice that all of them are part of the same group (blue). It is a large cluster that also contains Romanian and the 2 Baltic languages. In both dendrograms describing the languages according to lang2vec phylogenetic features (Figures 6.1 and 6.2), the Baltic languages are grouped with the Slavic ones as they share the same genealogical feature "F_Balto-Slavic". These two genera appear in the same cluster in both MarsaGram dendrograms presented in this section. However, the MarsaGram all properties analysis does not correspond precisely to the phylogenetic one. Croatian and Slovenian form a specific sub-cluster, but Czech is grouped with Slovak, and Russian with Bulgarian. Moreover, the Baltic languages are not adjacent in the dendrogram. This MarsaGram typological representation does not follow Hawkins (1983) classification, as Czech and Russian should be closer if it was the case (as both are considered as type 10 together with Dutch, Greek, and Icelandic).

2) Concerning non-Indo-European languages:
    a. Uralic family: When all MarsaGram properties are considered, all Uralic languages are part of the same cluster (purple) together with Turkish even though this language is OV and from a different language type in Hawkins (1983) analysis (i.e.: Turkish is considered as type 23 with the other OV languages, while the Finnic languages are type 15). When only MarsaGram linear patterns were analysed, Hungarian was not positioned in the same group as the other Uralic languages, thus, this representation is closer to the genealogical classification.
    b. Afro-Asiatic family: Maltese and Arabic are the two languages from this family in our language-set. However, they are not part of the same cluster in the dendrogram. Maltese is positioned in the pink cluster as a single branch closer to Germanic languages and Indonesian. When only linear patterns were analysed, it was grouped with OV languages. Arabic is part of the cluster in the middle of the dendrogram together with languages from many different linguistic families, including Irish. Both Irish and Arabic were considered as type 1 by Hawkins (1983).

Thus, it is possible to notice that when all MarsaGram properties are considered, the obtained classification shows more coherence to the phylogenetic one when compared to the analysis of only linear patterns. By analysing all properties, not only the word order inside the subtrees are quantified, but also some other linguistic phenomena. Moreover, the dendrogram shows that some specific word order features such as the ones considered by Hawkins (1983) or the basic verb and object ordering are not the most relevant ones in this analysis.

When the dissimilarity matrix is used to identify the possible optimized combination pairs for Croatian and the low-resourced EU languages, we have:

- Croatian: Slovenian (best);
- Hungarian: Latvian (best), Greek, German, and Turkish;
- Irish: Italian (best), Dutch, French, and Portuguese;
- Lithuanian: Slovak (best) and Czech;
- Maltese: Greek (best), German, French, Swedish.

For Hungarian, Irish, and Maltese, the criterion which includes the languages for which the difference between the distance from the other selected languages and the best candidate is lower than 10% would mean the selection of more than 10 combinations, thus, in these cases, we selected only the other 3 languages with the lowest distance values from the dissimilarity matrix. Moreover, besides Croatian, the other selected languages do not form a specific sub-cluster in the dendrogram, so no other language-pair is added to the list.

6.3.5. Discussion and language-pairs selection

In this section, we presented the application of the four selected corpus-based typological approaches to all European Union languages. Thus, we established four different language classifications regarding these 24 languages and 10 other worldwide ones.

Each dendrogram showed that when the analysis focuses on different syntactic phenomena, languages are classed differently. However, in all cases, similarities can be found to what has been described previously by Hawkins (1983), or to classic typological approaches (e.g.: VO and OV languages). The proposed quantitative studies presented here consider broader scenarios when compared with the classical methodologies as more syntactic phenomena are examined and all occurrences are considered (not only the standard attested word order).

The main objective of this analysis was to determine for each EU low-resourced language and Croatian, the language-pairs to be tested to improve dependency parsing results (LAS and MLAS). Each typological method provided a set of possible optimized combinations (with some overlap). Table 6.13 displays the ensemble of the selected languages.

	MarsaGram Linear (cos)	**Marsgaram all + Head and Dependent (Euc)**	**VO/OV (cos)**	**MarsaGram all (cos)**	**Total number of combinations**
hrv	slv, rus, eng, deu	ces	ces	slv	5
hun	ell, fra	nld, mlt, deu, eng, est, fin	deu	lav, ell, deu, tur	10
gle	fra	ron	mlt, pol, eng, rus, ind	ita, nld, fra, por	10
lit	hrv, por, ind	lav	ces	slk, ces	6
mlt	hrv, slv, eng, tur	dan, eng, nld, swe, tha	rus, pol, ita	ell, deu, fra, swe	14

Table 6.13. Overall description of the identified language-pairs for each corpus-based typological strategy for EU low-resourced languages and Croatian.

Therefore, following the criteria described in the previous sub-sections, the number of selected combinations varies from 5 for Croatian, and 14 for Maltese. Moreover:

- For Croatian, Slovenian is selected when both MarsaGram linear and all patterns are considered. Czech is the chosen language when the combined method is used, as well as when the verb and object features are analysed.
- For Hungarian, Greek is selected using the MarsaGram linear strategy and also when MarsaGram patterns are considered. German is chosen by the combined method and with the analysis of the verb and object positions.
- For Irish, French is the selected language for the MarsaGram linear method but is also a possible choice when all MarsaGram patterns are considered.
- For Lithuanian, Czech is the chosen language when the verb and object position features are analysed and is also a choice when all Marsgram patterns are examined.
- For Maltese, English is selected by both MarsaGram linear and the combined methods, while Swedish is chosen by the combined strategy and by the MarsaGram all properties one.

It is possible to notice that for each low-resourced language and Croatian, except for Maltese, at least one language from the same linguistic family is selected. These selected language-pairs were tested using UDify and the results are presented in the following sub-section.

6.3.6. Dependency parsing experiments

The objective of this sub-section is to analyse how the low-resourced languages and Croatian behave in terms of dependency parsing results when these languages are combined with the ones selected via the 4 different designated corpus-based typological methods.

For each language, the baseline for comparison is composed of the LAS and MLAS scores obtained when UDify is trained with the respective UD corpus. Table 6.14 presents the details concerning each corpus and the Table 6.15, the mean value of the obtained scores together with the standard deviations (calculated with the variation of the random seed value as described in section 5.2.1) and the scores published by the developers of UDify (using their multilingual model) (Kondratyuk and Straka, 2019).

		Training set		**Developement set**		**Test set**	
	UD Corpus (v.2.7)	**Sentences**	**Tokens**	**Sentences**	**Tokens**	**Sentences**	**Tokens**
hrv	SET	6,914	152,857	960	22,292	1,136	24,260
hun	Szeged	910	20,166	441	11,418	449	10,448
gle	IDT	4,005	95,860	451	10,000	454	10,109
lit	Alksnis	2,341	47,641	617	11,560	684	10,846
mlt	MUDT	1,123	22,880	433	10,209	518	11,073

Table 6.14. Description of the UD corpora in terms of the size of the selected languages for the dependency parsing improvement experiments.

The language with the smallest corpus is Hungarian (1,800 sentences in total), followed by Maltese (2,074 sentences). However, the total number of sentences is higher than the ones found for each language in the PUD collection. Irish (4,910) and Lithuanian (3,642) have larger corpora but are still considerably smaller than Croatian (9,010). In our experiments, we decided to keep the split in terms of training, development, and test sets as established in the UD database. Nevertheless, it is important to mention that these languages present different proportions: Hungarian and Maltese have training sets corresponding to around 50% of the total size, Lithuanian training set corresponds to 65%, while for Croatian and Irish, the training set is larger (76% and 80% respectively).

	UD Corpus (v.2.7)	**LAS**	**Std. Dev.**	**LAS (multilingual model)**	**MLAS**	**Std. Dev.**	**MLAS (multilingual model)**
hrv	SET	89.03	0.04	**89.79**	**78.93**	0.08	72.72
hun	Szeged	82.81	0.07	**84.88**	**67.22**	0.14	64.27
gle	IDT	**77.87**	0.07	69.28	**40.10**	0.16	34.39
lit	Alksnis	78.16	0.12	-	60.48	0.19	-
mlt	MUDT	73.79	0.23	**75.56**	57.50	0.36	**58.14**

Table 6.15. Baseline regarding the UDify results for the selected languages (monolingual models) in comparison to the results presented in the literature (Kondratyuk and Straka, 2019). Lithuanian score regarding Alksnis corpus is not provided in the literature. In bold are represented the best score when the monolingual results are compared to the multilingual ones.

In terms of LAS results, Croatian has the highest score which was expected as it has the largest corpus among the selected languages. Hungarian has the smallest training corpus, however, its LAS score is above 80, higher than the other low-resourced languages. The LAS scores obtained for Irish and Lithuanian are relatively close (higher than 77), while Maltese is the one with the worst result (73.79). When comparing these LAS scores to the ones presented by Kondratyuk and Straka (2019), it is possible to notice that the ones obtained with the multilingual model are higher than the monolingual ones except for Irish, which is coherent with what is stated by these authors (i.e.: the advantage of multilingual training for LAS).

When analysing the MLAS results, again Croatian has the highest score followed by Hungarian. However, for this specific metric, Irish has the worst result even though its corpus is considerably larger than the Maltese one. Furthermore, in terms of MLAS, the scores obtained in our experiments (monolingual) tend to be better than the ones provided by the multilingual model presented in the literature (Kondratyuk and Straka, 2019), the exception, in this case, is Maltese.

With the baseline established, we conducted the dependency parsing experiments using combined corpora using the same UDify parameters as described in Section 5.2.1. When PUD languages were associated, the size of the combined training set was 1,200 sentences (600 from language 1 and 600 from language 2). The development and test sets were composed of 200 sentences exclusively from language 1. In the experiments described in this section, we kept the development and test sets as detailed in Table 6.14 (also monolingual). For the combination of the training sets, we associate the whole training set of language 1 (Table 6.14) with the same number of sentences provided by UD corpora of language 2 (using also development and test sentences if the language 2 training set did not have enough data). In some particular cases, the associated language (language 2) does not have enough sentences to obey the determined ratio, thus, the experiments were conducted with all the UD available data (i.e.: with a larger proportion of language 1 in the combined corpus). These cases are detailed more precisely later when results are presented. When more than 1 corpus is available regarding language 2, we tried to respect as much as possible the genre of the language 1 corpus, specially avoiding spoken transcriptions.

In the subsequent paragraphs, the results obtained for each language are described separately, followed by a general discussion of the outcomes. To better understand how the combination of languages helps improving the dependency parsing results (LAS), for each low-resourced language and Croatian, we analysed in detail: a) which dependency relations were improved when compared to the monolingual model, b) which relations were negatively impacted, c) the impact of the language association in relation with the size of the sentences in the test-set, d) the impact of the combination in relation with the distances between the heads and the dependents (i.e.: positive distances corresponding to heads after the dependents in the sentence, and negative distances related to heads preceding the dependents). We compared the monolingual model to the best-identified combination of languages in terms of LAS improvement, using the annotated text from the first trained model (i.e.: with the UDify standard random seed value). For this analysis, we used the DependAble tool developed by Choi et al. (2015):

- First, the annotated corpus obtained via the monolingual model was compared to the gold test-set.
- Then, the corpus annotated with the best-identified model from the experiments with the language combinations was compared to the gold test-set.
- Finally, the scores were compared with the calculation of delta (i.e.: language association results minus monolingual results) .

a) Croatian

As previously mentioned, Croatian was not identified as a low-resourced language in terms of the criteria established for this thesis in terms of UD corpus size and dependency parsing results. This language was selected to test if the dependency parsing improvement strategies also present some advantages even for languages with some more resources.

By applying the different corpus-based typological strategies to Croatian, we identified the following languages to be combined with it: Slovenian, Russian, English, German, and Czech. Of the 5 selected languages, 3 are also Slavic (Slovenian being the closest one in terms of phylogenetic criteria), and 2 are Germanic, thus all of them are from the same family.

Table 6.17 presents the UD v.2.7 corpora chosen for the experiments regarding Croatian.

	UD corpus	**Total size (sentences)**	**Training corpus (sentences)**	**Genre**
ces	PDT	87,913	68,495	news, nonfiction, reviews
deu	GSD	15,590	13,814	news, reviews, wiki
eng	GUM	5,961	4,287	academic, blog, fiction, government, news, nonfiction, social, spoken, web, wiki
rus	SynTagRus	61,889	48,814	fiction, news, nonfiction
slv	SSJ	8,000	6,478	fiction, news, nonfiction

Table 6.17. Description of the UD v.2.7 corpora selected to be combined with Croatian SET training corpus (genres: news, web, and Wiki).

The size of the Croatian training corpus (SET) is 6,914 sentences, thus, for Russian, German, and Czech, we extracted the first 6,914 sentences of the respective training corpus described in Table 6.17. For Slovenian, as the SSJ training corpus has 6,478 sentences, 436 sentences were extracted from the development set. Moreover, for English, the whole GUM set was used plus 953 sentences from EWT English training corpus. Thus, in the end, the total size of the combined corpora for the experiments regarding Croatian is 13,828 sentences. The results of the language association experiments are presented in Table 6.18.

	Mean LAS	**stdev LAS**	**Delta LAS**	**p_value LAS**	**Mean MLAS**	**stdev MLAS**	**Delta MLAS**	**p_value MLAS**
hrv_ces	88.72	0.03	-0.31	0.00	78.31	0.03	-0.62	0.00
hrv_deu	88.92	0.05	-0.11	0.01	78.88	0.05	-0.05	0.32
hrv_eng	88.90	0.04	-0.13	0.00	78.21	0.03	-0.72	0.00
hrv_rus	88.96	0.05	-0.07	0.07	79.20	0.05	0.27	0.00
hrv_slv	88.95	0.08	-0.08	0.13	79.07	0.08	0.14	0.04

Table 6.18. Results obtained via the language association for Croatian. Positive deltas are identified in green, while negative ones are in red. When the p-values are lower than 0.01, the cells are highlighted in green.

From the results in Table 6.18, it is possible to observe that no improvement in terms of LAS is obtained when languages are combined. Instead, when Croatian is combined with Czech and

English, the LAS is decreased (with statistical significance). For the other 3 languages, deltas are also negative but the LAS results are not statistically different from the one obtained with the Croatian monolingual model.

In terms of MLAS, only the association with Russian presents a significant improvement (0.27). When Croatian is combined with Slovenian and German, results are statistically similar to the score of the monolingual model. As was the case for LAS, the associations with Czech and English also provide negative deltas.

Russian was selected as a candidate to be combined with Croatian by the MarsaGram linear method (cosine). With this method, Slovenian was identified as the closest language to Croatian, Russian being the second closest one. With this typological strategy, English and German were also selected as potential candidates. Thus, from the languages identified by this method, 1 presented an improvement in terms of MLAS (i.e.: Russian), 2 did not present any significant change in terms of LAS or MLAS (i.e.: German and Slovenian), and 1 presented a decrease on these metrics (i.e.: English).

With the MarsaGram all properties (cosine) strategy, only Slovenian was identified as a candidate. On the other hand, the optimized association of MarsaGram all properties with head and dependent positions and the verb and object position strategies identified Czech as the best candidate, but this language associated with Croatian presented a negative delta for both LAS and MLAS.

Table 6.19 presents the comparison between the LAS values obtained with the monolingual model and with the model trained with the combination of Croatian and Russian for each specific dependency relation label present in the test corpus. Although no combination experiment presented a significant improvement in terms of LAS, we decided to check the results obtained when Russian is combined with Croatian as it presented at least a positive delta for MLAS. It is interesting to notice that Russian is part of the same genus as Croatian but this characteristic alone does not guarantee a positive delta as for Slovenian no improvement was observed, and for Czech, results were degraded.

DEPREL	hrv	hrv + rus	Delta
flat:foreign	25.00	75.00	50.00
fixed	74.49	76.53	2.04
parataxis	62.88	64.88	2.00
discourse	65.38	66.83	1.45
ccomp	85.15	86.46	1.31
conj	81.07	81.95	0.88
acl	78.15	79.03	0.88
cc	92.45	93.24	0.79
appos	64.62	65.38	0.76
expl	96.36	97.02	0.66
det	92.27	92.52	0.25
obl	84.88	85.01	0.13
advmod	83.89	84.00	0.11
amod	95.26	95.31	0.05
advmod:emph	0.00	0.00	0.00
case	96.74	96.74	0.00
compound	0.00	0.00	0.00
csubj	68.29	68.29	0.00
iobj	55.00	55.00	0.00
orphan	7.69	7.69	0.00
vocative	0.00	0.00	0.00
root	96.39	96.21	-0.18
mark	93.63	93.41	-0.22
nmod	86.76	86.40	-0.36
aux	96.53	96.14	-0.39
nsubj	92.22	91.71	-0.51
xcomp	92.84	92.26	-0.58
punct	92.85	92.26	-0.59
cop	78.74	78.02	-0.72
obj	83.16	82.42	-0.74
nummod	77.48	75.07	-2.41
flat	85.58	82.85	-2.73
advcl	70.92	67.35	-3.57

Table 6.19. LAS values for each DEPREL label in the Croatian test-set regarding the monolingual model and the association with Russian. The delta values consist of the difference between the results obtained via the language association and the monolingual one. The color scale highlights the most positive delta values (green), and the most negative ones (red).

It is noticeable that the dependency relation with the highest improvement (+50.00) is flat:foreign[43] which is present only 4 times in the test-set. This tag is composed of a type and a subtype (i.e.: related to foreign MWE). The flat label (without any subtype) is also present in the corpus (659 occurrences) and presented a negative delta (-2.73). The dependency relation “fixed”, which also involves multiword expressions, also presents a positive delta (+2.04), however, an improvement cannot be observed for the third type of relation used for MWE (i.e.: “compound”). Three other relations showed an improvement higher than 1.00: parataxis, discourse, and clausal complement (ccomp). In total, 14 labels presented a positive delta for this language association.

In terms of negative delta, the relation that was most negatively impacted is the adverbial clause modifier (advcl) (-3.57), followed by flat (-2.73) and numerical modifier (nummod) (-2.41). All the other 9 negative deltas are comprised between -1.00 and 0.00, among these cases, we find labels that usually have a high frequency in corpora, such as nominal subject (nsubj), root, object (obj), auxiliary (aux), and even punctuation (punct).

The Tables 6.20, 6.21, and 6.22 present respectively the analysis of the LAS results in relation to the sentence length and the distances between head and dependents (positive and negative).

Overall, it is possible to notice that the LAS values decrease for longer sentences and when the heads are more distant from the dependents. When comparing the results obtained via the monolingual model with the ones obtained via the language association, we observe that the delta results present in Tables 6.20, 6.21, and 6.22 are mainly negative. In terms of sentence length, it is possible to observe that the language association decreases the LAS the most for very short sentences (less than 10 tokens), and very long ones (more than 40). When the distance between heads and dependents is considered, it is noticeable that when the head is after the dependent, the language combination model presents the highest decrease when the head is immediately after the dependent. A positive delta is observed when heads are positioned very far from the dependent (more than five tokens away). The opposite is observed for heads positioned before the dependents, as the most negative deltas correspond to the most distant heads.

---

[43] According to the Universal Dependencies guidelines: “The flat relation is one of three relations for multiword expressions (MWEs) in UD (the other two being fixed and compound). It is used for exocentric (headless) semi-fixed MWEs like names (*Hillary Rodham Clinton*) and dates (*24 December*). It contrasts with fixed, which applies to completely fixed grammaticized (function word-like) MWEs (like *in spite of*), and with compound, which applies to endocentric (headed) MWEs (like *apple pie*)”.

	<=10	<=20	<=30	<=40	<=50	>50
hrv	90.34	90.09	89.00	87.91	88.44	87.23
hrv + rus	89.27	89.92	89.02	87.83	87.94	86.25
Delta	-1.07	-0.17	0.02	-0.08	-0.5	-0.98

Table 6.20. LAS results for different sentence lengths. In green is highlighted the positive delta value (i.e.: the difference between the language association result and the monolingual one).

	1	2	3	4	5	>5
hrv	90.25	87.90	85.05	81.97	81.64	83.75
hrv + rus	89.84	87.75	85.00	81.92	81.53	84.01
Delta	-0.41	-0.15	-0.05	-0.05	-0.11	0.26

Table 6.21. LAS results for different positive distances between heads and dependents (i.e.: head after the dependent). In green is highlighted the positive delta value (i.e.: the difference between the language association result and the monolingual one).

	<-5	-5	-4	-3	-2	-1
hrv	87.60	87.04	87.21	87.87	90.57	93.60
hrv + rus	86.97	86.24	86.47	87.49	90.29	93.48
Delta	-0.63	-0.80	-0.74	-0.38	-0.28	-0.12

Table 6.22. LAS results for different negative distances between heads and dependents (i.e.: head before the dependent).

b) Hungarian:

As previously described, the Hungarian corpus has a total size of 1,800 sentences. Its training set has only 910 sentences, thus, for the 10 selected languages to be combined with Hungarian, we privileged PUD corpora when they were available (i.e.: German, English, Finnish, French, and Turkish). The Szeged Hungarian corpus is composed of texts from news, a genre which is also part of PUD. Moreover, this way it is possible to analyse these combinations in a more detailed way as the added sentences from these 5 languages are parallel.

The description of each selected UD corpus is described in Table 6.23. The final combined corpora have a total size of 1,820 sentences. Regarding the selected languages, it is possible to observe that the 2 languages from our language-set from the same genealogical family as Hungarian were selected (i.e.: Finnish and Estonian). In terms of Indo-European languages, 3 Germanic languages, 1 Romance language, 1 Baltic language, and Greek were also identified as possible candidates. Moreover, Maltese (Afro-Asiatic family) and Turkish (Altaic family) are also part of the list. Thus, showing a larger variety of selected languages when compared to Croatian.

	UD corpus	Total size (sentences)	Training corpus (sentences)	Genre
deu	PUD	1,000	-	news, wiki
ell	GDT	2,521	1,162	news, spoken, wiki
eng	PUD	1,000	-	news, wiki
est	EDT	30,972	24,633	academic, fiction, news, nonfiction
fin	PUD	1,000	-	news, wiki
fra	PUD	1,000	-	news, wiki
lav	LVTB	13,643	10,156	academic, fiction, legal, news, spoken
mlt	MUDT	2,074	1,123	fiction, legal, news, nonfiction, wiki
nld	Alpino	13,578	12,264	news
tur	PUD	1,000	-	news, wiki

Table 6.23. Description of the UD v.2.7 corpora selected to be combined with the Hungarian Szeged training corpus (genre: news). PUD corpora correspond only to test-sets composed of 1,000 sentences.

The results concerning the combination of Hungarian with the 10 selected languages are presented in Table 6.24.

Language pair	Mean LAS	Stdev LAS	Delta LAS	p_LAS	Mean MLAS	Stdev MLAS	Delta MLAS	p_value MLAS
hun_deu	83.11	0.06	0.30	0.00	67.61	0.07	0.40	0.00
hun_ell	83.18	0.12	0.37	0.00	67.29	0.17	0.07	0.44
hun_eng	83.08	0.20	0.27	0.01	67.32	0.32	0.10	0.49
hun_est	82.77	0.12	-0.03	0.56	67.16	0.13	-0.05	0.51
hun_fin	83.10	0.19	0.29	0.00	68.31	0.18	1.09	0.00
hun_fra	83.28	0.15	0.47	0.00	67.72	0.23	0.50	0.00
hun_lav	83.44	0.22	0.63	0.00	67.14	0.33	-0.08	0.61
hun_mlt	82.81	0.15	0.00	0.96	67.18	0.30	-0.03	0.81
hun_nld	83.31	0.06	0.50	0.00	68.49	0.18	1.27	0.00
hun_tur	83.19	0.19	0.38	0.00	68.12	0.26	0.90	0.00

Table 6.24. Results obtained via the language association for Hungarian. Positive deltas are identified in green, while negative ones are in red. When the p-values are lower than 0.01, the cells are highlighted in green.

In total, 7 associations presented a significant improvement in terms of LAS, and no association decreased this metric. The highest delta was obtained when Hungarian was combined with Latvian (+0.63). The other languages generating a positive delta are German, Greek, Finnish, French, Dutch, and Turkish. However, the improvement provided by these associations is significantly lower than the one obtained with Latvian (i.e.: p-value is lower than 0.01). English, Estonian, and Maltese did not present significant improvement or decrease for this metric. Latvian was identified as the closest language to Hungarian via the MarsaGram all properties method with which we also selected German, Greek, and Turkish (all with significant positive deltas). Moreover, it is possible to notice that the two languages selected with the MarsaGram linear strategy also provide positive deltas (i.e.: Greek and French). The method corresponding to the association of MarsaGram all properties and head and dependent positions is responsible for the identification of the 3 languages that did not improve the LAS results (together with 3 languages with a positive delta: Dutch, German, and Finnish).

Regarding MLAS, 5 languages provided a significant positive delta, the association of Hungarian and Dutch being the most favorable one (+1.27) with an improvement significantly higher than the ones from the other associations. The same languages which did not provide any improvement in terms of LAS, did not enhance MLAS either, together with Maltese and Latvian. It is interesting to notice that an improvement in terms of LAS (as the one observed with Latvian) does not guarantee the same result for MLAS. Moreover, no language association presented any significant decrease. For this specific metric, it was the optimized strategy composed of the MarsaGram all properties and head and dependent position methods that provided the best candidate (i.e.: Dutch was identified as the closest language). Of the 6 selected languages with this method, Maltese, English, and Estonian did not provide any improvement.

For Hungarian, the combinations providing the best LAS and MLAS results are composed of the association of this language with another one from a different family and genus. The association with Finnish presented a significant positive delta for both metrics but lower than the best-achieved scores. On the other hand, Estonian did not provide any improvement.

As the combination of Hungarian and Latvian was identified as the best one in terms of LAS, we conducted the analysis with the DependAble tool for this specific combination in comparison with the results obtained with the monolingual Hungarian model. The LAS results for each specific dependency relation tag are presented in Table 6.25.

DEPREL	hun	hun + lav	Delta
acl	30.56	50.00	19.44
aux	75.00	91.67	16.67
iobj	46.67	60.00	13.33
advcl	43.88	54.08	10.20
appos	44.68	53.19	8.51
obj:lvc	0.00	8.33	8.33
parataxis	1.37	8.22	6.85
ccomp:obl	53.13	59.38	6.25
ccomp:obj	51.52	57.58	6.06
conj	71.04	75.21	4.17
csubj	37.84	40.54	2.70
case	90.82	93.37	2.55
advmod	86.32	88.42	2.10
mark	87.97	89.87	1.90
xcomp	95.95	97.30	1.35
punct	78.90	80.05	1.15
root	93.32	94.21	0.89
cc	78.11	78.95	0.84
nmod:obl	84.95	85.48	0.53
det	92.60	93.07	0.47
nmod:att	80.51	80.68	0.17
advmod:mode	67.32	67.32	0.00
advmod:que	75.00	75.00	0.00
advmod:tfrom	16.67	16.67	0.00
advmod:to	0.00	0.00	0.00
advmod:tto	0.00	0.00	0.00
amod:obl	7.69	7.69	0.00
ccomp:pred	0.00	0.00	0.00
compound	97.50	97.50	0.00
compound:preverb	96.33	96.33	0.00
cop	73.17	73.17	0.00
goeswith	0.00	0.00	0.00
list	16.67	16.67	0.00
nmod:obllvc	0.00	0.00	0.00
nsubj:lvc	0.00	0.00	0.00
obj	95.73	95.73	0.00
orphan	4.17	4.17	0.00
obl	66.19	65.71	-0.48
amod:att	90.34	89.44	-0.90
flat:name	91.12	90.19	-0.93
nummod	80.65	79.57	-1.08
nsubj	90.78	89.22	-1.56
advmod:tlocy	80.87	77.83	-3.04
advmod:locy	65.63	62.50	-3.13
amod:mode	75.00	71.15	-3.85
nmod	27.27	18.18	-9.09
ccomp	23.08	7.69	-15.39

Table 6.25. LAS values for each DEPREL label in the Hungarian test-set for the monolingual model and the association with Latvian. The color scale highlights the most positive delta values (green), and the most negative ones (red).

In total, 21 relations present an improvement, and for 4 of them, the delta is higher than 10.00: adnominal clause (acl), auxiliary (aux), indirect object (iobj), and adverbial clause modifier (advcl). The advcl and aux labels were identified in Croatian as relations presenting negative deltas. Also, for Hungarian, the root and punctuation labels present a slight improvement, while its score was decreased in Croatian. Parataxis, on the other hand, is improved in both languages. It is also possible to notice that the association of Hungarian with Latvian allows both coordination to be improved: coordinating conjunction (cc) with a delta of 0.84, and conjunct (conj) with 4.17 as the delta. Furthermore, nominal subjects presented a slight decrease, however, objects remained with the same score.

The detailed analysis regarding LAS scores in terms of sentence length and head and dependent distances is presented in Tables 6.26, 6.27, and 6.28.

	<=10	<=20	<=30	<=40	<=50	>50
hun	86.03	83.78	83.36	82.27	80.79	63.65
hun + lav	84.38	84.83	83.97	82.45	81.02	68.79
Delta	-1.65	1.05	0.61	0.18	0.23	5.14

Table 6.26. LAS results for different sentence lengths. In green are highlighted the positive delta values (i.e.: the difference between the language association result and the monolingual one).

	1	2	3	4	5	>5
Hun	85.19	79.81	78.26	74.70	77.52	72.58
hun + lav	86.13	80.69	80.81	75.94	79.60	75.93
Delta	0.94	0.88	2.55	1.24	2.08	3.35

Table 6.27. LAS results for different positive distances between heads and dependents (i.e.: head after the dependent). In green are highlighted the positive delta values (i.e.: the difference between the language association result and the monolingual one).

	<-5	-5	-4	-3	-2	-1
Hun	74.92	77.99	77.98	80.24	83.82	88.16
hun + lav	74.25	76.40	78.37	80.94	83.70	88.33
Delta	-0.67	-1.59	0.39	0.70	-0.12	0.17

Table 6.28. LAS results for different negative distances between heads and dependents (i.e.: head before the dependent). In green are highlighted the positive delta values (i.e.: the difference between the language association result and the monolingual one).

When the results are analysed individually, the tendencies are the same as for Croatian (i.e.: lower LAS values for longer sentences and when heads are more distant from dependents).

On the other hand, when the delta values are examined, in terms of sentence length, it is possible to notice that the best improvements obtained via the language association concern long sentences. Delta values are positive for sentences with more than 20 tokens, however, a decrease in LAS is observed for short sentences with less than 10 tokens. The highest improvement is attested in sentences with more than 50 tokens. In Croatian, it was possible to observe that longer and shorter sentences presented the worst deltas.

When the head and dependent distances are analysed, it is possible to notice that when the head is after the dependent, higher deltas are obtained for longer distances. On the other hand, when the head is before the dependent, the most negative deltas are obtained when these elements are far away from each other (5 or more than 5 tokens away). These results are relatively similar to what was observed for Croatian, however, the deltas obtained for Hungarian tend to be positive while they were mostly negative for Croatian.

c) Irish:

The Irish IDT UD corpus has a training set composed of 4,005 sentences. With the corpus-based typological strategies, 10 languages were selected as possible candidates to be combined with Irish. Table 6.29 presents the selected corpus for each one of these languages.

	UD corpus	**Total size (sentences)**	**Training corpus (sentences)**	**Genre**
eng	GUM	5,961	4,287	academic, blog, fiction, government, news, nonfiction, social, spoken, web, wiki
fra	GSD	16,341	14,449	blog, news, reviews, wiki
ind	GSD	5,593	4,477	blog, news
ita	ISDT	14,167	13,121	legal, news, wiki
mlt	MUDT	2,074	1,123	fiction, legal, news, nonfiction, wiki
nld	Alpino	13,578	12,264	News
pol	PDB	22,152	17,722	fiction, news, nonfiction
por	Bosque	9,364	8,328	news
ron	RRT	9,524	8,043	academic, fiction, legal, medical, news, nonfiction, wiki
rus	SynTagRus	61,889	48,814	fiction, news, nonfiction

Table 6.29. Description of the UD v.2.7 corpora selected to be combined with the Irish IDT training corpus (genres: fiction, government, legal, news, web).

For all the languages in Table 6.29 except for Maltese, the number of sentences in the training set is higher than the amount in the Irish set. Thus, the combined corpora are composed of the original sentences from the Irish corpus associated with 4,005 sentences of the other training sets (a total of 8,010 sentences). For Maltese, even when the whole corpus is considered (i.e.: training, development, and test sets), still the total amount of sentences is lower than 4,005. Thus, the final training corpus regarding the association of Irish and Maltese has a size of 6,079 sentences.

As was the case for Hungarian, the group of selected languages to be combined with Irish also presents some variety in terms of genealogical families and genera. Irish is the only language in our language-set from the Celtic genus (Indo-European family), and the majority of the chosen languages are also part of the same family (2 Germanic, 4 Romance, and 2 Slavic languages). Besides them, Maltese (Afro-Asiatic family) and Indonesian are also part of the list. It is interesting to notice that our strategy concerning verb and object positions identified Indonesian which delivered the best results when used to train a delexicalized model to parse Irish sentences in the experiments conducted by Lynn et al. (2014).

Hawkins (1983) classified Irish as type 1. In our language-set, Arabic is also from this same group, however, in our typological strategies, this language was not identified as a potential candidate. The results obtained with the combination of Irish with each one of the 10 selected languages are presented in Table 6.30. From the data displayed in this table, it is possible to notice that Irish does not benefit from the corpora association: MLAS results are statistically similar to the value obtained with the monolingual model, and, in terms of LAS, 5 combinations deliver worse results, and the other 5 do not provide any improvement or decrease.

	Mean LAS	**Stdev LAS**	**Delta LAS**	**p_value LAS**	**Mean MLAS**	**Stdev MLAS**	**Delta MLAS**	**p_value MLAS**
gle_eng	77.81	0.08	-0.06	0.22	39.83	0.46	-0.27	0.20
gle_fra	77.95	0.17	0.08	0.34	40.26	0.25	0.17	0.20
gle_ind	77.68	0.07	-0.20	0.00	39.96	0.11	-0.14	0.10
gle_ita	77.67	0.12	-0.21	0.00	39.86	0.19	-0.23	0.04
gle_mlt	77.76	0.09	-0.11	0.05	39.94	0.19	-0.15	0.15
gle_nld	77.67	0.08	-0.21	0.00	39.97	0.21	-0.13	0.25
gle_pol	77.37	0.13	-0.51	0.00	39.97	0.23	-0.12	0.30
gle_por	77.62	0.09	-0.26	0.00	40.45	0.23	0.36	0.01
gle_ron	77.74	0.07	-0.14	0.01	39.97	0.28	-0.12	0.37
gle_rus	77.89	0.12	0.02	0.79	40.03	0.28	-0.07	0.61

Table 6.30. Results obtained via the language association for Irish. Positive deltas are identified in green, while negative ones are in red. When the p-values are lower than 0.01, the cells are highlighted in green.

The five languages with a negative synergy with Irish regarding dependency parsing are Indonesian, Italian, Dutch, Polish, and Portuguese. They were selected via the verb and object strategy and the method concerning MarsaGram all properties. These strategies also provide other candidates which lead to non-significant deltas (i.e.: Maltese, English, Russian, and French).

It is interesting to notice that Lynn et al. (2014) obtained the best results in terms of LAS for Irish with Indonesian. However, their experiments were conducted differently. Delexicalized monolingual corpora from different languages were trained to parse the also delexicalized Irish corpus. In our study, we analyse the combination of different languages, and although Indonesian was identified as a potential candidate with the verb and object strategy, it decreased significantly the LAS score.

Moreover, from the results presented in Table 6.31, it is noticeable that the association with Portuguese has the best MLAS delta result, with a p-value equal to 0.01, even though in terms of LAS, the value is significantly lower than the one from the monolingual model. Even though the results for this specific association are on the threshold of statistical significance, we decided to analyse in detail the impact of the association for each dependency relation and in relation to the sentence lengths, as well as, with the distance between heads and dependents (as presented in Tables 6.32, 6.33, and 6.34 respectively).

The relations which are most negatively impacted when Irish is associated with Portuguese are discourse, parataxis, flat, clausal complement (ccomp), and adverbial clause modifier (advcl). Some improvement can be seen, specially for: vocative, list, vocative particle (case:vocative), appositional modifier (appos), clausal subject in copular constructions (csubj:cop), flat regarding foreign MWE (flat:foreign), and numeric modifier (nummod). It is also possible to notice that root and object (obj) present slight negative deltas, however, nominal subjects (nsubj) are not impacted.

Irish follows the same pattern as Croatian in terms of LAS values tendencies: lower scores for longer sentences and when heads and dependents are distant. It is possible to notice that for longer sentences (more than 40 tokens), there is a slight tendency for better results for the combination of Irish and Portuguese. Furthermore, the delta values displayed in Tables 6.33 and 6.34 show that the negative impact of the combination concern mostly the cases where the heads are positioned after the dependent (deltas lower than 8.00).

DEPREL	gle	gle + por	Delta
vocative	30.00	40.00	10.00
list	61.90	71.43	9.53
case:voc	54.55	63.64	9.09
appos	45.00	52.50	7.50
csubj:cop	73.91	80.43	6.52
flat:foreign	43.75	50.00	6.25
nummod	78.00	84.00	6.00
obl:tmod	32.56	37.21	4.65
csubj:cleft	52.00	56.00	4.00
conj	67.49	68.90	1.41
cop	86.34	87.58	1.24
acl:relcl	66.67	67.17	0.50
nmod	57.06	57.56	0.50
obl:prep	82.87	83.33	0.46
cc	88.38	88.80	0.42
obl	74.02	74.19	0.17
case	94.45	94.45	0.00
compound:prt	20.00	20.00	0.00
dislocated	0.00	0.00	0.00
flat:name	86.73	86.73	0.00
nsubj	88.14	88.14	0.00
orphan	0.00	0.00	0.00
root	90.97	90.75	-0.22
obj	78.07	77.81	-0.26
det	92.95	92.68	-0.27
mark	80.47	80.13	-0.34
xcomp:pred	65.91	65.34	-0.57
mark:prt	81.84	81.27	-0.57
xcomp	71.71	70.93	-0.78
nmod:poss	85.96	85.09	-0.87
amod	68.28	66.99	-1.29
fixed	84.13	82.54	-1.59
punct	76.65	74.84	-1.81
advmod	71.69	69.12	-2.57
compound	7.41	3.70	-3.71
advcl	52.38	48.30	-4.08
ccomp	61.96	57.61	-4.35
flat	64.73	59.38	-5.35
parataxis	41.18	35.29	-5.89
discourse	14.29	0.00	-14.29

Table 6.31. LAS values for each DEPREL label in the Irish test-set regarding the monolingual model and the association with Portuguese. The delta values consist of the difference between the results obtained via the language association and the monolingual ones. The color scale highlights the most positive delta values (green), and the most negative ones (red).

	<=10	<=20	<=30	<=40	<=50	>50
gle	82.77	80.44	78.87	77.90	78.08	69.64
gle + por	81.77	80.17	77.83	77.70	78.20	69.79
Delta	-1.00	-0.27	-1.04	-0.20	0.12	0.15

Table 6.32. LAS results for different sentence lengths. In green are highlighted the positive delta values (i.e.: the difference between the language association result and the monolingual one).

	1	2	3	4	5	>5
gle	91.33	85.91	77.95	77.15	71.97	68.74
gle + por	82.52	75.80	69.80	66.83	63.78	60.67
Delta	-8.81	-10.11	-8.15	-10.32	-8.19	-8.07

Table 6.33. LAS results for different positive distances between heads and dependents (i.e.: head after the dependent). In green are highlighted the positive delta values (i.e.: the difference between the language association result and the monolingual one).

	<-5	-5	-4	-3	-2	-1
gle	45.89	62.07	55.81	61.62	77.61	88.25
gle + por	45.00	54.84	56.69	60.44	77.10	88.13
Delta	-0.89	-7.23	0.88	-1.18	-0.51	-0.12

Table 6.34. LAS results for different negative distances between heads and dependents (i.e.: head before the dependent). In green are highlighted the positive delta values (i.e.: the difference between the language association result and the monolingual one).

c) Lithuanian:

Lithuanian ALSKINS corpus has a training set composed of 2,341 sentences from different genres (fiction, legal, news, and nonfiction). In total, 6 languages were identified via the different corpus-based typological approaches. All these languages have corpora with training sets larger than the Lithuanian one as described in Table 6.35. Thus, the sentences extracted from these corpora to compose the combined ones come exclusively from the training sets. In total, the combined corpora have 4,682 sentences.

It is possible to notice that Latvian (also a Baltic language) is present in the list of selected languages. The other identified languages are also from the Indo-European family (except for Indonesian), mainly from the Slavic genus.

Lithuanian was characterized as type 11 by Hawkins (1983), the same group as 2 Germanic languages (i.e.: Dutch and Swedish). These languages, however, were not selected via our typological methods.

	UD corpus	**Total size (sentences)**	**Training corpus (sentences)**	**Genre**
ces	PDT	87,913	68,495	news, nonfiction, reviews
hrv	SET	9,010	6,914	news, web, wiki
ind	GSD	5,593	4,477	blog, news
lav	LVTB	13,643	10,156	academic, fiction, legal, news, spoken
por	Bosque	9,364	8,328	News
slk	SNK	10,604	8,483	fiction, news, nonfiction

Table 6.35. Description of the UD v.2.7 corpora selected to be combined with Lithuanian ALSKINS training corpus (genres: fiction, legal, news, nonfiction).

The results regarding the language combination experiments are displayed in Table 6.36.

Language pair	**Mean LAS**	**Stdev LAS**	**Delta LAS**	**p_value LAS**	**Mean MLAS**	**Stdev MLAS**	**Delta MLAS**	**p_value MLAS**
lit_ces	77.92	0.10	-0.23	0.00	61.21	0.25	0.73	0.00
lit_hrv	77.94	0.10	-0.22	0.01	61.42	0.12	0.94	0.00
lit_ind	77.99	0.13	-0.17	0.04	61.51	0.21	1.03	0.00
lit_lav	77.42	0.13	-0.74	0.00	60.67	0.27	0.19	0.18
lit_por	78.19	0.11	0.03	0.64	62.35	0.25	1.86	0.00
lit_slk	77.23	0.16	-0.92	0.00	60.82	0.15	0.34	0.01

Table 6.36. Results obtained via the language association for Lithuanian. Positive deltas are identified in green, while negative ones are in red. When the p-values are lower than 0.01, the cells are highlighted in green.

As was the case for Croatian, no significant improvement is observed in terms of LAS. Portuguese and Indonesian are the only 2 languages that present a statistically similar LAS value when compared to the one obtained with the monolingual model. The combinations with the other 4 languages (i.e.: Czech, Croatian, Latvian, and Slovak) present a negative synergy. Portuguese and Indonesian (together with Croatian) were selected via the MarsaGram linear method, however, the closest language to Lithuanian identified with this strategy is Croatian (which provided a negative delta). The other typological methods indicated the other languages from the list, all of them generating a negative synergy.

The scenario is different for MLAS as almost all associations provide significantly better results than the monolingual model (except for Latvian). The best score is achieved when Lithuanian is associated with Portuguese (score statistically higher than the rest), followed by Indonesian and Croatian (all three selected via MarsaGram linear strategy). Latvian was identified as a possible candidate via the association of MarsaGram all properties and head and dependent strategy.

DEPREL	lit	lit + por	Delta
iobj	0.00	33.33	33.33
acl:relcl	72.09	81.40	9.31
flat	38.00	46.00	8.00
nummod:gov	23.08	30.77	7.69
det	84.24	87.88	3.64
advmod:emph	60.06	63.61	3.55
advcl	49.12	52.21	3.09
appos	16.28	18.60	2.32
obl:arg	67.24	69.09	1.85
nmod	78.77	80.38	1.61
advmod	77.82	78.95	1.13
obj	83.51	84.57	1.06
acl	74.80	75.20	0.40
conj	70.83	70.96	0.13
case	90.05	90.05	0.00
cc	85.53	85.53	0.00
ccomp	66.67	66.67	0.00
compound	0.00	0.00	0.00
cop	89.22	89.22	0.00
dep	0.00	0.00	0.00
discourse	0.00	0.00	0.00
flat:foreign	0.00	0.00	0.00
nsubj:pass	64.29	64.29	0.00
punct	82.27	81.88	-0.39
nsubj	83.56	83.05	-0.51
amod	84.20	83.37	-0.83
xcomp	93.75	92.86	-0.89
mark	84.78	83.39	-1.39
root	87.43	85.67	-1.76
obl	52.95	50.59	-2.36
nummod	77.00	74.00	-3.00
parataxis	30.17	25.86	-4.31
csubj	29.69	21.88	-7.81

Table 6.37. LAS values for each DEPREL label in the Lithuanian test-set regarding the monolingual model and the association with Portuguese. The colour scale highlights the most positive delta values (green), and the most negative ones (red).

In Table 6.37, we present the individual LAS score for each dependency relation present in the Lithuanian test-set for the monolingual model and the one trained with the association of this language and Portuguese.

From the 10 dependency relations showing an improvement via the language association, the one with the highest delta is the indirect object (iobj) which has 6 occurrences in the test-set. This relation also presented a positive delta for Hungarian. Moreover, both adnominal clause (acl) and adnominal clause involving relative clauses (acl:relcl) present positive deltas, as it was the case for Croatian and Hungarian (this last one does not use the subtype "relcl"). The object (obj) relation also presents an improvement, while for Croatian the delta was negative, and 0.00 for Hungarian. Moreover, there is an improvement regarding the flat relation (concerning MWE), but not for foreign ones (flat:foreign).

In terms of the observed negative delta, again nominal subject presents a decrease in the LAS when compared to the monolingual model (as it was the case for both Croatian and Hungarian). It is also possible to notice that relations such as root and punctuation have also a negative delta (similar to Croatian). Oblique nominal relation has a negative delta, a phenomenon also observed for Hungarian (while the delta was 0.00 for Croatian). The most negative delta concerns the clausal subject relation (-7.81) which occurs 139 times in the test-set. For Hungarian, the delta was positive (2.70) while 0.00 for Croatian.

Thus, it seems that overall there are more similitudes in terms of the dependency relation deltas between Croatian and Lithuanian, even though many differences can still be found. In both cases, the comparison was done between the monolingual model and the association which provided the best MLAS score but showed no improvement in terms of LAS (as no combination provided any significantly positive delta for this metric).

The LAS results in terms of sentence length and head and dependent distances are presented in Tables 6.38, 6.39, and 6.40.

	<=10	<=20	<=30	<=40	<=50	>50
lit	80.39	78.7	74.73	74.83	71.35	65.59
lit + por	80.66	77.98	75.2	75.26	73.37	68.81
Delta	0.27	-0.72	0.47	0.43	2.02	3.22

Table 6.38. LAS results for different sentence lengths. In green are highlighted the positive delta values (i.e.: the difference between the language association result and the monolingual one).

	1	2	3	4	5	>5
lit	77.76	74.04	72.14	68.65	65.04	66.93
lit + por	77.41	73.92	71.62	70.07	65.19	66.88
Delta	-0.35	-0.12	-0.52	1.42	0.15	-0.05

Table 6.39. LAS results for different positive distances between heads and dependents (i.e.: head after the dependent). In green are highlighted the positive delta values (i.e.: the difference between the language association result and the monolingual one).

	<-5	-5	-4	-3	-2	-1
lit	66.25	66.96	72.21	77.93	77.51	85.00
lit + por	65.97	68.00	73.33	76.79	78.32	85.43
Delta	-0.28	1.04	1.12	-1.14	0.81	0.43

Table 6.40. LAS results for different negative distances between heads and dependents (i.e.: head before the dependent). In green are highlighted the positive delta values (i.e.: the difference between the language association result and the monolingual one).

Overall, when the LAS results are analysed individually, the tendencies follow the same patterns as observed for Croatian and Hungarian. In terms of sentence length, it is possible to observe that, again, the highest deltas concern the longest sentences. However, the delta is also positive for shorter sentences (less than 10 tokens). The only negative delta is observed for sentences containing from 10 to 20 tokens. When the heads are positioned after the dependents, it is possible to see that the language association provides better LAS scores than the monolingual one when the distance is either 4 or 5 tokens (while slightly negative for larger distances). Again, it is possible to observe that when the distance is one or two tokens, the monolingual model provides better results. On the other hand, when the heads are located before the dependents, the only negative deltas concern the cases where the head is far away from the dependent, or 3 tokens away, a slightly different tendency than the one observed for Croatian and Hungarian.

d) Maltese:

Maltese has the largest number of selected candidates for the dependency parsing improvement experiments (14 in total). The corpora from which the sentences were extracted to compose the combined training sets are described in Table 6.41. Maltese training set (MUDT corpus) has 1,123 sentences. From the selected corpora, only Thai does not have enough sentences to respect the established ratio (i.e.: 50% from Maltese and 50% from language 2). Thus, the total size of the combined corpora is 2,246 sentences, except for the one composed of Maltese and Thai which has a final size of 2,123 sentences.

	UD corpus	**Total size (sentences)**	**Training corpus (sentences)**	**Genre**
dan	DDT	5,512	4,383	fiction, news, nonfiction, spoken
deu	GSD	15,590	13,814	news, reviews, wiki
ell	GDT	2,521	1,162	news, spoken, wiki
eng	GUM	5,961	4,287	academic, blog, fiction, government, news, nonfiction, social, spoken, web, wiki
fra	GSD	16,341	14,449	blog, news, reviews, wiki
hrv	SET	9,010	6,914	news, web, wiki
ita	ISDT	14,167	13,121	legal, news, wiki
nld	Alpino	13,578	12,264	news
pol	PDB	22,152	17,722	fiction, news, nonfiction
rus	SynTagRus	61,889	48,814	fiction, news, nonfiction
slv	SSJ	8,000	6,478	fiction, news, nonfiction
swe	Talbanken	6,026	4,303	news, nonfiction
tha	PUD	1,000	-	news, wiki
tur	IMST	5,635	3,664	news, nonfiction

Table 6.41. Description of the UD v.2.7 corpora selected to be combined with Maltese MUDT training corpus (genres: fiction, legal, news, nonfiction, wiki). PUD corpora correspond only to test-sets composed of 1,000 sentences.

The closest language to Maltese in our language-set in terms of phylogenetic features is Arabic. However, this language was not identified as a possible candidate by the corpus-based typological approaches. Instead, our methods selected 12 Indo-European languages (i.e.: 5 Germanic, 4 Slavic, 2 Romance, 1 Greek), Thai, and Turkish. The experiment results concerning the association of Maltese with the listed languages are presented in Table 6.42.

It is noticeable that Maltese benefits from the language association experiments as in 11 cases LAS is statistically improved, and for 12 combinations, MLAS is significantly increased. For both metrics, the best score is obtained when Maltese is combined with French (a score significantly higher than the one obtained with the second-best combination: Maltese and Greek). The improvements are the highest ones observed for the low-resourced languages selected in this thesis (i.e.: 2.51 for LAS, and 4.05 for MLAS).

French was selected as a candidate via the MarsaGram all properties method which established Greek (second-best scores) as the closest language to Maltese. This method also presented German and Swedish as potential languages to be combined with Maltese, and in both cases, the deltas regarding LAS and MLAS are positive.

	Mean LAS	**Stdev LAS**	**Delta LAS**	**p_value LAS**	**Mean MLAS**	**Stdev MLAS**	**Delta MLAS**	**p_value MLAS**
mlt_dan	74.02	0.13	0.23	0.06	58.38	0.23	0.88	0.00
mlt_deu	74.36	0.13	0.57	0.00	58.95	0.18	1.45	0.00
mlt_ell	76.18	0.21	2.39	0.00	61.17	0.28	3.68	0.00
mlt_eng	75.21	0.12	1.42	0.00	60.01	0.21	2.51	0.00
mlt_fra	76.31	0.06	2.51	0.00	61.55	0.17	4.05	0.00
mlt_hrv	75.52	0.17	1.73	0.00	59.65	0.30	2.15	0.00
mlt_ita	75.64	0.10	1.85	0.00	60.23	0.16	2.73	0.00
mlt_nld	75.11	0.18	1.31	0.00	59.62	0.25	2.12	0.00
mlt_pol	73.71	0.13	-0.08	0.48	57.27	0.34	-0.22	0.29
mlt_rus	75.34	0.03	1.54	0.00	60.09	0.19	2.59	0.00
mlt_slv	74.12	0.20	0.32	0.03	57.97	0.20	0.47	0.02
mlt_swe	74.52	0.12	0.73	0.00	59.29	0.26	1.80	0.00
mlt_tha	75.62	0.11	1.83	0.00	60.42	0.33	2.93	0.00
mlt_tur	74.39	0.06	0.59	0.00	58.60	0.07	1.10	0.00

Table 6.42. Results obtained via the language association for Maltese. Positive deltas are identified in green, while negative ones are in red. When the p-values are lower than 0.01, the cells are highlighted in green.

The combination with Danish did not provide any LAS improvement but increased MLAS. Furthermore, the two languages which did not present any positive synergy (or negative) with Maltese are: Polish and Slovenian. Polish was selected with the verb and object method, while Slovenian, with the MarsaGram linear strategy. These two methods proposed also other languages which provided some improvement for both metrics. The closest language to Maltese identified via the verb and object method was Russian (with an improvement of 1.54 for LAS, and 2.59 for MLAS), and for MarsaGram linear, the closest language is Croatia (positive delta of 1.73 for LAS, and 2.15 for MLAS).

It is interesting to notice that from the selected low-resourced languages, Maltese is the only one that is not represented in multilingual BERT, however, it is the one with the best improvements in terms of dependency parsing when corpora are combined. Moreover, Maltese MUDT corpus is the only one without FEATS annotation, thus, this criterion does not affect the calculation of MLAS.

DEPREL	mlt	mlt + fra	Delta
fixed	6.15	47.69	41.54
appos	13.33	36.67	23.34
parataxis	35.09	54.39	19.30
advcl	35.61	47.73	12.12
list	44.44	55.56	11.12
conj	55.42	65.44	10.02
iobj	0.00	10.00	10.00
nmod	25.00	34.43	9.43
ccomp	53.89	62.28	8.39
obl:agent	4.17	12.50	8.33
advmod	55.96	62.18	6.22
acl	56.70	61.68	4.98
nsubj	64.91	69.12	4.21
nummod	72.41	75.86	3.45
aux:neg	85.29	88.24	2.95
obl:arg	2.74	5.48	2.74
root	86.10	88.42	2.32
nsubj:pass	4.08	6.12	2.04
cop	72.50	74.38	1.88
nmod:poss	88.16	89.92	1.76
cc	84.11	85.65	1.54
mark	84.69	86.15	1.46
det	95.87	97.31	1.44
case	90.12	91.20	1.08
aux	90.20	91.18	0.98
punct	72.74	73.39	0.65
advmod:neg	89.24	89.87	0.63
amod	80.85	81.09	0.24
cop:expl	0.00	0.00	0.00
csubj	0.00	0.00	0.00
dep	0.00	0.00	0.00
dislocated	0.00	0.00	0.00
expl	0.00	0.00	0.00
vocative	0.00	0.00	0.00
obl	60.33	59.88	-0.45
obj	68.85	67.76	-1.09
case:det	96.04	94.53	-1.51
aux:part	96.72	95.08	-1.64
xcomp	62.03	59.31	-2.72
flat:name	93.85	89.23	-4.62
aux:pass	90.00	80.00	-10.00
discourse	58.97	47.44	-11.53
flat	36.36	18.18	-18.18
compound	21.05	2.63	-18.42

Table 6.43. LAS values for each DEPREL label in the Maltese test-set regarding the monolingual model and the association with French. The color scale highlights the most positive delta values (green), and the most negative ones (red).

When each label present in the Maltese test-set is analysed in terms of LAS improvement (as displayed in Table 6.43), it is possible to notice that the number of relations with positive delta corresponds to 63.6% (28 out of the 44 attested relations). The number of negative deltas is much lower (10 out of 44, 22.7%). The highest improvements are observed for fixed, parataxis, and appositional modifier (appos). Nominal subject (nsubj) which presented negative deltas for Croatian, Hungarian and Lithuanian, are improved for Maltese when it is combined with French. The identification of the root is also improved with this language association, however, as it was seen before for Croatian and Irish, objects present a negative delta.

The analysis of LAS in terms of sentence length and distance between heads and dependents for Maltese confirms what was observed for other languages: LAS values are lower for longer sentences and when heads are more distant from dependents (right or left side) as it can be seen in Tables 6.44, 6.45, and 6.46.

	<=10	<=20	<=30	<=40	<=50	>50
mlt	75.41	75.43	72.61	72.02	69.97	69.57
mlt + fra	76.91	78.25	74.91	74.24	70.93	73.02
Delta	1.50	2.82	2.30	2.22	0.96	3.45

Table 6.44. LAS results for different sentence lengths. In green are highlighted the positive delta values (i.e.: the difference between the language association result and the monolingual one).

	1	2	3	4	5	>5
mlt	69.59	66.62	60.59	56.02	47.37	60.57
mlt + fra	70.68	69.15	62.73	61.48	57.20	67.20
Delta	1.09	2.53	2.14	5.46	9.83	6.63

Table 6.45. LAS results for different positive distances between heads and dependents (i.e.: head after the dependent). In green are highlighted the positive delta values (i.e.: the difference between the language association result and the monolingual one).

	<-5	-5	-4	-3	-2	-1
mlt	55.70	64.17	68.26	72.77	78.64	88.44
mlt + fra	62.34	68.64	70.05	72.23	79.03	89.05
Delta	6.64	4.47	1.79	-0.54	0.39	0.61

Table 6.46. LAS results for different negative distances between heads and dependents (i.e.: head before the dependent). In green are highlighted the positive delta values (i.e.: the difference between the language association result and the monolingual one).

When the deltas are considered, it is possible to see that in all cases there is an improvement in terms of LAS when Maltese is combined with French. In terms of sentence length, although the lowest delta concerns sentences with 40 to 50 tokens, the highest one is attested for longer sentences (more than 50 tokens). Maltese and Lithuanian were the only languages for which an improvement was observed also for sentences with less than 10 tokens, however, the delta value is higher for Maltese (1.50 versus 0.27).

Another similarity observed between Maltese and the other languages concerns the tendency of LAS values to be improved for the combined corpora when the heads are far from the dependents on the right side (i.e.: after the dependents). However, when the heads are positioned before the dependents, Maltese has a unique behaviour: for the other languages, deltas tended to be higher when heads and dependents were closer, while for Maltese, it is the opposite. The highest delta (6.64) is attested for heads positioned more than 5 tokens away from the dependents.

6.4. General Discussion

In this section, we provided 4 different corpus-based typological analysis of the 24 official European Union languages using the selected methods from the previous section which presented the best results regarding the improvement of dependency parsing results. The EU languages were classified using a clustering algorithm together with 10 other worldwide languages (from PUD collection), a total of 34 languages. Each corpus-based strategy considers different syntactic attested structures which are extracted from annotated corpora with the same size in terms of sentences.

The main objective of the new typological classifications was the identification of the closest languages to the 4 identified EU low-resourced languages regarding dependency parsing criteria (i.e.: the size of available UD corpora and UDify LAS and MLAS results): Hungarian, Irish, Lithuanian, and Maltese. Moreover, Croatian was also selected to check the efficiency of corpora-combination using the typological strategies for a language with more data and better parsing results than the low-resourced ones.

Therefore, each typological method provided different propositions in terms of language combination (i.e.: closest ones in terms of distances from the dissimilarity matrices and the obtained dendrograms).

From the dependency parsing results obtained via the corpora combination, it was possible to observe that not all languages showed statistically significant improvements in terms of LAS and MLAS. Croatian and Lithuanian results were improved only regarding MLAS, while Hungarian and Maltese presented better results for both metrics. Irish did not show any improvement for both metrics, although the combination with Portuguese obtained the best MLAS score with a p-value equal to 0.01.

In terms of corpus size, it is possible to notice that languages with larger training corpus tend to present no improvement in terms of LAS. Croatian has 6,914 sentences in its training-set, Irish has 4,005, and Lithuanian, 2,341. On the other hand, Hungarian and Maltese have smaller training-sets (910 and 1,123 sentences respectively). However, at least some improvement was observed in terms of MLAS for Croatian and Lithuanian.

When comparing the best scores obtained with the language associations with the UDify results presented by Kondratyuk and Straka (2019) with their multilingual model (Table 6.15), it is possible to observe that for:

- Croatian: the UDify multilingual model provides a better LAS than the monolingual one trained only with Croatian (baseline). As no combination resulted in LAS improvement, UDify multilingual model is still the most accurate. However, in terms of MLAS, the monolingual model already presented a better score, and as with the association of Croatian and Russian a positive delta was obtained, this combination provides the best result for this metric.
- Hungarian: the score of the multilingual model proposed by Kondratyuk and Straka (2019) was higher than the monolingual one. Even though LAS is improved with the combination of Hungarian with other languages (especially Latvian), it is still lower than the state-of-art. For MLAS, it is the same case as Croatian: the monolingual model has already a better score than the multilingual one and is mostly improved via the combination of Hungarian and Dutch.
- Irish: For this language, the monolingual model presented better results than the multilingual one (Kondratyuk and Straka, 2019). Only Portuguese may be considered as having a tendency for increasing MLAS.
- Lithuanian: In their study, Kondratyuk and Straka (2019) did not provide results for the corpus that was considered in our experiments.

- Maltese: the results obtained with the combination of this language with French are higher than the ones observed for the monolingual and multilingual models.

It is interesting to notice that the best associations did not follow the phylogenetic classification of the associated languages. The only case where a language from the same family and genus was identified as the best pair was for the combination between Croatian and Russian. However, Slovenian, which is even genetically closer to Croatian, did not improve the results when this association was tested.

When analysing which corpus-based methods succeed in identifying the best associations, we observed that MarsaGram all properties and MarsaGram linear patterns (both with cosine distance) provided the best results (3 and 2 correct results respectively). The optimized association of MarsaGram all properties and head and dependent method was responsible for the good identification of the pair Hungarian and Dutch (MLAS improvement). The verb and object position strategy did not provide any correct result. This specific method was tested as it showed some interesting results for some languages which did not present a good correlation between language distance and dependency parsing results (specially the ones with no close languages in terms of phylogenetic features). However, in the cases where improvements were attested, all methods present at least one candidate with a positive delta.

Furthermore, we could observe that the increase in LAS does not determine an improvement in terms of MLAS, and not always the language-pair providing the best LAS is also the one with the best MLAS value. Additionally, as MLAS metric also consider UPOS and FEATS, it is possible to observe that usually in cases where this score is improved, there is usually at least an increase in the FEATS as can be seen in Table 6.47. For Croatian and Irish, the UPOS is even significantly decreased.

	Mean UPOS	**Stdev UPOS**	**Delta UPOS**	**p_value UPOS**	**Mean FEATS**	**Stdev FEATS**	**Delta FEATS**	**p_value FEATS**
hrv_rus	98.16	0.02	-0.08	0.00	95.32	0.04	0.30	0.00
hun_nld	96.13	0.05	0.14	0.00	90.80	0.14	0.78	0.00
gle_por	93.93	0.06	-0.34	0.00	72.37	0.22	0.70	0.00
lit_por	95.28	0.06	0.10	0.05	86.45	0.13	1.44	0.00
mlt_fra	92.77	0.06	0.28	0.00	99.97	0.01	-0.03	0.00

Table 6.47. UPOS and FEATS results for the best-identified language combinations.

Maltese present a slight decrease in its FEATS value as its original corpus does not present any annotation for this but French does. Thus, a solution would be removing FEATS tags from the French corpus before combining it with Maltese.

Finally, we have analysed for each language the LAS improvement for each label present in the test-set and how this metric varies in terms of sentence length and distance between heads and dependents. What we could observe is that even for cases where the overall LAS metric is not significantly improved, the delta is positive for very long sentences (i.e.: more than 50 tokens). Thus, the language combinations tend to be better when sentences are more complex to be analysed. We also noticed a tendency of better LAS results for heads that are positioned far away from the dependents specifically when the heads come after the dependent in the sentence. On the other hand, when the heads precede the dependents, usually better scores are obtained when they are positioned nearby the dependents (except for Maltese).

Therefore, it is possible to affirm that the potential of language association for dependency parsing improvement is confirmed specially for languages with very small training corpora (around 1,000 sentences) and that, in general, MLAS metrics were more improved than LAS. However, it has been shown that in all cases, the association of 2 languages is a good way to improve LAS results for more complex sentences. Moreover, it seems that the information provided by MarsaGram (considering all properties or only the linear one) is the most useful for the identification of the best language-pairs.

7. Conclusion

The main objective of this thesis was to propose innovative corpus-based typological analyses of syntactic structures and examine them as potential strategies to improve dependency parsing results via corpora combination of close languages. In order to achieve it, we divided this work into three main sections. First, we presented the different typological approaches and the language classifications obtained for each method in comparison to the classical phylogenetic characterization and the state-of-the-art syntactic classification attained with the comparison of word-order features from typological databases. Secondly, we described the dependency parsing experiments conducted with corpora associations of 20 worldwide languages (10 official European Union ones) using parallel corpora and compared the language-pairs synergy regarding the improvement of dependency parsing metrics to the language classifications obtained in the first step. Finally, we extended the typological analysis with the best corpus-based methods identified in the previous part to the other European Union languages and provided the results of dependency parsing improvement for a set of low-resourced languages and Croatian.

In the introductory section regarding the theoretical background and related work, we started with the presentation of general typological principles and detailed the major contributions in the field of syntactic typology provided by Greenberg (1963), Hawkins (1983), and Dryer (1992). As presented by Moravcsik (2012), the main objective of typologists is to find generalities that can be observed either in all human languages, or for the majority of them, or even only for a specific subset of them. Usually, typological studies concern both universals and types and confront the possibility of the occurrences of determined phenomena with reality. Greenberg was a pioneer in proposing a set of universals using implicational statements, most of them related directly or indirectly to the position of verbs and objects. Hawkins (1983) developed Greenberg's research for a larger number of languages, extended the analysis to other grammatical elements questioning the supremacy of the verb and object position, and proposed an alignment between his observations and the X-bar theory of generative grammar. On the other hand, Dryer (1992) reviewed Hawkins' statements proposing a theory based on the branching tendencies (left or right) which is more suitable for explaining attested phenomena than the usage of the concepts of heads and dependents. In these studies, languages are compared and classified in types according to "basic word orders" as described by Hawkins (1983). Thus, it means that only the most general word order phenomena are considered.

In terms of corpus-based typology, Levshina (2022) presented an extended overview of how quantitative methods can provide valuable typological information in addition to classic theoretical studies, especially concerning studies focusing on the analysis of language complexity, however, the usage of corpus-based approaches in Natural Language Processing was not examined in this study. Ponti et al. (2019) focused on the usage of typology for improving NLP scores in various tasks involving cross-lingual transfer, showing that multilingual models tend to present better results than monolingual ones. When it comes to the specific NLP task of dependency parsing, by analysing the related work concerning typological methods for parsing scores, we showed that, in most cases, language comparison is conducted with the usage of syntactic features available in typological databases and with quantitative analysis of part-of-speech sequences. These studies are usually focused on languages with no or low resources, and delexicalized corpora are combined. Moreover, the tools which are trained do not correspond to the state-of-the-art regarding dependency parsing (i.e.: deep-learning systems associated with language models).

The usage of typological databases to compare languages has the advantage of not requiring the availability of annotated corpora, however, not all languages have detailed descriptions regarding all the syntactic features. On the other hand, the studies concerning the analysis of part-of-speech sequences require at least some portion of text to be annotated with this type of information. The corpus-based typological strategies presented in this thesis are based on the analysis of datasets annotated with part-of-speech and dependency relations, thus, a more complex type of annotation. However, as of today, in version 2.11 of the Universal Dependencies collection (released on November 2022), 138 languages have at least one corpus with the necessary data for our methods to be applied. Another observation regarding the usage of typology for NLP applications is the lack of detailed analysis regarding the different contributions of the different factors influencing the observed positive synergies.

The first contribution of this thesis was the typological analysis of languages with quantitative methods which consider a great variety of syntactic phenomena and which have never been considered as strategies for dependency parsing improvement. Moreover, we provided a detailed comparison of these methods with the genealogical classification and with the one obtained using typological databases (i.e.: URIEL, Littell et al., 2017). Four different strategies have been chosen. Two of them are based on the extraction of syntactic patterns provided by the MarsaGram software (Blache et al., 2016) that allows languages to be compared syntactically using a quantitative method of extraction of context-free grammars (CFG) from treebanks: one method is based on all properties extracted from corpora (i.e.: linear, require, exclude, and unicity), and the other one considers just the linear patterns (i.e.: if one element precedes the other at the sentence level). The patterns extracted with this tool concern elements which are part of the same syntactic sub-structre. While the linear patterns describe specific word order phenomena, the other properties are linked to other attested phenomena (i.e.: if one element excludes or requires the other, and if the same element can be attested more than once inside the subtree). The third method is a quantitative extension of the analysis proposed by Hawkins (1983) as it is based on the relative position of heads and dependents at the surface level. While Hawkins based his analysis on the most general word order of nominal terms and their heads, our analysis considers all attested heads and dependents inside the selected corpora. Finally, the fourth corpus-based strategy is based on the relative position of verbs and objects. Dryer (1992) showed the correlation between the order of these two elements and the position of other ones, thus, the idea is to analyse if the quantitative analysis of solely these two components can be used for our NLP purpose.

These four different strategies were applied to a language-set composed of 20 languages (9 distinct linguistic families) which compose the Parallel Universal Dependencies collection (1,000 sentences per language) and contain 10 European Union languages. We decided to use these parallel corpora to avoid bias regarding size and genre as all corpora present the same semantic content. In quantitative methods for language comparison, usually, languages are represented by vectors composed by features associated with frequency values which are, then, compared in terms of distances. In this thesis, we decided to consider two different distance metrics (i.e.: Euclidean and cosine) as they compare vectors differently, thus, providing different classifications. With the dissimilarity matrices built with the calculated distance between the languages, we classified the languages in clusters using the Ward linkage algorithm (Ward, 1963).

As expected, the dendrograms obtained via the clustering analysis of the distance matrices showed that each proposed corpus-based strategy presents a different classification of the selected languages. Moreover, each distance metric provides a different classification. However, in every case, some similarities can be found between the classification obtained via the new methods and the ones concerning phylogenetic characteristics, syntactic features of typological databases, and the language-types proposed by Hawkins (1983). The PUD collection is composed of 12 Indo-European languages with more than 2 languages for the Romance, Slavic, and Germanic genera. On the other hand, the other 8 languages do not have closely related ones. Thus, the similarities found concern mainly the abovementioned genera, however, their clusters do not follow exactly the phylogenetic classification.

When all MarsaGram properties are considered (a total of 158,755 patterns of which only 78 occur in all 20 corpora), we observed that with Euclidean distances, Romance languages are clustered together in an isolated group. The Germanic cluster can also be identified but it also contains Russian. The other 2 Slavic languages (i.e.: Czech and Polish) form a distinct group that is close to the Germanic one. The languages with no genealogically close related ones form a large central cluster, with no clear distribution between VO and OV languages and not following Hawkins classification. When the cosine metric was used to compare languages, again the Romance cluster could be easily identified, however, forming a larger group with Germanic languages (except Icelandic) and Indonesian. This latter is classified by Hawkins (1983) as type 9 as are all Romance languages. All Slavic languages are grouped in the same cluster but not close to Romance and Germanic clusters. Icelandic was grouped with Finnish and Turkish, being part of the large cluster of non-related languages. Chinese and Korean also formed a distinct cluster, separated from all the other PUD languages.

Most of the patterns extracted using MarsaGram correspond to the exclude property (81.37%), only 13.38% (21,242 patterns) concern the linear property, while unicity and require patterns represent around 5%. Of the extracted linear patterns, only 10 of them are present in all corpora and they are basically composed of word order structures concerning nominal subjects and objects, coordinative conjunctions and roots, and appositional modifiers and roots in subtrees whose heads are either verbs, nouns, or proper nouns.

In the dendrograms obtained when only linear patterns are considered, we could observe that the obtained clusters were less coherent to the phylogenetic classification than the classification obtained with all MarsaGram patterns. In the Euclidean dendrogram, while Italian and Spanish form a specific cluster, French and Portuguese are classed with English and Swedish. German was positioned in a cluster with the Slavic languages and Thai. In this scenario, Icelandic and Japanese are the most isolated languages. In the cosine dendrogram, it was also possible to identify the mix between Germanic and Romance languages, however with Portuguese forming a specific sub-cluster with Spanish. Again, German was positioned with the Slavic languages and Thai, however, Icelandic was not isolated, forming a cluster with Finnish. Furthermore, it was possible to identify a separate cluster formed by OV languages and Chinese.

As previously mentioned, the third typological method consisted in identifying patterns regarding the relative position of heads and dependents (i.e.: heads preceding dependents or following them). In total 2,890 patterns were identified, with a balanced number between cases where the heads come after the dependents and cases where they precede them. Moreover, 98 patterns occur in all 20 PUD corpora and are related to adverbial modifiers and coordinative conjunctions preceding the heads, subjects preceding verbs (coherent to the feature provided by URIEL database), appositional modifiers following the heads, and punctuation either preceding or after the heads. In terms of the overall tendency for being left or right-branching (i.e.: dependents preceding the head and vice-versa), we observed that Arabic, Thai, Indonesian, and Japanese have more patterns where the head precedes the dependent. Polish and Icelandic have a relatively similar distribution of right and left-branching patterns, while the other languages have a higher percentage of left-branching ones, especially German, Turkish, and Korean. These tendencies do not correspond to the VO/OV classification, although the left-branching tendency is strongly present in 2 OV languages, and in German, which is described as a language with no dominant order regarding these two elements.

The analysis of the dendrograms regarding the head and dependent relative position strategy showed that again the Romance cluster is clearly identified, with the proximity of Portuguese and Spanish well characterized. For this specific corpus-based approach, both dendrograms (i.e.: Euclidean and cosine) present quite similar representations. A large cluster can be identified containing two sub-clusters: German and Slavic. However, Icelandic is positioned in both cases with the Slavic languages, closer to Polish. Furthermore, OV languages form a cluster that also includes Finnish and Turkish as a specific sub-cluster (although being different in terms of overall right and left-branching tendencies).

When only verb and object positions are analysed, 13 patterns of objects preceding verbs are identified, and 12 cases of the opposite order are found. Only Indonesian does not present any pattern where the verb is positioned after the object, and solely Turkish and Korean do not present any case where the object is after the verb. In summary, the PUD OV languages present very low or zero VO patterns and vice-versa. German has relatively balanced percentage between VO and OV occurrences (which is coherent with its status as no dominant order language described in WALS). For both cases, the frequency of occurrences is much higher in features concerning dependents which are nouns, pronouns, and proper nouns. The clear separation between VO and OV languages was also observed in both dendrograms, with German being classed closer to the OV ones. The obtained classifications do not follow the phylogenetic characteristics. The Euclidean dendrogram presented a large cluster formed of Arabic, Slavic, and Romance languages, while Germanic languages (except German) form a separate cluster with Finnish and Indonesian. Thai and Chinese are part of a specific cluster in the VO group. On the other hand, the cosine dendrogram presented a specific group formed by Finnish, Czech, and French. Other Romance languages are classed together in a cluster containing also Arabic and Thai. In this scenario, Chinese and Indonesian are associated with the Germanic group (except German).

Each corpus-based method extracted different syntactic features, thus the number of patterns analysed in each case varied considerably (the largest number concerns the MarsaGram all properties strategy). Moreover, we also observed that the number patterns attested in each language differed significantly which is also linked to the total number of tokens present in each corpus. We also noticed that the language classifications provided by the dendrograms are a useful way for visualizing the data provided by the dissimilarity matrices, however, when it comes to the identification of the closest language-pairs, sometimes the information obtained via the graphs does not correspond to the one extracted from the distance matrices as the dendrograms are built with the comparison of the variance regarding the distances between clusters via the Ward algorithm, not necessarily positioning side-by-side the closest languages.

The second major contribution of this thesis concern the language combination experiments for dependency parsing improvement and the analysis of the obtained results in correlation with the language distances provided by each corpus-based typological strategy. For this aim, we first established the baseline which is composed of the LAS and MLAS scores obtained with monolingual models trained with the different PUD corpora. The selected deep-learning tool is the UDify software developed by Dan Kondratyuk and Milan Straka (2019) which is capable of predicting lemmas, part-of-speech, morphological features, and dependencies relations using multilingual BERT (Devlin et al., 2018) in its embedding, encoder and projection layers. It was possible to observe that even though monolingual training corpora had the same size in terms of sentences, results varied considerably.

The best monolingual LAS scores were achieved for Japanese, French, and Spanish (higher than 91.00), while the worst ones were obtained for Thai and Chinese (lower than 75.00). For MLAS, the best results concern the same languages as was the case for LAS, while Icelandic has the worst score (lower than 50.00). While some languages present some consistency in terms of their positioning in terms of LAS and MLAS results among PUD languages, some tend to perform much better for one specific metric. It is the case of Icelandic, German, Czech, and Polish which present better positioning in terms of LAS, while for Korean and Indonesian the opposite is observed. In general, Romance languages and Japanese have good scores for both dependency parsing evaluation metrics, which is probably due to their syntactic characteristics (with more rigid word order structures). When these results were correlated with the language representation size in multilingual BERT language-model, we found that there is a strong Spearman's correlation between LAS and the language size in the language-model, while in terms of MLAS, a moderate Spearman's correlation was attested. Thus, the influence of how well the language is represented in mBERT influences more the LAS metric. However, it does not mean that the language with the largest representation has the best score. Moreover, no moderate or strong correlation has been found between the LAS and MLAS results and the number of labels in each corpus in terms of part-of-speech, morphosyntactic features, and dependency relations.

Regarding the corpora association experiments, PUD languages were combined in pairs, a total of 380 combinations. The size of the training-set doubled (from 600 to 1,200), while the development and test-sets remained the same as for the monolingual experiments. In terms of LAS, Russian and Finnish were the languages with the highest number of cases where the combination with other languages provided a statistically significant positive delta (i.e.: the difference between the score of the language association and the monolingual one). On the other hand, for Hindi, Japanese, and Korean, no significant improvement was obtained. Some negative synergy was also observed, especially for Korean (14 cases), Japanese (6 cases), and Thai (6 cases). In terms of MLAS, again no improvement was achieved for Japanese and Korean, nor for Thai. These languages were also the ones with the highest number of cases where a negative delta was observed (12, 17, and 7 respectively). On the other hand, for Arabic and Turkish, a significant improvement was obtained for all the tested combinations (i.e.: with every other 19 PUD languages). We also observed that while in some cases the genealogical proximity can define the language with the best LAS and MLAS delta (e.g.: Portuguese and Spanish), it is not always the case (e.g.: for English, the best results concern the association with French, not with another Germanic language). Regarding the magnitude of the obtained improvements, the best LAS delta was 1.95 (Finnish and Russian combination in comparison with Finnish monolingual model), while the association of Russian and French provided the best MLAS improvement (4.11) in comparison with Russian alone.

With the set of delta values obtained for each language and the language distances provided by the dissimilarity matrices, we calculated both Pearson's and Spearman's correlation scores and analysed which corpus-based method provided the greatest number of moderate and high correlations for both LAS and MLAS. We observed that no corpus-based method correlates with the observed synergy in terms of parsing results for all languages. However, it was possible to observe that MarsaGram linear strategy using the cosine distance metric presented the best results (i.e.: moderate or strong Pearson's correlation for 10 out of the 20 PUD languages). This specific method also presented the highest number of moderate and strong correlations for MLAS (i.e.: for 9 out of the 20 PUD languages), together with the MarsaGram all properties one. These correlation results are better than the ones achieved for the state-of-the-art typological method (i.e.: using syntactic information from typological databases). The MarsaGram linear method with cosine distances also presents at least some correlation for 3 other languages in terms of LAS, and 6 other cases for MLAS. Thus, it seems that the linear patterns concerning elements inside the same syntactic subtree have some influence in the parsing results when different languages are combined. Overall, OV languages presented less correlation than the VO ones, however, they correspond to only 4 of the 20 selected languages. Also, although languages with genealogical close ones in the language-set tend to provide better correlation scores, this was not always observed (e.g.: Italian).

To test how dependency parsing results from language combination experiments correlate with the language distances obtained with each corpus-based approach for other dependency parsing architectures, we conducted the same type of experiments described previously with UDPipe 1.0 which does not rely on language-models. In this case, almost no moderate or strong correlation could be found, indicating that the proposed typological analyses are more efficient for deep-learning tools based on the same architecture as UDify.

As previously mentioned, MarsaGram linear method was identified as the one with the best overall results in terms of correlation, however, we also decided to verify which method selects the best choice of language-pair conducting to the highest positive empirical delta. Thus, for each corpus-based strategy, we verified if: a) the closest languages provided the best delta (or statistically similar to it), b) at least a statistically significant improvement, c) if it significantly decreased LAS and MLAS. Each method was examined alone and in association with others (via linear regression). In terms of LAS, we observed that the highest number of right choices regarding the language pairs providing the best scores (i.e.: 9 out of the 20 PUD languages, with no choice conducting to a negative synergy) was obtained via a specific association of both MarsaGram all properties and head and dependent methods (Euclidean distances). On the other hand, for MLAS, the best results were obtained with MarsaGram linear (cosine) (i.e.: the right choice for 9 languages, however with one language-pair providing a negative delta). Besides this method, when all MarsaGram properties are considered, results are also interesting: the same number of right choices, 2 negative deltas but a higher number of overall significant positive deltas. When compared to the state-of-the-art typological strategy (i.e.: using lang2vec language vectors composed with syntactic features from typological databases), the combination of MarsaGram all and head and dependent strategies provides better results for LAS, however, the lang2vec method is better for MLAS (i.e.: 10 right-choices and no negative synergy). We also observed that in many cases where the best method did not select the best alternative for LAS improvement, the verb and object relative position strategy (cosine) provided a better candidate.

Thus, we can conclude that when the typological classifications are evaluated in terms of providing valuable information for dependency parsing improvement, the syntactic information provided by MarsaGram is the most effective. MarsaGram linear strategy (cosine) was identified as the one with the better scores in terms of correlation and is the method with better results for identifying the best language-pairs. In terms of LAS, using all MarsaGram properties associated with the relative position of heads and dependents was the best-identified method even though these two strategies did not present the best correlation values. Although lang2vec method proved to be the most efficient strategy for MLAS improvement, it presents the limitation of being dependent on the availability of information regarding the syntactic features in URIEL database. For the PUD language-set, the 20 languages have values for 41 out of the 103 average syntactic features, however, when the other EU languages are added to this set, there is no feature with valid values which are common to all of them.

The third contribution of this thesis concerns the extension of the empirical typological analyses to all the remaining EU languages which are not present in the PUD collection. The three corpus-based methods which provided the best results in terms of correlations and identification of the best language-pairs (i.e.: MarsaGram linear with cosine distances, combination of MarsaGram all and head and dependent methods with Euclidean distances, MarsaGram all properties with cosine distances) were applied to the ensemble of the PUD and EU languages (34 in total). Moreover, we decided to include the verb and object relative position strategy as it provided interesting results for exceptional languages (i.e.: the ones for which the other selected methods did not select the best language-pair in terms of LAS). The aim was to provide new typological classifications of all EU languages (in a worldwide scenario as non-EU ones were also included) and test the selected methods in a less controlled scenario as non-parallel corpora were used for the 14 non-PUD languages. Moreover, to avoid any bias regarding the size of the corpora, we selected 1,000 sentences of the available UD corpora regarding these languages.

Regarding the MarsaGram linear (cosine) strategy, 31,339 patterns were extracted in total. Of this amount, 67.67% are present in only one corpus, and only 2 occur in all languages (i.e.: coordinating conjunctions (CCONJ) preceding the heads in subtrees ruled by verbs and nominal appositional modifiers preceding a noun in subtrees ruled by nouns). The dendrogram showed that Maltese and Hungarian were grouped with OV languages and Chinese. Dutch and Greek were clustered with Romance languages (except Romanian) and English. An eclectic large group was formed by Arabic, Bulgarian, Swedish, Irish, Danish, and Romanian. Croatian, Slovenian, and Slovak formed a concise Slavic cluster, however, closer to Latvian (while Lithuanian formed a specific sub-group with Indonesian). The Finnic languages (i.e.: Finnish and Estonian) also formed a sub-cluster of a diverse group formed with the remaining languages.

When MarsaGram all properties method was combined with the head and dependent one, the obtained dendrogram showed an empirical language classification quite coherent with the phylogenetic one. The Romance languages were grouped in the same cluster (except Romanian which formed a distant sub-cluster with Irish and Arabic). The Germanic cluster was also identified (with Greek inside it). Moreover, the Slavic cluster contains most Slavic languages (except for Polish, clustered with Icelandic and Indonesian and closer to Thai and Maltese). The Baltic languages formed a large group together with the Uralic ones. A cluster formed with OV non-related languages was also identified (except Korean which was clustered with Chinese).

The dendrogram obtained using the verb and object relative position features presented a specific OV cluster closer to a group formed by Dutch, German and Hungarian (these three languages are the ones for which the overall analysis of VO and OV tendencies showed balanced results). Dutch is considered in WALS as a language with no dominant order regarding verbs and objects (same as German), however, Hungarian is classed in this database as VO. Regarding the other VO languages, a specific cluster is formed by two Slavic languages (i.e.: Slovak and Slovenian) and the Finnic ones. The other languages form diverse clusters without any connection to the genealogical classification.

Finally, regarding the dendrogram obtained with MarsaGram all properties features (with cosine distances), again we observed that Romanian is classed outside the Romance cluster. The Germanic languages formed a concise group, closer to Maltese which is positioned in between these two clusters. Romanian was grouped with the Slavic languages, in a group that also contains the Baltic ones. The Uralic languages formed a separate cluster together with Turkish.

Besides the typological analyses regarding all EU languages, we also conducted dependency parsing experiments regarding language association for 4 low-resourced EU languages with the best language-pairs identified via the dissimilarity matrices and the obtained dendrograms. To define most low-resourced languages, we analysed the availability of UD corpora and the state-of-the-art parsing results (i.e.: LAS and MLAS presented by the UDify authors, Kondratyuk and Straka, 2019). Four languages were selected: Hungarian, Irish, Lithuanian, and Maltese. We also decided to include Croatian in the list to test how well the typological methods function for a language with more resources although not comparable to the most resourced ones. Maltese is an interesting case as it is not present in multilingual BERT, thus, it was also the opportunity to check if this specificity would jeopardize the results.

The obtained results showed that the best improvements were obtained for Hungarian and Maltese, the languages with the smallest training corpora. For these two languages, both LAS and MLAS presented a statistically significant positive delta. Lithuanian and Croatian presented an improvement regarding MLAS (although for Croatian the p-value was slightly higher than 0.01). On the other hand, Irish did not present any improvement. We also noticed that the best scores were obtained with the language-pairs proposed by MarsaGram all properties and MarsaGram linear strategies (both with cosine distance). Thus, these results corroborate to our previous statement regarding the efficiency of the syntactic information provided by the MarsaGram tool. We could also check that the fact of Maltese being absent from multilingual BERT did not compromise the results as this language presented the best improvements in terms of LAS and MLAS.

Moreover, when we analysed how the language association impacted the LAS in terms of sentence length and distance between heads and dependents, we observed that overall, when languages are associated, LAS tends to be improved especially for longer sentences. Also, when the heads are positioned before the dependents, there is a tendency to have better LAS (or less negative impact) in cases where these elements are relatively close, while when the heads are on the right side of the dependents, the opposite is observed, better LAS are obtained when heads are distant from the dependents.

In conclusion, we can affirm that this thesis presented major contributions in the typological field by presenting new corpus-based methods which bring new light to language comparison regarding syntactic structures and, in the Natural Language Processing field, by detailing a complete series of experiments for a better understanding of the syntactic factors influencing dependency parsing results when languages are combined. The examples displayed regarding low-resourced EU languages showed the potential of the identified strategies for LAS and MLAS results especially for languages with training-sets with a size comparable to Maltese and Hungarian (i.e.: around 1,000 sentences).

In terms of perspectives for future work, the presented corpus-based typological analysis could be applied to larger and more diverse language-sets. Moreover, other deep-learning algorithms with multilingual BERT or other language-models should be tested to check the extend of the dependency parsing improvements obtained with UDify. Also, delexicalized experiments could be useful to check if the presence of words and lemmas have an influence in the overall results.

In this thesis, we focused on the combination of language in pairs, however, it would be interesting to test multilingual corpora with more than 2 languages, beginning with the association of the closest languages identified via the dissimilarity matrices.

Complementary analysis with larger language-sets of overall tendencies regarding the position of heads and dependents and verbs and objects could also be envisaged. A more detailed analysis of the advantage in terms of LAS improvement regarding more complex sentences (i.e.: longer ones) could also be conducted which could have as an outcome a hybrid system using monolingual models for parsing short sentences and multilingual for longer ones.

Moreover, as our focus was on the improvement of LAS and MLAS, we focused on the dependency parsing task, however, part-of-speech and morphosyntactic labels are also affected when corpora are combined. Thus, an extended analysis of the consequences of language associations regarding other CoNNL-U labels could also be useful.

8. References

1. Agić, Ž. (2017, May). Cross-lingual parser selection for low-resource languages. In *Proceedings of the NoDaLiDa 2017 workshop on universal dependencies (UDW 2017)* (pp. 1-10).
2. Agić, Ž., Hovy, D., & Søgaard, A. (2015, July). If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. 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)* (pp. 268-272).
3. Agić, Ž., Johannsen, A., Plank, B., Alonso, H. M., Schluter, N., & Søgaard, A. (2016). Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, *4*, 301-312.
4. Agić, Ž., Tiedemann, J., Merkler, D., Krek, S., Dobrovoljc, K., & Može, S. (2014, October). Cross-lingual dependency parsing of related languages with rich morphosyntactic tagsets. In *Proceedings of the EMNLP'2014 Workshop on Language Technology for Closely Related Languages and Language Variants* (pp. 13-24).
5. Aho, A. V., & Ullman, J. D. (1973). *The theory of parsing, translation, and compiling* (Vol. 1, p. 309). Englewood Cliffs, NJ: Prentice-Hall.
6. Alves, D., Tadić, M., & Bekavac, B. (2022, June). Multilingual Comparative Analysis of Deep-Learning Dependency Parsing Results Using Parallel Corpora. In *Proceedings of the BUCC Workshop within LREC 2022* (pp. 33-42).
7. Alves, D., Bekavac, B., & Tadić, M. (2021, December). Typological Approach to Improve Dependency Parsing for Croatian Language. In Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021) (pp. 1-11).
8. Alexopoulos, E. C. (2010). Introduction to multivariate regression analysis. *Hippokratia*, *14*(Suppl 1), 23.
9. Allassonnière-Tang, M. (2020, August). Optimal parameters for extracting constituent order. In *BOOK OF ABSTRACTS* (p. 440).
10. Ammar, W., Mulcaire, G., Ballesteros, M., Dyer, C., & Smith, N. A. (2016). Many languages, one parser. *Transactions of the Association for Computational Linguistics*, *4*, 431-444.

11. Anderson, J. M., & Grammar, O. C. (1977). Prolegomena to a theory of grammatical relations. *Croom Helm. London*.
12. Aronoff, M. (2007). In the beginning was the word. *Language*, *83*(4), 803-830.
13. Bakker, D., Rendón, J. G., & Hekking, E. (2008). Spanish meets Guaraní, Otomí and Quichua: a multilingual confrontation. *Empirical approaches to language typology*, *35*, 165.
14. Barzilay, R., & Zhang, Y. (2015). Hierarchical low-rank tensors for multilingual transfer parsing. Association for Computational Linguistics.
15. Baumgärtner, K. (1970). Konstituenz und Dependenz. Zur Integration der beiden grammatischen Prinzipien. *Vorschläge für eine strukturelle Grammatik des Deutschen*, 52-77.
16. Baumgärtner, K. (1965). Spracherklärung mit den Mitteln der Abhängigkeitsstruktur. *Beiträge zur Sprachkunde und Informationsverarbeitung*, *5*, 31-53.
17. Bender, E. M. (2016). Linguistic typology in natural language processing. *Linguistic Typology*, *20*(3), 645-660.
18. Bender, E. M. (2009, March). Linguistically naïve!= language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?* (pp. 26-32).
19. Bender, E. M. (2011). On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, *6*.
20. Benedetto, D., Caglioti, E., & Loreto, V. (2002). Language trees and zipping. *Physical Review Letters*, *88*(4).
21. Bentz, C., & Ferrer Cancho, R. (2016). Zipf's law of abbreviation as a language universal. In *Proceedings of the Leiden workshop on capturing phylogenetic algorithms for linguistics* (pp. 1-4). University of Tübingen.
22. Berzak, Y., Reichart, R., & Katz, B. (2016). Contrastive analysis with predictive power: Typology driven estimation of grammatical error distributions in ESL. *arXiv preprint arXiv:1603.07609*.
23. Bharadwaj, A., Mortensen, D. R., Dyer, C., & Carbonell, J. G. (2016, November). Phonologically aware neural model for named entity recognition in low resource transfer settings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1462-1472).

24. Bhat, I. A., Bhat, R. A., Shrivastava, M., & Sharma, D. M. (2017). Joining hands: Exploiting monolingual treebanks for parsing of code-mixing data. *arXiv preprint arXiv:1703.10772*.
25. Bickel, B. (2007). Typology in the 21st century: Major current developments.
26. Björkelund, A., Falenska, A., Yu, X., & Kuhn, J. (2017, August). IMS at the CoNLL 2017 UD shared task: CRFs and perceptrons meet neural networks. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 40-51).
27. Blache, P., Rauzy, S., & Montcheuil, G. (2016, May). MarsaGram: an excursion in the forests of parsing trees. In *Language Resources and Evaluation Conference* (No. 10, p. 7).
28. Bladier, T., Evang, K., Kallmeyer, L., Möllemann, R., & Osswald, R. (2019). Creating RRG treebanks through semi-automatic conversion of annotated corpora.
29. Bonfante, G., Guillaume, B., & Perrier, G. (2018). *Application of Graph Rewriting to Natural Language Processing*. John Wiley & Sons.
30. Bresnan, J., & Mchombo, S. A. (1995). The lexical integrity principle: Evidence from Bantu. *Natural Language & Linguistic Theory, 13*(2), 181-254.
31. Bresnan, J., Asudeh, A., Toivonen, I., & Wechsler, S. (2015). *Lexical-functional syntax*. John Wiley & Sons.
32. Bresnan, J., & Bresnan, J. W. (Eds.). (1982). *The mental representation of grammatical relations* (Vol. 1). MIT press.
33. Buchholz, S., & Marsi, E. (2006, June). CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the tenth conference on computational natural language learning (CoNLL-X)* (pp. 149-164).
34. Bybee, J. L. (1985). Morphology: A study of the relation between meaning and form.
35. Čéplö, S. (2018). Constituent order in Maltese: A quantitative analysis.
36. Chen, D., & Manning, C. D. (2014, October). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 740-750).
37. Choi, J. D., Tetreault, J., & Stent, A. (2015, July). It depends: Dependency parser comparison using a web-based evaluation tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 387-396).

38. Chomboon, K., Chujai, P., Teerarassamee, P., Kerdprasop, K., & Kerdprasop, N. (2015, March). An empirical study of distance metrics for k-nearest neighbor algorithm. In *Proceedings of the 3rd international conference on industrial application engineering* (pp. 280-285).
39. Chomsky, N. (1968). *Remarks on nominalization*. Linguistics Club, Indiana University.
40. Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT press.
41. Chomsky, N. (1976). *Reflections on language*. London: Temple Smith.
42. Chu, Y. J. (1965). On the shortest arborescence of a directed graph. *Scientia Sinica*, *14*, 1396-1400.
43. Cilibrasi, R., & Vitanyi, P. M. B. (2005). Clustering by Compression IEEE Transaction on Information Theory 51.
44. Colas, C., Sigaud, O., & Oudeyer, P. Y. (2018). How many random seeds? statistical power analysis in deep reinforcement learning experiments. *arXiv preprint arXiv:1806.08295*.
45. Collins, C., & Kayne, R. (2009). Syntactic structures of the world's languages. *New York: New York University*.
46. Comrie, B. (2005). Alignment of case marking. In *The world atlas of language structures* (pp. 398-405). Oxford Univ. Press.
47. Comrie, B. (1989). *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.
48. Corbett, G. G. (2012). *Features*. Cambridge University Press.
49. Covington, M. A. (2001). A fundamental algorithm for dependency parsing. In *Proceedings of the 39th annual ACM southeast conference* (Vol. 1).
50. Dahl, Ö. (2004). *The growth and maintenance of linguistic complexity* (Vol. 10). Amsterdam: John Benjamins.
51. Daiber, J., Stanojević, M., & Sima'an, K. (2016, December). Universal reordering via linguistic typology. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 3167-3176).
52. de Lhoneux, M., Bjerva, J., Augenstein, I., & Søgaard, A. (2018). Parameter sharing between dependency parsers for related languages. *arXiv preprint arXiv:1808.09055*.
53. de Lhoneux, M., Shao, Y., Basirat, A., Kiperwasser, E., Stymne, S., Goldberg, Y., & Nivre, J. (2017, August). From Raw Text to Universal Dependencies-Look, No Tags!. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 207-217).

54. De Marneffe, M. C., & Manning, C. D. (2008, August). The Stanford typed dependencies representation. In *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation* (pp. 1-8).
55. de Marneffe, M. C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., & Manning, C. D. (2014). Universal Dependencies: A cross-linguistic typology.
56. De Marneffe, M. C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. *Computational linguistics*, *47*(2), 255-308.
57. Deri, A., & Knight, K. (2016, August). Grapheme-to-phoneme models for (almost) any language. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 399-408).
58. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
59. Dozat, T., & Manning, C. D. (2016). Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.
60. Dryer, M. S., & Haspelmath, M. (2013). The world atlas of language structures online. Leipzig: Max Planck Institute for Evolutionary Anthropology. *Online: http://wals. info*.
61. Dryer, M. S. (1992). The Greenbergian word order correlations. *Language*, *68*(1), 81-138.
62. Dryer, M. (2005). Order of subject, object and verb. *The world atlas of language structures*, 330-333.
63. Eder, M. (2017). Visualization in stylometry: cluster analysis using networks. *Digital Scholarship in the Humanities*, *32*(1), 50-64.
64. Edmonds, J. (1967). Optimum branchings. *Journal of Research of the national Bureau of Standards B*, *71*(4), 233-240.
65. Szmrecsanyi, B. (2016). An information-theoretic approach to assess linguistic complexity. *Complexity, isolation, and variation*, *57*, 71.
66. Evang, K., Bladier, T., Kallmeyer, L., & Petitjean, S. (2021, December). Bootstrapping role and reference grammar treebanks via Universal Dependencies. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)* (pp. 30-48).
67. Fang, M., & Cohn, T. (2017). Model transfer for tagging low-resource languages using a bilingual dictionary. *arXiv preprint arXiv:1705.00424*.
68. Ferrer-i-Cancho, R. F. (2006). Why do syntactic links not cross?. *EPL (Europhysics Letters)*, *76*(6), 1228.

69. Fillmore, C. J. (1967). The case for case.
70. Fisch, A., Guo, J., & Barzilay, R. (2019). Working hard or hardly working: Challenges of integrating typology into neural dependency parsers. *arXiv preprint arXiv:1909.09279*.
71. Fitialov, S. J. (1962). O modelirovanii sintaksisa v strukturnoj lingvistike. *Problemy strukturnoj lingvistiki, Moskva*, 100-114.
72. Futrell, R., Mahowald, K., & Gibson, E. (2015, August). Quantifying word order freedom in dependency corpora. In *Proceedings of the third international conference on dependency linguistics (Depling 2015)* (pp. 91-100).
73. Gaifman, H. (1965). Dependency systems and phrase-structure systems. *Information and control*, *8*(3), 304-337.
74. Ganchev, K., & Das, D. (2013). Cross-lingual discriminative learning of sequence models with posterior regularization.
75. Georgi, R., Xia, F., & Lewis, W. (2010, August). Comparing language similarity across genetic and typologically-based groupings. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)* (pp. 385-393).
76. Gerdes, K., Guillaume, B., Kahane, S., & Perrier, G. (2019a, August). Improving Surface-syntactic Universal Dependencies (SUD): surface-syntactic relations and deep syntactic features. In *TLT 2019-18th* International Workshop on Treebanks and Linguistic Theories.
77. Gerdes, K., Kahane, S., & Chen, X. (2019, August). Rediscovering Greenberg's Word Order Universals in UD. In UDW, Universal Dependencies Workshop 2019, Syntaxfest.
78. Gerdes, K., Kahane, S., & Chen, X. (2021). Typometrics: From implicational to quantitative universals in word order typology. Glossa: a journal of general linguistics, 6(1).
79. Gil, D. (2005). Numeral classifiers. In *The world atlas of language structures* (pp. 226-229). Oxford Univ. Press.
80. Glavaš, G., & Vulić, I. (2021, August). Climbing the tower of treebanks: Improving low-resource dependency parsing via hierarchical source selection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 4878-4888).
81. Greenberg, J. H. (1960). A quantitative approach to the morphological typology of language. *International journal of American linguistics*, *26*(3), 178-194.

82. Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language*, *2*, 73-113.
83. Greenberg, J. H. (1969). Some methods of dynamic comparison in linguistics. *Substance and structure of language*, 147-203.
84. Greenberg, J. H. (1995). The Diachronic typological approach. *Approaches to language typology*, 145-66.
85. Naranjo, M. G., & Becker, L. (2018, December). Quantitative word order typology with UD. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018), December 13–14, 2018, Oslo University, Norway* (No. 155, pp. 91-104). Linköping University Electronic Press.
86. Hagège, C. (1988). Contribution des recherches typologiques à. l'étude diachronique des langues. *A. Joly (éd.) La linguistique génétique. Histoire et Théories*.
87. Haider, H., Fábregas, A., Mateu, J., & Putnam, M. (2015). Head directionality. *The handbook of parameters*, 73-97.
88. Haig, G., Schnell, S., & Wegener, C. (2011). Comparing corpora from endangered language projects: Explorations in language typology based on original texts. *Documenting endangered languages*, 55-86.
89. Hajič, J. (1998). Building a syntactically annotated corpus: The prague dependency treebank. *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová*, 106-132.
90. Hajic, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M. A., Màrquez, L., ... & Zhang, Y. (2009, June). The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task* (pp. 1-18).
91. Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2015). Glottolog 2.6. Jena, Germany: Max Planck Institute for the Science of Human History.
92. Haspelmath, M., Dryer, M. S., Gil, D., & Comrie, B. (2005). *The world atlas of language structures*. OUP Oxford.
93. Hawkins, J. A. (1983). *Word order universals*. Academic Press.
94. Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. OUP Oxford.
95. Hays, D. G. (1964). Dependency theory: A formalism and some observations. *Language*, *40*(4), 511-525.
96. Hays, D. G. (1960). Grouping and dependency theories. In *Proceedings of the National Symposium on Machine Translation*.

97. Heringer, H. J. (1970). Einige Ergebnisse und Probleme der Dependenzgrammatik. *Der Deutschunterricht*, *22*(4), 42-98.

98. Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

99. Hudson, R. (1984). *Word Grammar*. Oxford: Oxford University Press.

100. Hudson, R. (1990). *English Word Grammar*. Oxford: Basil Blackwell.

101. Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., & Kolak, O. (2005). Bootstrapping parsers via syntactic projection across parallel texts. *Natural language engineering*, *11*(3), 311-325.

102. Lidija, I. (1963). O nekotoryx svojstvax pravil'noj sintaksiceskoj struktury (na materiale russkogo jazyka). *On some properties of the correct syntactic structure (on the basis of Russian)], Voprosy jazykoznanija*, *4*, 102-12.

103. Nuessel Jr, F. H. (1979). X̄ syntax: a study of phrase structure. Linguistic Inquiry Monograph Two: Ray Jackendoff, MIT Press, Cambridge, Mass., 1977. xii, 249 pp. *Lingua*, *49*(2-3), 255-259.

104. Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., ... & Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, *5*, 339-351.

105. Juola, P. (1998). Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics*, *5*(3), 206-213.

106. Juola, P. (2008). Assessing linguistic complexity. *Language complexity: Typology, contact, change*, 89-108.

107. Jurafsky, D., Martin, J. H. (2021). *Speech and Language Processing*, 3rd edition (draft).

108. Keenan, E. L. (1974). The Functional Principle: Generalizing the Notion of Subject-of. In *Papers from the 10th Regional Meeting of the Chicago Linguistic Society*. Chicago: Chicago Linguistic Society.

109. Keenan, E. L., & Comrie, B. (1977). Noun phrase accessibility and universal grammar. *Linguistic inquiry*, *8*(1), 63-99.

110. Keenan, E. L. (1978). *Language variation and the logical structure of universal grammar*.

111. Keenan, E. L. (1979). On surface form and logical form. *Studies in the Linguistic Sciences Urbana, Ill*, *8*(2), 163-203.

112. Khapra, M. M., Joshi, S., Chatterjee, A., & Bhattacharyya, P. (2011, June). Together we can: Bilingual bootstrapping for WSD. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 561-569).

113. Kiperwasser, E., & Goldberg, Y. (2016). Easy-first dependency parsing with hierarchical tree LSTMs. *Transactions of the Association for Computational Linguistics*, *4*, 445-461.

114. Kiperwasser, E., & Goldberg, Y. (2016). Simple and accurate dependency parsing using bidirectional LSTM feature representations. *Transactions of the Association for Computational Linguistics*, *4*, 313-327.

115. Kondratyuk, D., & Straka, M. (2019). 75 languages, 1 model: Parsing universal dependencies universally. *arXiv preprint arXiv:1904.02099*.

116. Kudo, T., & Matsumoto, Y. (2002). Japanese dependency analysis using cascaded chunking. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

117. Kulmizev, A., de Lhoneux, M., Gontrum, J., Fano, E., & Nivre, J. (2019). Deep Contextualized Word Embeddings in Transition-Based and Graph-Based Dependency Parsing--A Tale of Two Parsers Revisited. *arXiv preprint arXiv:1908.07397*.

118. Kunze, J. (1972). Die Komponenten der Darstellung syntaktischer Strukturen in einer Abhängigkeitsgrammatik. *Prague Bulletin of Mathematical Linguistics*, *18*, 5-27.

119. Lecerf, Y. (1960). Programme des conflits - modèle des conflits. Rapport CETIS., N.4, Euratom. p. 1-24.

120. Lehman, A. (2005). *JMP for basic univariate and multivariate statistics: a step-by-step guide*. SAS Institute.

121. Lehmann, W. P. (1973). A structural principle of language and its implications. *Language*, 47-66.

122. Lehmann, W. P. (1974). *Proto-indo-european syntax*. University of Texas Press.

123. Lehmann, W. P. (Ed.). (1978). *Syntactic typology: Studies in the phenomenology of language* (Vol. 10). University of Texas Press.

124. Lemay, D. L'ordre des mots en slovaque expliqué aux apprenants francophones The Word Order in Slovak Language Explained to French Speaking Students. *Linguae. eu*, 9.

125. Levshina, N. (2019). Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology*, *23*(3), 533-572.

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

127. Lewis, M. P., Simons, G. F., & Fennig, C. D. (2015). Ethnologue: languages of Ecuador. *SIL International, Dallas*.

128. Li, M., Chen, X., Li, X., Ma, B., & Vitányi, P. M. (2004). The similarity metric. *IEEE transactions on Information Theory*, *50*(12), 3250-3264.

129. Li, S., Graça, J., & Taskar, B. (2012, July). Wiki-ly supervised part-of-speech tagging. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 1389-1398).

130. Lightfoot, D. W. (1979). Principles of diachronic syntax. *Cambridge Studies in Linguistics London*, *23*.

131. Litschko, R., Vulić, I., Agić, Ž., & Glavaš, G. (2020). Towards instance-level parser selection for cross-lingual transfer of dependency parsers. *arXiv preprint arXiv:2004.07642*.

132. Littell, P., Mortensen, D. R., Lin, K., Kairis, K., Turner, C., & Levin, L. (2017, April). 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* (pp. 8-14).

133. Liu, Z. (2020). Mixed evidence for crosslinguistic dependency length minimization. *STUF-Language Typology and Universals*, *73*(4), 605-633.

134. Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, *9*(2), 159-191.

135. Liu, H., & Xu, C. (2012). Quantitative typological analysis of Romance languages. *Poznań Studies in Contemporary Linguistics*, *48*(4), 597-625.

136. Liu, H. (2010). Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, *120*(6), 1567-1578.

137. Lui, M., & Baldwin, T. (2012, July). langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations* (pp. 25-30).

138. Lynn, T., Foster, J., Dras, M., & Tounsi, L. (2014, August). Cross-lingual transfer parsing for low-resourced languages: An Irish case study. In *Proceedings of the First Celtic Language Technology Workshop* (pp. 41-49).

139. Malaviya, C., Neubig, G., & Littell, P. (2017). Learning language representations for typology prediction. *arXiv preprint arXiv:1707.09569*.

140. Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., ... & Schasberger, B. (1994). The Penn treebank: Annotating predicate argument structure. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11*.

141. Marcus, S. (1965). Sur la notion de projectivité. *Mathematical Logic Quarterly*, *11*(2), 181-192.

142. Mayer, T., & Cysouw, M. (2012, April). Language comparison through sparse multilingual word alignment. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH* (pp. 54-62).

143. McDonald, R., & Nivre, J. (2011). Analyzing and integrating dependency parsers. *Computational Linguistics, 37*(1), 197-230.

144. McDonald, R., Pereira, F., Ribarov, K., & Hajič, J. (2005, October). Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing* (pp. 523-530).

145. McDonald, R., Petrov, S., & Hall, K. (2011, July). Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 62-72).

146. McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, *2*(11), 205.

147. McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

148. Mel’čuk, I. (1974). Opyt teorii lingvističeskix modelej “Smysl⇔ Tekst”. Semantika, Sintaksis.[Outline of a Theory of “Meaning-Text” Type Linguistic Models. Semantics, Syntax].

149. Melʹčuk, I. A. (1979). *Studies in dependency syntax* (Vol. 2). Karoma Pub.

150. Mel'čuk, I. A. (1988). *Dependency syntax: theory and practice*. SUNY press.

151. Mel’čuk, I. (1997). Vers une linguistique Sens-Texte. Leçon inaugurale. Paris:Collège de France.

152. Mel’čuk, I. (2009). Dependency in natural language. *Dependency in linguistic description*, *111*, 1.

153. Moran, S., McCloy, D., & Wright, R. (2014). PHOIBLE online.

154. Moravcsik, E. A. (2012). *Introducing language typology*. Cambridge University Press.

155. Naseem, T., Barzilay, R., & Globerson, A. (2012). Selective sharing for multilingual dependency parsing. The Association for Computational Linguistics.

156. Nettle, D., & Romaine, S. (2000). *Vanishing voices: The extinction of the world's languages*. Oxford University Press on Demand.

157. Nivre, J., Agić, Ž., Aranzabe, M. J., Asahara, M., Atutxa, A., Ballesteros, M., ... & Zhu, H. (2015). Universal Dependencies 1.2.

158. Nivre, J. (2003, April). An efficient algorithm for projective dependency parsing. In *Proceedings of the eighth international conference on parsing technologies* (pp. 149-160).

159. Nivre, J., Hall, J., & Nilsson, J. (2006, May). Maltparser: A data-driven parser-generator for dependency parsing. In *LREC* (Vol. 6, pp. 2216-2219).

160. Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., & Yuret, D. (2007, June). 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)* (pp. 915-932).

161. Nivre, J. (2009, August). Non-projective dependency parsing in expected linear time. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (pp. 351-359).

162. Norman, J. (2009). A New Look at Altaic [Review of Etymological Dictionary of the Altaic Languages. 3 volumes. Handbook of Oriental Studies: Section 8, Uralic and Central Asian Studies; no. 8, by S. Starostin, A. Dybo, O. Mudrak, I. Gruntov, & V. Glumov]. Journal of the American Oriental Society, 129(1), 83–89.

163. Nuzzo, R. (2014). Statistical errors. *Nature*, *506*(7487), 150.

164. O'Horan, H., Berzak, Y., Vulić, I., Reichart, R., & Korhonen, A. (2016). Survey on the use of typological information in natural language processing. *arXiv preprint arXiv:1610.03349*.

165. Östling, R. (2015, July). 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)* (pp. 205-211).

166. Otter, D. W., Medina, J. R., & Kalita, J. K. (2019). A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, *32*(2), 604-624.

167. Padó, S., & Lapata, M. (2009). Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research, 36*, 307-340.

168. Padučeva, E. V. (1964). O sposobax predstavlenija sintaksičeskoj struktury predloženija. *Voprosy jazykoznanija*, *13*(2), 99-113.

169. Pappas, N., & Popescu-Belis, A. (2017). Multilingual hierarchical attention networks for document classification. *arXiv preprint arXiv:1707.00896*.

170. Perlmutter, D. M., & Rosen, C. G. (Eds.). (1983). *Studies in relational grammar 1* (Vol. 1). University of Chicago Press.

171. Petkevic, V. (1995). A new formal specification of underlying structures. *Theoretical linguistics*, *21*(1), 7-61.

172. Petrov, S., Das, D., & McDonald, R. (2011). A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

173. Pierce, J. R., & Carroll, J. B. (1966). Language and machines: Computers in translation and linguistics.

174. Pierrel, J. M. (2014). Ortolang. Une infrastructure de mutualisation de ressources linguistiques écrites et orales. *Recherches en didactique des langues et des cultures. Les cahiers de l'Acedle*, *11*(11-1).

175. Plank, F., & Filimonova, E. (2000). The Universals Archive: A Brief Introduction for Prospective Users. *STUF-Language Typology and Universals*, *53*(1), 109-123.

176. Polguère, A., Mel'čuk, I. (2009). *Dependency in Natural Language*.

177. Ponti, E. M., O'horan, H., Berzak, Y., Vulić, I., Reichart, R., Poibeau, T., ... & Korhonen, A. (2019). Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, *45*(3), 559-601.

178. Ponti, E., Reichart, R., Korhonen, A. L., & Vulic, I. (2018, July). Isomorphic transfer of syntactic structures in cross-lingual NLP. Association for Computational Linguistics.

179. Przepiórkowski, A., & Patejuk, A. (2020). From lexical functional grammar to enhanced universal dependencies. *Language Resources and Evaluation*, *54*(1), 185-221.

180. Rehm, G. et al. (2012). *META-NET White Paper Series*, available at http://www.meta-net.eu/whitepapers/overview.

181. Rijkhoff, J. (2002). Verbs and nouns from a cross-linguistic perspective. *Rivista di linguistica*, *14*(1), 115-147.

182. Robinson, J. J. (1970). Dependency structures and transformational rules. *Language*, 259-285.

183. Rosa, R., & Žabokrtský, Z. (2015, July). Klcpos3-a language similarity measure for delexicalized parser transfer. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 243-249).

184. Rosén, V., Dyvik, H., Meurer, P., & De Smedt, K. (2020). Creating and exploring LFG treebanks.

185. Ruszkowski, M. (2003). Behaghel's Law. *Polonica*, (22-23), 117-121.

186. Santorini, B. (1990). Part-of-speech tagging guidelines for the Penn Treebank Project.

187. Sapir, E. (1921). An introduction to the study of speech. *Language*, *1*.

188. Say, S. (2014). Bivalent verb classes in the languages of Europe: A quantitative typological study. *Language dynamics and change*, *4*(1), 116-166.

189. Schlegel, F. V. On the Language and Wisdom of the Indians [1808]. *The Aesthetic and Miscellaneous Works of Friedrich von Schlegel, trans. EJ Millington (London: Bohn, 1849)*.

190. Schluter, N., & Agić, Ž. (2017). Empirically sampling universal dependencies. *NEALT (Northern European Association of Language Technology) Proceedings Series*, *31*, 117-122.

191. Scholivet, M., Dary, F., Nasr, A., Favre, B., & Ramisch, C. (2019, June). Typological features for multilingual delexicalised dependency parsing. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3919-3930). Association for Computational Linguistics.

192. Schone, P., & Jurafsky, D. (2001). Language-independent induction of part of speech class labels using only language universals. *Machine Learning: Beyond Supervision*.

193. Sgall, P., Hajicová, E., Hajicová, E., Panevová, J., & Panevova, J. (1986). *The meaning of the sentence in its semantic and pragmatic aspects*. Springer Science & Business Media.

194. Shibatani, M. (2015). Linguistic Typology. In: James D. Wright (editor-in-chief), *International Encyclopedia of the Social & Behavioral Sciences*, 2nd edition, Vol14. Oxford: Elsevier. pp. 208–214.

195. Siewierska, A., & Bakker, D. (1996). The distribution of subject and object agreement and word order type. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, *20*(1), 115-161.

196. Sinnemäki, K. (2008). Complexity trade-offs in core argument marking. *Language complexity: Typology, contact, change*, *67,* 88.

197. Skalička, V. (1979). A "typological construct". In *Typological Studies* (pp. 335-341). Vieweg + Teubner Verlag.

198. Spencer, N. H. (2013*). Essentials of multivariate data analysis*. CRC press.

199. Starosta, S. (1988). *The Case for Lexicase*. London: Pinter.

200. Straka, M., Hajič, J., & Straková, J. (2016, May). UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 4290-4297).

201. Straka, M. (2018, October). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 197-207).

202. Stymne, S., de Lhoneux, M., Smith, A., & Nivre, J. (2018). Parser training with heterogeneous treebanks. *arXiv preprint arXiv:1805.05089*.

203. Szmrecsanyi, B. (2009). Typological parameters of intralingual variability: Grammatical analyticity versus syntheticity in varieties of English. *Language Variation and Change*, *21*(3), 319-353.

204. Täckström, O., McDonald, R., & Nivre, J. (2013, June). Target language adaptation of discriminative transfer parsers. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1061-1071).

205. Takamura, H., Nagata, R., & Kawasaki, Y. (2016, May). Discriminative analysis of linguistic features for typological study. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 69-76).

206. Teh, Y., Daume III, H., & Roy, D. M. (2007). Bayesian agglomerative clustering with coalescents. *Advances in neural information processing systems, 20.*

207. Tesnière, L. (1959). *Eléments de syntaxe structurale*. Paris.

208. Tsvetkov, Y., Sitaram, S., Faruqui, M., Lample, G., Littell, P., Mortensen, D., ... & Dyer, C. (2016). Polyglot neural language models: A case study in cross-lingual phonetic representation learning. *arXiv preprint arXiv:1605.03832*.

209. Tutorials, S. P. S. S. (2022). Pearson correlation. Retrieved on October, 4th, 2022 at https://libguides.library.kent.edu/SPSS/PearsonCorr.

210. Urrutia, A. T. (2018). A Proposal to Describe Fuzziness in Natural. *Logic and Algorithms in Computational Linguistics 2018 (LACompLing2018)*, 87.

211. Urrutia, A. T. (2017). Extracción de propiedades para una sintaxis del español. In *Recerca en humanitats 2017* (pp. 191-201). Publicacions URV.

212. Üstün, A., Bisazza, A., Bouma, G., & van Noord, G. (2020). UDapter: Language adaptation for truly Universal Dependency parsing. *arXiv preprint arXiv:2004.14327.*

213. Vennemann, T. (1972). Analogy in generative grammar: the origin of word order. In *Proceedings of the eleventh international congress of linguists* (Vol. 2, pp. 79-83).

214. Vennemann, T. (1973). Explanation in syntax. In J. Kimball (Ed.), *Syntax and Semantics (Vol. 2)*. New York: Academic Press.

215. Vennemann, T. l974. Topics, Subjects, and Word Order: From SXV to SVX Via TVX. *Anderson and Jones (eds.), Historical_Linguistics L., Amsterdam: North Holland Publishing Co*, 559-76.

216. Vennemann, T. (1976). Categorial grammar and the order of meaningful elements. *Linguistic studies offered to Joseph Greenberg on the occasion of his sixtieth birthday, 3*, 615-634.

217. Vennemann, T. (1981). Typology, Universals and Change of Language. *International Conference on Historical Syntax*, Poznan.

218. Wackernagel, J., “Über ein Gesetz der indo-germanischen Wortstellung”. Indogermanische Forschungen 1. 333-436, 1892.

219. Wagner, R. A., & Fischer, M. J. (1974). The string-to-string correction problem. *Journal of the ACM (JACM)*, *21*(1), 168-173.

220. Wälchli, B., & Cysouw, M. (2012). Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics*, *50*(3), 671-710.

221. Wälchli, B. (2009). Data reduction typology and the bimodal distribution bias.

222. Wang, D., & Eisner, J. (2018). Synthetic data made to order: The case of parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 1325-1337).

223. Wang, D., & Eisner, J. (2016). The galactic dependencies treebanks: Getting more data by synthesizing new languages. *Transactions of the Association for Computational Linguistics*, *4*, 491-505.

224. Wang, M., & Manning, C. D. (2014). Cross-lingual projected expectation regularization for weakly supervised learning. *Transactions of the Association for Computational Linguistics*, *2*, 55-66.

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

226. Wisniewski, G., Pécheux, N., Gahbiche-Braham, S., & Yvon, F. (2014, October). Cross-lingual part-of-speech tagging through ambiguous learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1779-1785).

227. Wu, S., & Dredze, M. (2020). Are all languages created equal in multilingual BERT?. *arXiv preprint arXiv:2005.09093*.

228. Xia, F., & Palmer, M. (2001). *Converting dependency structures to phrase structures*. Pennsylvania Univ Philadelphia.

229. Yamada, H., & Matsumoto, Y. (2003, April). Statistical dependency analysis with support vector machines. In *Proceedings of the eighth international conference on parsing technologies* (pp. 195-206).

230. Yarowsky, D., Grace, N., Richard W. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8.

231. Zeman, D., & Resnik, P. (2008). Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

232. Zeman, D., HajiČ, J., Popel, M., Potthast, M., Straka, M., Ginter, F., ... & Petrov, S. (2018, October). 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* (pp. 1-21).

233. Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gökırmak, M., Nedoluzhko, A., Cinková, S., Hajič jr., J., Hlaváčová, J., Kettnerová, V., Urešová, Z. et al. (2017). CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

234. Zeman, D. (2008, May). Reusable Tagset Conversion Using Tagset Drivers. In *LREC* (Vol. 2008, pp. 28-30).

235. Zhang, Y., Reichart, R., Barzilay, R., & Globerson, A. (2012, July). Learning to map into a universal pos tagset. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 1368-1378).

236. Zipf, G., *The psychobiology of Language: An introduction to dynamic philology*. Cambridge, MA: MIT Press; 1965 [1935].

Annex 1.

UPOS tag	Part-of-speech	Type
ADJ	Adjective	Open class
ADP	Adposition (preposition/postposition)	Closed class
ADV	Adverb	Open class
AUX	Auxiliary verb or other tense, aspect, or mood particle	Closed class
CCONJ	Coordinating conjunction	Closed class
DET	Determiner (including article)	Closed class
INTJ	Interjection	Open class
NOUN	Common Noun	Open class
NUM	Numeral (cardinal)	Closed class
PART	Particle (special single word markers in some languages)	Closed class
PRON	Pronoun	Closed class
PROPN	Proper noun	Open class
PUNCT	Punctuation	Other
SCONJ	Subordinating conjunction	Closed class
SYM	Non-punctuation symbol (e.g.: # or emojis)	Other
VERB	Main Verb	Open class
X	Other (e.g.: foreign words)	Other

Table A.1. List of labels composing the universal part-of-speech tag-set of Universal Dependencies framework.

Annex 2.

FEAT tag	Feature	Type	Sub-type	Values
Abbr	Abbreviation	Lexical	-	"Yes"
Animacy	Animacy	Inflectional	Nominal	"Animate", "Inanimate", "Inanimate", "Non-human"
Aspect	Aspect	Inflectional	Verbal	"Habitual", "Imperfect", "Iterative/frequentative", "Perfect", "Progressive", "Prospective"
Case	Case	Inflectional	Nominal	"Absolutive", "Accusative", "Ergative", "Nominative", "Abessive/caritative/privative", "Benefactive/destinative", "Causative/motivative/purposive", "Comparative", "Considerative", "Comitative/associative", "Dative", "Distributive", "Equative", "Genitive", "Instrumental/instructive", "Partitive", "Temporal", "Translative/factive", "Vocative", "Ablative/adelative", "Additive", "Adessive", "Allative", "Delative/superelative", "Elative/inelative", "Essive/prolative", "Illative/inlative", "Inessive", "Lative/directional allative", "Locative", "Perlative", "Subessive", "Sublative", "Superlative", "Subelative", "Supressive", "Terminative/terminal allative"

Table A.2. Part I of the list of labels and possible values composing the universal features tag-set of Universal Dependencies framework.

Annex 3.

FEAT tag	Feature	Type	Sub-type	Values
Clusivity	Clusivity	Inflectional	Verbal	"Exclusive", "Inclusive"
Definite	Definite	Inflectional	Nominal	"Indefinite", "Specific indefinite", "Definite", "Construct state/reduced definiteness", "Complex"
Degree	Degree	Inflectional	Nominal	"Positive/first degree", "Equative", "Comparative/second degree", "Superlative/third degree", "Absolute superlative"
Evident	Evidentiality	Inflectional	Verbal	"Firsthand", "Non-firsthand"
Foreign	Foreign	Lexical	-	"Yes"
Gender	Gender	Inflectional	Nominal	"Masculine", "Feminine", "Neuter", "Common"
Mood	Mood	Inflectional	Verbal	"Indicative", "Imperative", "Conditional", "Potential", "Subjunctive/conjunctive", "Jussive/injunctive", "Purposive", "Quotative", "Optative", "Desiderative", "Necessitative", "Irrealis", "Admirative"
NounClass	Noun class	Inflectional	Nominal	"Bantu 1 to 23", "Wol 1 to 12"
Number	Number	Inflectional	Nominal	"Singular", "Plural", "Dual", "Trial", "Paucal", "Greater paucal", "Inverse number", "Count plural", "Plurale tantum", "Collective/mass/singulare tantum"
NumType	Numeral type	Lexical	-	"Cardinal", "Ordinal", "Multiplicative", "Fraction", "Sets", "Distributive", "Range"

Table A.3. Part II of the list of labels and possible values composing the universal features tag-set of Universal Dependencies framework.

Annex 4.

FEAT tag	**Feature**	**Type**	**Sub-type**	**Values**
Person	Person	Inflectional	Verbal	"0", "1", "2", "3", "4"
Polarity	Polarity	Inflectional	Verbal	"Positive", "Negative"
Polite	Polite	Inflectional	Verbal	"Informal", "Formal", "Referent elevating", "Speaker humbling"
Poss	Possessive	Lexical	-	"Yes"
PronType	Pronominal type	Lexical	-	"Personal or possessive personal pronoun or determiner", "Reciprocal", "Article", "Interrogative pronoun, determiner, numeral or adverb", "Relative pronoun, determiner, numeral or adverb", "Exclamative determiner", "Demonstrative pronoun, determiner, numeral or adverb", "Emphatic determiner", "Total (collective) pronoun, determiner or adverb, "Negative pronoun, determiner or adverb", "Indefinite pronoun, determiner, numeral or adverb"
Reflex	Reflexive	Lexical	-	"Yes"
Tense	Tense	Inflectional	Verbal	"Past/preterite/aorist", "Present/non-past tense/aorist", "Future", "Pluperfect"
Typo	Misspelled word	Lexical	-	"Yes"
VerbForm	Form of verb or deverbative	Inflectional	Verbal	"Finite", "Infinitive", "Supine", "Participle", "Converb/transgressive/adverbial participle", "Gerundive", "Gerund", "Verbal noun"
Voice	Voice	Inflectional	Verbal	"Active", "Middle voice", "Reciprocal", "Passive", "Antipassive", "Location-focus", "Beneficiary-focus", "Direct", "Inverse", "Causative",

Table A.4. Part III of the list of labels and possible values composing the universal features tag-set of Universal Dependencies framework.

Annex 5.

DEPREL tag	Dependency Relation	Relation to the head	Structural category
acl	clausal modifier of noun (adnominal clause)	Nominal dependent	Clause
acl:relcl	relative clause modifier	Nominal dependent	Clause
advcl	adverbial clause modifier	Non-core dependent	Clause
advmod	adverbial modifier	Non-core dependent	Modifier
advmod:emph	emphasizing word, intensifier	Non-core dependent	Modifier
advmod:lmod	locative adverbial modifier	Non-core dependent	Modifier
amod	adjectival modifier	Nominal dependent	Modifier
appos	appositional modifier	Nominal dependent	Nominal
aux	auxiliary	Non-core dependent	Function
aux:pass	passive auxiliary	Non-core dependent	Function
case	case marking	Nominal dependent	Function
cc	coordinating conjunction	-	Coordination
cc:preconj	Preconjunct	-	Coordination
ccomp	clausal complement	Core argument	Clause
clf	classifier	Nominal dependent	Function
compound	compound	-	Multiword expression
compound:lvc	light verb construction	-	Multiword expression
compound:prt	phrasal verb particle	-	Multiword expression
compound:redup	reduplicated compounds	-	Multiword expression
compound:svc	serial verb compounds	-	Multiword expression
conj	conjunct	-	Coordination
cop	copula	Non-core dependent	Function

Table A.5. Part I of the list of DEPREL labels and classification of the Universal Dependencies tag-set.

Annex 6.

DEPREL tag	Dependency Relation	Relation to the head	Structural category
csubj	clausal subject	Core argument	Clause
csubj:pass	clausal passive subject	Core argument	Clause
dep	unspecified dependency	-	Other
det	determiner	Nominal dependent	Function
det:numgov	pronominal quantifier governing the case of the noun	Nominal dependent	Function
det:nummod	pronominal quantifier agreeing in case with the noun	Nominal dependent	Function
det:poss	possessive determiner	Nominal dependent	Function
discourse	discourse element	Non-core dependent	Modifier
dislocated	dislocated elements	Non-core dependent	Nominal
expl	expletive	Non-core dependent	Nominal
expl:impers	impersonal expletive	Non-core dependent	Nominal
expl:pass	reflexive pronoun used in reflexive passive	Non-core dependent	Nominal
expl:pv	reflexive clitic with an inherently reflexive verb	Non-core dependent	Nominal
fixed	fixed multiword expression	-	Multiword expression
flat	flat multiword expression	-	Multiword expression
flat:foreign	foreign words	-	Multiword expression
flat:name	names	-	Multiword expression
goeswith	goes with	-	Special
iobj	indirect object	Core argument	Nominal
list	list	-	Loose
mark	marker	Non-core dependent	Function

Table A.6. Part II of the list of DEPREL labels and classification of the Universal Dependencies tag-set.

Annex 7.

DEPREL tag	Dependency Relation	Relation to the head	Structural category
nmod	nominal modifier	Nominal dependent	Nominal
nmod:poss	possessive nominal modifier	Nominal dependent	Nominal
nmod:tmod	temporal modifier	Nominal dependent	Nominal
nsubj	nominal subject	Core argument	Nominal
nsubj:pass	passive nominal subject	Core argument	Nominal
nummod	numeric modifier	Nominal dependent	Nominal
nummod:gov	numeric modifier governing the case of the noun	Nominal dependent	Nominal
obj	object	Core argument	Nominal
obl	oblique nominal	Non-core dependent	Nominal
obl:agent	agent modifier	Non-core dependent	Nominal
obl:arg	oblique argument	Non-core dependent	Nominal
obl:lmod	locative modifier	Non-core dependent	Nominal
obl:tmod	temporal modifier	Non-core dependent	Nominal
orphan	orphan	-	Special
parataxis	parataxis	-	Loose
punct	punctuation	-	Other
reparandum	overridden disfluency	-	Other
root	root	-	Other
vocative	vocative	Non-core dependent	Nominal
xcomp	open clausal complement	Core argument	Clause

Table A.7. Part III of the list of DEPREL labels and classification of the Universal Dependencies tag-set.

Annex 8.

Language	Number of Features	Descriptive Genealogical Features (lang2vec)
arb	6	F_Afro-Asiatic, F_Semitic, F_West_Semitic, F_Central_Semitic, F_Arabian, F_Arabic
cmn	4	F_Sino-Tibetan, F_Sinitic, F_Northern_Chinese, F_Mandarinic
ces	6	F_Indo-European, F_Balto-Slavic, F_Slavic, F_West_Slavic, F_Czech-Slovak, F_Czech-Lach
eng	9	F_Indo-European, F_Germanic, F_Northwest_Germanic, F_West_Germanic, F_North_Sea_Germanic, F_Anglo-Frisian, F_Anglian, F_Mercian, F_Macro-English
fin	3	F_Uralic, F_Finnic, F_Nuclear_Finnish
fra	12	F_Indo-European, F_Italic, F_Latino-Faliscan, F_Latinic, F_Imperial_Latin, F_Romance, F_Italo-Western_Romance, F_Western_Romance, F_Shifted_Western_Romance, F_Northwestern_Shifted_Romance, F_Gallo-Rhaetian, F_Oil
deu	6	F_Indo-European, F_Germanic, F_Northwest_Germanic, F_West_Germanic, F_Franconian
hin	7	F_Indo-European, F_Indo-Iranian, F_Indo-Aryan, F_Indo-Aryan_Central_zone, F_Subcontinental_Central_Indo-Aryan, F_Western_Hindi, F_Hindustani
isl	6	F_Indo-European, F_Germanic, F_Northwest_Germanic, F_North_Germanic, F_West_Scandinavian, F_Icelandic-Faroese
ind	8	F_Austronesian, F_Nuclear_Austronesian, F_Malayo-Polynesian, F_Malayo-Sumbawan, F_North_and_East_Malayo-Sumbawan, F_Malayic, F_Nuclear_Malayic, F_Indonesian_Archipelago_Malay
ita	9	F_Indo-European, F_Italic, F_Latino-Faliscan, F_Latinic, F_Imperial_Latin, F_Romance, F_Italo-Western_Romance, F_Italo-Dalmatian, F_Italian_Romance
jpn	3	F_Japonic, F_Japanese, F_Japan-Taiwan_Japanese

Table A.8. Part I of the list of genealogical features of PUD languages (with value equal to 1.0 in lang2vec tool).

Annex 9.

Language	Number of Features	Descriptive Genealogical Features (lang2vec)
kor	1	F_Koreanic
pol	5	F_Indo-European, F_Balto-Slavic, F_Slavic, F_West_Slavic, F_Lechitic
por	11	F_Indo-European, F_Italic, F_Latino-Faliscan, F_Latinic, F_Imperial_Latin, F_Romance, F_Italo-Western_Romance, F_Western_Romance, F_Shifted_Western_Romance, F_Southwestern_Shifted_Romance, F_West_Ibero-Romance
rus	4	F_Indo-European, F_Balto-Slavic, F_Slavic, F_East_Slavic
spa	12	F_Indo-European, F_Italic, F_Latino-Faliscan, F_Latinic, F_Imperial_Latin, F_Romance, F_Italo-Western_Romance, F_Western_Romance, F_Shifted_Western_Romance, F_Southwestern_Shifted_Romance, F_West_Ibero-Romance, F_Castilic
swe	6	F_Indo-European, F_Germanic, F_Northwest_Germanic, F_North_Germanic, F_East_Scandinavian, F_Macro-Swedish
tha	10	F_Tai-Kadai, F_Kam-Tai, F_Be-Tai, F_Daic, F_Central-Southwestern_Tai, F_Wenma-Southwestern_Tai, F_Sapa-Southwestern_Tai, F_Southwestern_Tai, F_Sputhwestern_Thai_PH, F_Lao-Thai
tur	6	F_Turkic, F_Common_Turkic, F_Oghuz-Kipchak-Uyghur, F_Oghuz, F_West_Oghuz, F_Nuclear_West_Oghuz

Table A.9. Part II of the list of genealogical features of PUD languages (with value equal to 1.0 in lang2vec tool).

Annex 10.

	arb	cmn	ces	eng	fin	fra	deu	hin	isl	ind	ita	jpn	kor	pol	por	rus	spa	swe	tha	tur
arb	0.00	3.16	3.46	3.87	3.00	4.24	3.46	3.61	3.46	3.74	3.87	3.00	2.65	3.32	4.12	3.16	4.24	3.46	4.00	3.46
cmn	3.16	0.00	3.16	3.61	2.65	4.00	3.16	3.32	3.16	3.46	3.61	2.65	2.24	3.00	3.87	2.83	4.00	3.16	3.74	3.16
ces	3.46	3.16	0.00	3.61	3.00	4.00	3.16	3.32	3.16	3.74	3.61	3.00	2.65	1.73	3.87	2.00	4.00	3.16	4.00	3.46
eng	3.87	3.61	3.61	0.00	3.46	4.36	2.65	3.74	3.00	4.12	4.00	3.46	3.16	3.46	4.24	3.32	4.36	3.00	4.36	3.87
fin	3.00	2.65	3.00	3.46	0.00	3.87	3.00	3.16	3.00	3.32	3.46	2.45	2.00	2.83	3.74	2.65	3.87	3.00	3.61	3.00
fra	4.24	4.00	4.00	4.36	3.87	0.00	4.00	4.12	4.00	4.47	2.65	3.87	3.61	3.87	2.24	3.74	2.45	4.00	4.69	4.24
deu	3.46	3.16	3.16	2.65	3.00	4.00	0.00	3.32	2.45	3.74	3.61	3.00	2.65	3.00	3.87	2.83	4.00	2.45	4.00	3.46
hin	3.61	3.32	3.32	3.74	3.16	4.12	3.32	0.00	3.32	3.87	3.74	3.16	2.83	3.16	4.00	3.00	4.12	3.32	4.12	3.61
isl	3.46	3.16	3.16	3.00	3.00	4.00	2.45	3.32	0.00	3.74	3.61	3.00	2.65	3.00	3.87	2.83	4.00	2.00	4.00	3.46
ind	3.74	3.46	3.74	4.12	3.32	4.47	3.74	3.87	3.74	0.00	4.12	3.32	3.00	3.61	4.36	3.46	4.47	3.74	4.24	3.74
ita	3.87	3.61	3.61	4.00	3.46	2.65	3.61	3.74	3.61	4.12	0.00	3.46	3.16	3.46	2.45	3.32	2.65	3.61	4.36	3.87
jpn	3.00	2.65	3.00	3.46	2.45	3.87	3.00	3.16	3.00	3.32	3.46	0.00	2.00	2.83	3.74	2.65	3.87	3.00	3.61	3.00
kor	2.65	2.24	2.65	3.16	2.00	3.61	2.65	2.83	2.65	3.00	3.16	2.00	0.00	2.45	3.46	2.24	3.61	2.65	3.32	2.65
pol	3.32	3.00	1.73	3.46	2.83	3.87	3.00	3.16	3.00	3.61	3.46	2.83	2.45	0.00	3.74	1.73	3.87	3.00	3.87	3.32
por	4.12	3.87	3.87	4.24	3.74	2.24	3.87	4.00	3.87	4.36	2.45	3.74	3.46	3.74	0.00	3.61	1.00	3.87	4.58	4.12
rus	3.16	2.83	2.00	3.32	2.65	3.74	2.83	3.00	2.83	3.46	3.32	2.65	2.24	1.73	3.61	0.00	3.74	2.83	3.74	3.16
spa	4.24	4.00	4.00	4.36	3.87	2.45	4.00	4.12	4.00	4.47	2.65	3.87	3.61	3.87	1.00	3.74	0.00	4.00	4.69	4.24
swe	3.46	3.16	3.16	3.00	3.00	4.00	2.45	3.32	2.00	3.74	3.61	3.00	2.65	3.00	3.87	2.83	4.00	0.00	4.00	3.46
tha	4.00	3.74	4.00	4.36	3.61	4.69	4.00	4.12	4.00	4.24	4.36	3.61	3.32	3.87	4.58	3.74	4.69	4.00	0.00	4.00
tur	3.46	3.16	3.46	3.87	3.00	4.24	3.46	3.61	3.46	3.74	3.87	3.00	2.65	3.32	4.12	3.16	4.24	3.46	4.00	0.00

Table A.10. Euclidean dissimilarity matrix considering genealogical features of lang2vec for all PUD corpora.

Annex 11.

	arb	cmn	ces	eng	fin	fra	deu	hin	isl	ind	ita	jpn	kor	pol	por	rus	spa	swe	tha	tur
arb	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
cmn	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ces	1.00	1.00	0.00	0.86	1.00	0.88	0.83	0.85	0.83	1.00	0.86	1.00	1.00	0.27	0.88	0.39	0.88	0.83	1.00	1.00
eng	1.00	1.00	0.86	0.00	1.00	0.90	0.46	0.87	0.59	1.00	0.89	1.00	1.00	0.85	0.90	0.83	0.90	0.59	1.00	1.00
fin	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
fra	1.00	1.00	0.88	0.90	1.00	0.00	0.88	0.89	0.88	1.00	0.33	1.00	1.00	0.87	0.22	0.86	0.25	0.88	1.00	1.00
deu	1.00	1.00	0.83	0.46	1.00	0.88	0.00	0.85	0.50	1.00	0.86	1.00	1.00	0.82	0.88	0.80	0.88	0.50	1.00	1.00
hin	1.00	1.00	0.85	0.87	1.00	0.89	0.85	0.00	0.85	1.00	0.87	1.00	1.00	0.83	0.89	0.81	0.89	0.85	1.00	1.00
isl	1.00	1.00	0.83	0.59	1.00	0.88	0.50	0.85	0.00	1.00	0.86	1.00	1.00	0.82	0.88	0.80	0.88	0.33	1.00	1.00
ind	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ita	1.00	1.00	0.86	0.89	1.00	0.33	0.86	0.87	0.86	1.00	0.00	1.00	1.00	0.85	0.30	0.83	0.33	0.86	1.00	1.00
jpn	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
kor	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
pol	1.00	1.00	0.27	0.85	1.00	0.87	0.82	0.83	0.82	1.00	0.85	1.00	1.00	0.00	0.87	0.33	0.87	0.82	1.00	1.00
por	1.00	1.00	0.88	0.90	1.00	0.22	0.88	0.89	0.88	1.00	0.30	1.00	1.00	0.87	0.00	0.85	0.04	0.88	1.00	1.00
rus	1.00	1.00	0.39	0.83	1.00	0.86	0.80	0.81	0.80	1.00	0.83	1.00	1.00	0.33	0.85	0.00	0.86	0.80	1.00	1.00
spa	1.00	1.00	0.88	0.90	1.00	0.25	0.88	0.89	0.88	1.00	0.33	1.00	1.00	0.87	0.04	0.86	0.00	0.88	1.00	1.00
swe	1.00	1.00	0.83	0.59	1.00	0.88	0.50	0.85	0.33	1.00	0.86	1.00	1.00	0.82	0.88	0.80	0.88	0.00	1.00	1.00
tha	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00
tur	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00

Table A.11. Cosine dissimilarity matrix considering genealogical features of lang2vec for all PUD corpora.

Annex 12.

Common syntactic features in all PUD languages
S_SVO
S_SOV
S_VSO
S_VOS
S_OVS
S_OSV
S_SUBJECT_BEFORE_VERB
S_SUBJECT_AFTER_VERB
S_OBJECT_AFTER_VERB
S_OBJECT_BEFORE_VERB
S_SUBJECT_BEFORE_OBJECT
S_SUBJECT_AFTER_OBJECT
S_ADPOSITION_BEFORE_NOUN
S_ADPOSITION_AFTER_NOUN
S_POSSESSOR_BEFORE_NOUN
S_POSSESSOR_AFTER_NOUN
S_ADJECTIVE_BEFORE_NOUN
S_ADJECTIVE_AFTER_NOUN
S_DEMONSTRATIVE_WORD_BEFORE_NOUN
S_DEMONSTRATIVE_WORD_AFTER_NOUN
S_DEMONSTRATIVE_PREFIX
S_DEMONSTRATIVE_SUFFIX
S_NUMERAL_BEFORE_NOUN
S_NUMERAL_AFTER_NOUN
S_NEGATIVE_WORD_BEFORE_VERB
S_NEGATIVE_PREFIX
S_NEGATIVE_WORD_AFTER_VERB
S_NEGATIVE_SUFFIX
S_TEND_PREFIX
S_TEND_SUFFIX
S_CASE_PREFIX
S_CASE_SUFFIX
S_CASE_PROCLITIC
S_CASE_ENCLITIC
S_CASE_MARK
S_DEGREE_WORD_BEFORE_ADJECTIVE
S_DEGREE_WORD_AFTER_ADJECTIVE
S_POLARQ_MARK_INITIAL
S_POLARQ_MARK_FINAL
S_SUBORDINATOR_WORD_BEFORE_CLAUSE
S_SUBORDINATOR_WORD_AFTER_CLAUSE

Table A.12. List of syntactic features which are common to all PUD languages.

Annex 13.

	arb	cmn	ces	eng	fin	fra	deu	hin	isl	ind	ita	jpn	kor	pol	por	rus	spa	swe	tha	tur
arb	0.00	3.86	3.45	3.30	3.46	3.04	3.39	3.71	3.16	3.08	2.70	4.66	4.14	2.64	2.68	2.94	2.92	3.54	3.62	4.25
cmn	3.86	0.00	2.59	1.95	1.99	3.14	2.78	2.83	2.84	3.10	2.52	3.10	2.80	2.82	2.49	2.41	2.39	2.02	3.42	3.12
ces	3.45	2.59	0.00	2.41	1.72	2.97	2.33	2.94	2.14	3.33	2.75	3.56	3.17	2.23	2.69	1.67	2.59	2.47	3.96	3.03
eng	3.30	1.95	2.41	0.00	2.42	2.27	2.21	3.31	2.05	2.43	1.74	3.60	3.37	2.35	1.50	1.86	1.56	1.11	3.00	3.52
fin	3.46	1.99	1.72	2.42	0.00	2.98	2.47	2.42	2.26	3.33	2.79	3.21	2.59	2.33	2.69	1.75	2.59	2.26	3.97	2.66
fra	3.04	3.14	2.97	2.27	2.98	0.00	2.60	3.61	2.06	2.40	2.09	4.12	3.79	2.13	1.90	2.36	2.11	2.40	3.52	4.12
deu	3.39	2.78	2.33	2.21	2.47	2.60	0.00	2.78	1.87	3.24	2.51	3.50	2.76	2.11	2.62	1.68	2.66	2.12	3.82	3.19
hin	3.71	2.83	2.94	3.31	2.42	3.61	2.78	0.00	3.35	3.84	3.40	2.54	1.83	2.95	3.48	2.96	3.65	3.28	4.46	2.30
isl	3.16	2.84	2.14	2.05	2.26	2.06	1.87	3.35	0.00	2.65	2.37	3.79	3.26	1.81	2.20	1.53	2.13	2.12	3.33	3.56
ind	3.08	3.10	3.33	2.43	3.33	2.40	3.24	3.84	2.65	0.00	1.91	4.46	4.14	2.40	1.69	2.80	2.09	2.74	2.37	4.38
ita	2.70	2.52	2.75	1.74	2.79	2.09	2.51	3.40	2.37	1.91	0.00	3.98	3.63	1.89	0.75	2.07	0.99	2.15	2.92	3.76
jpn	4.66	3.10	3.56	3.60	3.21	4.12	3.50	2.54	3.79	4.46	3.98	0.00	1.94	3.96	4.07	3.70	4.03	3.64	4.68	1.88
kor	4.14	2.80	3.17	3.37	2.59	3.79	2.76	1.83	3.26	4.14	3.63	1.94	0.00	3.33	3.74	3.02	3.70	3.26	4.47	1.85
pol	2.64	2.82	2.23	2.35	2.33	2.13	2.11	2.95	1.81	2.40	1.89	3.96	3.33	0.00	1.92	1.39	1.99	2.60	3.52	3.46
por	2.68	2.49	2.69	1.50	2.69	1.90	2.62	3.48	2.20	1.69	0.75	4.07	3.74	1.92	0.00	1.99	0.80	1.96	2.82	3.92
rus	2.94	2.41	1.67	1.86	1.75	2.36	1.68	2.96	1.53	2.80	2.07	3.70	3.02	1.39	1.99	0.00	1.86	2.04	3.53	3.24
spa	2.92	2.39	2.59	1.56	2.59	2.11	2.66	3.65	2.13	2.09	0.99	4.03	3.70	1.99	0.80	1.86	0.00	2.01	3.00	3.85
swe	3.54	2.02	2.47	1.11	2.26	2.40	2.12	3.28	2.12	2.74	2.15	3.64	3.26	2.60	1.96	2.04	2.01	0.00	3.41	3.63
tha	3.62	3.42	3.96	3.00	3.97	3.52	3.82	4.46	3.33	2.37	2.92	4.68	4.47	3.52	2.82	3.53	3.00	3.41	0.00	4.75
tur	4.25	3.12	3.03	3.52	2.66	4.12	3.19	2.30	3.56	4.38	3.76	1.88	1.85	3.46	3.92	3.24	3.85	3.63	4.75	0.00

Table A.13. Euclidean dissimilarity matrix considering syntactic features (“syntax_average”) of lang2vec for all PUD corpora.

Annex 14.

	arb	cmn	ces	eng	fin	fra	deu	hin	isl	ind	ita	jpn	kor	pol	por	rus	spa	swe	tha	tur
arb	0.00	0.47	0.37	0.36	0.39	0.28	0.33	0.43	0.31	0.31	0.24	0.68	0.53	0.21	0.24	0.28	0.28	0.41	0.45	0.56
cmn	0.47	0.00	0.21	0.13	0.13	0.30	0.22	0.26	0.26	0.32	0.21	0.31	0.25	0.24	0.21	0.19	0.19	0.14	0.41	0.31
ces	0.37	0.21	0.00	0.18	0.09	0.26	0.15	0.27	0.14	0.36	0.24	0.39	0.31	0.15	0.23	0.09	0.22	0.19	0.53	0.28
eng	0.36	0.13	0.18	0.00	0.20	0.16	0.14	0.36	0.14	0.20	0.10	0.43	0.37	0.17	0.08	0.11	0.09	0.04	0.32	0.41
fin	0.39	0.13	0.09	0.20	0.00	0.28	0.18	0.19	0.17	0.38	0.27	0.34	0.22	0.17	0.25	0.10	0.24	0.18	0.57	0.23
fra	0.28	0.30	0.26	0.16	0.28	0.00	0.19	0.39	0.13	0.18	0.13	0.51	0.43	0.13	0.11	0.17	0.14	0.18	0.40	0.51
deu	0.33	0.22	0.15	0.14	0.18	0.19	0.00	0.22	0.10	0.31	0.18	0.36	0.22	0.12	0.20	0.08	0.21	0.13	0.45	0.29
hin	0.43	0.26	0.27	0.36	0.19	0.39	0.22	0.00	0.35	0.49	0.38	0.21	0.11	0.26	0.41	0.28	0.45	0.35	0.69	0.17
isl	0.31	0.26	0.14	0.14	0.17	0.13	0.10	0.35	0.00	0.23	0.18	0.45	0.33	0.10	0.16	0.07	0.15	0.15	0.38	0.40
ind	0.31	0.32	0.36	0.20	0.38	0.18	0.31	0.49	0.23	0.00	0.12	0.66	0.56	0.18	0.10	0.26	0.15	0.26	0.20	0.63
ita	0.24	0.21	0.24	0.10	0.27	0.13	0.18	0.38	0.18	0.12	0.00	0.52	0.43	0.11	0.02	0.14	0.03	0.16	0.31	0.46
jpn	0.68	0.31	0.39	0.43	0.34	0.51	0.36	0.21	0.45	0.66	0.52	0.00	0.12	0.47	0.56	0.44	0.55	0.44	0.76	0.11
kor	0.53	0.25	0.31	0.37	0.22	0.43	0.22	0.11	0.33	0.56	0.43	0.12	0.00	0.33	0.46	0.29	0.45	0.35	0.68	0.11
pol	0.21	0.24	0.15	0.17	0.17	0.13	0.12	0.26	0.10	0.18	0.11	0.47	0.33	0.00	0.11	0.06	0.12	0.21	0.40	0.36
por	0.24	0.21	0.23	0.08	0.25	0.11	0.20	0.41	0.16	0.10	0.02	0.56	0.46	0.11	0.00	0.13	0.02	0.14	0.30	0.51
rus	0.28	0.19	0.09	0.11	0.10	0.17	0.08	0.28	0.07	0.26	0.14	0.44	0.29	0.06	0.13	0.00	0.12	0.14	0.44	0.33
spa	0.28	0.19	0.22	0.09	0.24	0.14	0.21	0.45	0.15	0.15	0.03	0.55	0.45	0.12	0.02	0.12	0.00	0.14	0.33	0.49
swe	0.41	0.14	0.19	0.04	0.18	0.18	0.13	0.35	0.15	0.26	0.16	0.44	0.35	0.21	0.14	0.14	0.14	0.00	0.42	0.43
tha	0.45	0.41	0.53	0.32	0.57	0.40	0.45	0.69	0.38	0.20	0.31	0.76	0.68	0.40	0.30	0.44	0.33	0.42	0.00	0.77
tur	0.56	0.31	0.28	0.41	0.23	0.51	0.29	0.17	0.40	0.63	0.46	0.11	0.11	0.36	0.51	0.33	0.49	0.43	0.77	0.00

Table A.14. Cosine dissimilarity matrix considering syntactic features (“syntax_average”) of lang2vec for all PUD corpora.

Annex 15.

Linear properties in all PUD languages	
VERB-+_exclude_NOUN-obj_VERB-ccomp	VERB-+_exclude_NOUN-obj_PRON-obj
NOUN-+_exclude_NUM-nummod_NOUN-appos	NOUN-+_exclude_ADJ-amod_NOUN-appos
VERB-+_exclude_ADV-advmod_PROPN-obj	VERB-+_unicity_CCONJ-cc_-
VERB-+_exclude_PRON-nsubj_ADJ-ccomp	VERB-+_unicity_VERB-ccomp_-
NOUN-+_exclude_CCONJ-cc_PROPN-appos	VERB-+_exclude_VERB-advcl_NUM-obj
VERB-+_exclude_PROPN-obj_NUM-obj	NOUN-+_unicity_NOUN-appos_-
PROPN-+_precede_*_NOUN-appos	VERB-+_exclude_NOUN-nsubj_ADJ-advcl
VERB-+_exclude_PRON-obj_NUM-obj	VERB-+_unicity_PROPN-obj_-
PROPN-+_unicity_AUX-cop_-	NOUN-+_unicity_PROPN-nsubj_-
VERB-+_exclude_PROPN-nsubj_PROPN-obj	NOUN-+_exclude_NUM-nummod_ADV-advmod
VERB-+_exclude_PRON-nsubj_NUM-obj	VERB-+_exclude_NOUN-obj_NUM-obj
NOUN-+_exclude_ADJ-amod_CCONJ-cc	VERB-+_precede_PRON-nsubj_NOUN-obj
VERB-+_exclude_PRON-nsubj_PROPN-obj	VERB-+_exclude_PRON-nsubj_ADJ-advcl
NOUN-+_unicity_PROPN-appos_-	VERB-+_unicity_PRON-obj_-
VERB-+_exclude_PUNCT-punct_ADJ-advcl	VERB-+_exclude_PUNCT-punct_ADJ-ccomp
PROPN-+_precede_CCONJ-cc_*	VERB-+_exclude_NOUN-nsubj_PROPN-nsubj
NOUN-+_exclude_ADJ-amod_PROPN-appos	VERB-+_unicity_ADJ-ccomp_-
VERB-+_precede_NOUN-nsubj_NOUN-obj	VERB-+_exclude_NOUN-obj_PROPN-obj
VERB-+_exclude_PRON-obj_ADJ-ccomp	NOUN-+_exclude_NUM-nummod_PROPN-nsubj
VERB-+_exclude_ADV-advmod_NOUN-advcl	PROPN-+_unicity_PROPN-appos_-
PROPN-+_precede_*_PROPN-appos	VERB-+_unicity_PRON-nsubj_-
VERB-+_exclude_NOUN-obj_ADJ-ccomp	VERB-+_exclude_ADV-advmod_ADJ-ccomp
VERB-+_exclude_NOUN-obj_ADJ-advcl	NOUN-+_unicity_AUX-cop_-
VERB-+_exclude_VERB-ccomp_NUM-obj	VERB-+_exclude_ADV-advmod_ADJ-advcl
VERB-+_exclude_ADV-advmod_NUM-obj	PROPN-+_unicity_CCONJ-cc_-
VERB-+_exclude_VERB-advcl_ADJ-advcl	NOUN-+_precede_CCONJ-cc_*
NOUN-+_precede_*_NOUN-appos	VERB-+_exclude_ADJ-ccomp_NUM-obj
VERB-+_unicity_NUM-obj_-	VERB-+_unicity_NOUN-nsubj_-
NOUN-+_precede_*_PROPN-appos	VERB-+_exclude_NOUN-nsubj_NUM-obj
VERB-+_unicity_NOUN-advcl_-	VERB-+_exclude_ADV-advmod_VERB-ccomp
ADJ-+_unicity_VERB-advcl_-	VERB-+_exclude_PROPN-nsubj_NUM-obj
VERB-+_unicity_NOUN-obj_-	VERB-+_exclude_NOUN-nsubj_ADJ-ccomp
VERB-+_precede_PROPN-nsubj_NOUN-obj	VERB-+_exclude_NOUN-nsubj_PROPN-obj
VERB-+_exclude_VERB-ccomp_ADJ-ccomp	VERB-+_precede_CCONJ-cc_*
NOUN-+_exclude_CCONJ-cc_NOUN-appos	NOUN-+_exclude_AUX-cop_NOUN-appos
VERB-+_unicity_PROPN-nsubj_-	VERB-+_exclude_VERB-advcl_ADJ-ccomp
PROPN-+_exclude_PUNCT-punct_PROPN-appos	VERB-+_exclude_PROPN-nsubj_ADJ-advcl
NOUN-+_unicity_CCONJ-cc_-	VERB-+_exclude_NOUN-nsubj_CCONJ-cc
VERB-+_exclude_PROPN-nsubj_ADJ-ccomp	VERB-+_exclude_PUNCT-punct_NUM-obj

Table A.15. List of MarsaGram patterns which occur in all PUD languages (i.e.: frequency higher than 0.0). The “*” symbol indicates that the element corresponds to the head of the subtree.

Annex 16.

	arb	cmn	ces	eng	fin	fra	deu	hin	isl	ind	ita	jpn	kor	pol	por	rus	spa	swe	tha	tur
arb	0.00	19.18	18.64	18.99	18.90	18.89	18.93	20.38	19.65	17.36	19.50	18.67	18.04	19.14	18.93	18.86	19.30	17.77	17.42	17.91
cmn	19.18	0.00	19.90	19.82	18.86	19.74	19.36	21.44	20.51	17.78	20.86	19.38	16.70	19.87	20.41	20.01	20.46	19.19	17.81	18.07
ces	18.64	19.90	0.00	18.92	18.65	19.40	18.56	21.55	19.31	17.92	19.48	19.61	18.83	17.75	19.34	18.07	19.25	18.02	18.71	18.05
eng	18.99	19.82	18.92	0.00	18.98	17.91	17.08	20.37	19.15	16.90	17.36	19.79	19.19	19.96	17.34	18.50	18.45	16.17	18.11	18.97
fin	18.90	18.86	18.65	18.98	0.00	18.92	18.43	21.30	18.59	17.07	20.11	19.38	17.39	19.47	19.66	18.39	19.71	17.64	17.75	17.41
fra	18.89	19.74	19.40	17.91	18.92	0.00	18.16	21.31	20.11	17.39	17.81	20.07	19.15	19.76	17.24	18.59	17.54	17.75	18.38	19.19
deu	18.93	19.36	18.56	17.08	18.43	18.16	0.00	20.74	19.80	17.24	18.24	20.22	19.06	19.65	18.20	18.40	18.37	17.58	18.86	18.37
hin	20.38	21.44	21.55	20.37	21.30	21.31	20.74	0.00	21.78	19.90	21.18	21.01	20.84	21.89	20.64	21.38	21.24	19.90	19.90	20.94
isl	19.65	20.51	19.31	19.15	18.59	20.11	19.80	21.78	0.00	18.35	20.34	20.30	19.48	20.49	20.13	19.58	20.66	17.53	18.76	19.27
ind	17.36	17.78	17.92	16.90	17.07	17.39	17.24	19.90	18.35	0.00	18.13	18.04	16.20	18.18	17.46	17.47	18.24	16.73	16.35	17.16
ita	19.50	20.86	19.48	17.36	20.11	17.81	18.24	21.18	20.34	18.13	0.00	20.55	20.31	20.60	17.99	19.54	18.13	18.43	19.12	19.77
jpn	18.67	19.38	19.61	19.79	19.38	20.07	20.22	21.01	20.30	18.04	20.55	0.00	17.64	20.19	20.40	20.31	20.46	18.89	17.86	18.82
kor	18.04	16.70	18.83	19.19	17.39	19.15	19.06	20.84	19.48	16.20	20.31	17.64	0.00	18.71	19.44	19.33	19.50	18.38	15.94	16.78
pol	19.14	19.87	17.75	19.96	19.47	19.76	19.65	21.89	20.49	18.18	20.60	20.19	18.71	0.00	20.35	19.31	20.07	19.55	18.95	18.88
por	18.93	20.41	19.34	17.34	19.66	17.24	18.20	20.64	20.13	17.46	17.99	20.40	19.44	20.35	0.00	19.29	15.98	18.61	18.68	19.55
rus	18.86	20.01	18.07	18.50	18.39	18.59	18.40	21.38	19.58	17.47	19.54	20.31	19.33	19.31	19.29	0.00	19.45	17.39	18.72	18.66
spa	19.30	20.46	19.25	18.45	19.71	17.54	18.37	21.24	20.66	18.24	18.13	20.46	19.50	20.07	15.98	19.45	0.00	18.93	18.77	19.65
swe	17.77	19.19	18.02	16.17	17.64	17.75	17.58	19.90	17.53	16.73	18.43	18.89	18.38	19.55	18.61	17.39	18.93	0.00	17.46	18.16
tha	17.42	17.81	18.71	18.11	17.75	18.38	18.86	19.90	18.76	16.35	19.12	17.86	15.94	18.95	18.68	18.72	18.77	17.46	0.00	17.77
tur	17.91	18.07	18.05	18.97	17.41	19.19	18.37	20.94	19.27	17.16	19.77	18.82	16.78	18.88	19.55	18.66	19.65	18.16	17.77	0.00

Table A.16. Euclidean dissimilarity matrix considering all MarsaGram properties for all PUD corpora.

Annex 17.

	arb	cmn	ces	eng	fin	fra	deu	hin	isl	ind	it	jpn	kor	pol	por	rus	spa	swe	tha	tur
arb	0.00	0.79	0.67	0.67	0.77	0.67	0.66	0.67	0.70	0.70	0.66	0.74	0.90	0.73	0.65	0.66	0.68	0.62	0.73	0.71
cmn	0.79	0.00	0.80	0.76	0.81	0.77	0.72	0.77	0.80	0.78	0.79	0.84	0.82	0.82	0.79	0.78	0.80	0.75	0.80	0.76
ces	0.67	0.80	0.00	0.63	0.71	0.67	0.61	0.73	0.65	0.70	0.63	0.77	0.91	0.60	0.65	0.58	0.65	0.60	0.78	0.67
eng	0.67	0.76	0.63	0.00	0.70	0.55	0.50	0.62	0.61	0.58	0.48	0.75	0.89	0.72	0.50	0.58	0.57	0.46	0.69	0.71
fin	0.77	0.81	0.71	0.70	0.00	0.70	0.66	0.77	0.66	0.72	0.73	0.84	0.90	0.79	0.74	0.66	0.74	0.64	0.80	0.71
fra	0.67	0.77	0.67	0.55	0.70	0.00	0.57	0.69	0.69	0.63	0.51	0.78	0.90	0.72	0.50	0.60	0.52	0.57	0.73	0.74
deu	0.66	0.72	0.61	0.50	0.66	0.57	0.00	0.65	0.66	0.61	0.53	0.78	0.87	0.70	0.55	0.58	0.57	0.55	0.75	0.66
hin	0.67	0.77	0.73	0.62	0.77	0.69	0.65	0.00	0.71	0.70	0.64	0.74	0.88	0.76	0.63	0.69	0.67	0.62	0.72	0.75
isl	0.70	0.80	0.65	0.61	0.66	0.69	0.66	0.71	0.00	0.68	0.65	0.78	0.89	0.75	0.67	0.64	0.71	0.54	0.73	0.72
ind	0.70	0.78	0.70	0.58	0.72	0.63	0.61	0.70	0.68	0.00	0.62	0.79	0.87	0.74	0.61	0.63	0.68	0.61	0.75	0.75
ita	0.66	0.79	0.63	0.48	0.73	0.51	0.53	0.64	0.65	0.62	0.00	0.76	0.92	0.72	0.51	0.61	0.52	0.57	0.71	0.72
jpn	0.74	0.84	0.77	0.75	0.84	0.78	0.78	0.74	0.78	0.79	0.76	0.00	0.91	0.84	0.78	0.79	0.79	0.72	0.80	0.81
kor	0.90	0.82	0.91	0.89	0.90	0.90	0.87	0.88	0.89	0.87	0.92	0.91	0.00	0.93	0.89	0.90	0.90	0.87	0.88	0.86
pol	0.73	0.82	0.60	0.72	0.79	0.72	0.70	0.76	0.75	0.74	0.72	0.84	0.93	0.00	0.74	0.68	0.72	0.73	0.83	0.76
por	0.65	0.79	0.65	0.50	0.74	0.50	0.55	0.63	0.67	0.61	0.51	0.78	0.89	0.74	0.00	0.62	0.42	0.60	0.72	0.74
rus	0.66	0.78	0.58	0.58	0.66	0.60	0.58	0.69	0.64	0.63	0.61	0.79	0.90	0.68	0.62	0.00	0.64	0.54	0.74	0.69
spa	0.68	0.80	0.65	0.57	0.74	0.52	0.57	0.67	0.71	0.68	0.52	0.79	0.90	0.72	0.42	0.64	0.00	0.63	0.73	0.75
swe	0.62	0.75	0.60	0.46	0.64	0.57	0.55	0.62	0.54	0.61	0.57	0.72	0.87	0.73	0.60	0.54	0.63	0.00	0.68	0.69
tha	0.73	0.80	0.78	0.69	0.80	0.73	0.75	0.72	0.73	0.75	0.71	0.80	0.88	0.83	0.72	0.74	0.73	0.68	0.00	0.82
tur	0.71	0.76	0.67	0.71	0.71	0.74	0.66	0.75	0.72	0.75	0.72	0.81	0.86	0.76	0.74	0.69	0.75	0.69	0.82	0.00

Table A.17. Cosine dissimilarity matrix considering all MarsaGram properties for all PUD corpora.

Annex 18.

	arb	cmn	ces	eng	fin	fra	deu	hin	isl	ind	ita	jpn	kor	pol	por	rus	spa	swe	tha	tur
arb	0.00	3.60	3.16	2.82	3.70	2.87	3.04	3.71	5.29	3.53	3.47	4.89	3.05	3.10	3.07	3.13	3.66	2.88	2.85	3.73
cmn	3.60	0.00	3.97	3.59	4.12	3.73	3.74	4.27	5.66	4.05	4.26	4.95	3.31	3.94	3.92	3.88	4.39	3.46	3.63	3.86
ces	3.16	3.97	0.00	2.97	3.76	3.07	2.53	3.98	5.26	3.86	3.71	5.15	3.44	3.12	3.29	2.86	3.57	3.04	3.19	3.90
eng	2.82	3.59	2.97	0.00	3.55	2.34	2.64	3.77	5.05	3.53	2.92	4.96	3.12	3.11	2.70	2.85	3.17	2.54	3.01	3.59
fin	3.70	4.12	3.76	3.55	0.00	3.54	3.40	3.95	5.38	4.03	4.32	5.18	3.51	3.91	3.98	3.34	4.13	3.64	3.50	4.12
fra	2.87	3.73	3.07	2.34	3.54	0.00	2.66	3.84	5.15	3.49	2.81	5.01	3.20	3.09	2.36	2.56	3.05	2.66	3.00	3.79
deu	3.04	3.74	2.53	2.64	3.40	2.66	0.00	3.67	5.04	3.57	3.28	4.96	3.14	3.17	3.03	2.64	3.07	2.91	2.79	3.67
hin	3.71	4.27	3.98	3.77	3.95	3.84	3.67	0.00	5.72	4.34	4.36	4.54	3.58	4.03	3.89	3.86	4.08	3.89	3.51	3.99
isl	5.29	5.66	5.26	5.05	5.38	5.15	5.04	5.72	0.00	5.65	5.52	6.61	5.37	5.24	5.41	5.07	5.51	5.25	5.13	5.65
ind	3.53	4.05	3.86	3.53	4.03	3.49	3.57	4.34	5.65	0.00	3.91	5.05	3.63	3.82	3.51	3.69	4.05	3.55	3.36	4.08
ita	3.47	4.26	3.71	2.92	4.32	2.81	3.28	4.36	5.52	3.91	0.00	5.33	3.87	3.71	3.23	3.61	2.92	3.52	3.62	4.38
jpn	4.89	4.95	5.15	4.96	5.18	5.01	4.96	4.54	6.61	5.05	5.33	0.00	4.51	5.12	5.15	5.13	5.43	5.01	4.81	4.80
kor	3.05	3.31	3.44	3.12	3.51	3.20	3.14	3.58	5.37	3.63	3.87	4.51	0.00	3.41	3.42	3.40	3.95	3.18	3.15	3.39
pol	3.10	3.94	3.12	3.11	3.91	3.09	3.17	4.03	5.24	3.82	3.71	5.12	3.41	0.00	3.34	3.06	3.70	3.23	3.22	3.99
por	3.07	3.92	3.29	2.70	3.98	2.36	3.03	3.89	5.41	3.51	3.23	5.15	3.42	3.34	0.00	3.21	2.73	2.96	3.28	4.03
rus	3.13	3.88	2.86	2.85	3.34	2.56	2.64	3.86	5.07	3.69	3.61	5.13	3.40	3.06	3.21	0.00	3.31	3.06	2.75	3.88
spa	3.66	4.39	3.57	3.17	4.13	3.05	3.07	4.08	5.51	4.05	2.92	5.43	3.95	3.70	2.73	3.31	0.00	3.56	3.55	4.49
swe	2.88	3.46	3.04	2.54	3.64	2.66	2.91	3.89	5.25	3.55	3.52	5.01	3.18	3.23	2.96	3.06	3.56	0.00	3.10	3.76
tha	2.85	3.63	3.19	3.01	3.50	3.00	2.79	3.51	5.13	3.36	3.62	4.81	3.15	3.22	3.28	2.75	3.55	3.10	0.00	3.59
tur	3.73	3.86	3.90	3.59	4.12	3.79	3.67	3.99	5.65	4.08	4.38	4.80	3.39	3.99	4.03	3.88	4.49	3.76	3.59	0.00

Table A.18. Euclidean dissimilarity matrix considering MarsaGram linear properties for all PUD corpora.

Annex 19.

	arb	cmn	ces	eng	fin	fra	deu	hin	isl	ind	ita	jpn	kor	pol	por	rus	spa	swe	tha	tur
arb	0.00	0.90	0.76	0.68	0.96	0.66	0.79	0.90	0.88	0.81	0.69	0.99	0.97	0.75	0.68	0.74	0.74	0.68	0.70	0.97
cmn	0.90	0.00	0.93	0.82	0.94	0.86	0.90	0.96	0.92	0.87	0.88	0.87	0.81	0.93	0.88	0.89	0.90	0.75	0.86	0.82
ces	0.76	0.93	0.00	0.62	0.84	0.63	0.45	0.89	0.80	0.84	0.70	0.99	0.98	0.63	0.67	0.52	0.62	0.63	0.72	0.90
eng	0.68	0.82	0.62	0.00	0.81	0.40	0.54	0.86	0.75	0.76	0.45	0.97	0.91	0.69	0.48	0.56	0.51	0.49	0.71	0.83
fin	0.96	0.94	0.84	0.81	0.00	0.78	0.74	0.83	0.82	0.86	0.91	0.96	0.93	0.93	0.91	0.66	0.80	0.84	0.80	0.94
fra	0.66	0.86	0.63	0.40	0.78	0.00	0.52	0.87	0.78	0.72	0.41	0.97	0.90	0.65	0.36	0.44	0.46	0.51	0.68	0.89
deu	0.79	0.90	0.45	0.54	0.74	0.52	0.00	0.81	0.75	0.77	0.58	0.97	0.92	0.71	0.61	0.48	0.48	0.64	0.61	0.87
hin	0.90	0.96	0.89	0.86	0.83	0.87	0.81	0.00	0.92	0.95	0.88	0.70	0.89	0.93	0.82	0.84	0.75	0.90	0.76	0.84
isl	0.88	0.92	0.80	0.75	0.82	0.78	0.75	0.92	0.00	0.90	0.81	0.99	0.96	0.80	0.85	0.73	0.79	0.82	0.79	0.92
ind	0.81	0.87	0.84	0.76	0.86	0.72	0.77	0.95	0.90	0.00	0.71	0.88	0.93	0.84	0.67	0.77	0.74	0.75	0.70	0.88
ita	0.69	0.88	0.70	0.45	0.91	0.41	0.58	0.88	0.81	0.71	0.00	0.93	0.94	0.71	0.51	0.66	0.36	0.66	0.72	0.93
jpn	0.99	0.87	0.99	0.97	0.96	0.97	0.97	0.70	0.99	0.88	0.93	0.00	0.85	1.00	0.97	0.99	0.95	0.98	0.91	0.82
kor	0.97	0.81	0.98	0.91	0.93	0.90	0.92	0.89	0.96	0.93	0.94	0.85	0.00	0.99	0.92	0.95	0.93	0.91	0.95	0.85
pol	0.75	0.93	0.63	0.69	0.93	0.65	0.71	0.93	0.80	0.84	0.71	1.00	0.99	0.00	0.70	0.61	0.68	0.73	0.75	0.96
por	0.68	0.88	0.67	0.48	0.91	0.36	0.61	0.82	0.85	0.67	0.51	0.97	0.92	0.70	0.00	0.63	0.35	0.57	0.73	0.93
rus	0.74	0.89	0.52	0.56	0.66	0.44	0.48	0.84	0.73	0.77	0.66	0.99	0.95	0.61	0.63	0.00	0.53	0.64	0.53	0.89
spa	0.74	0.90	0.62	0.51	0.80	0.46	0.48	0.75	0.79	0.74	0.36	0.95	0.93	0.68	0.35	0.53	0.00	0.65	0.66	0.94
swe	0.68	0.75	0.63	0.49	0.84	0.51	0.64	0.90	0.82	0.75	0.66	0.98	0.91	0.73	0.57	0.64	0.65	0.00	0.74	0.89
tha	0.70	0.86	0.72	0.71	0.80	0.68	0.61	0.76	0.79	0.70	0.72	0.91	0.95	0.75	0.73	0.53	0.66	0.74	0.00	0.84
tur	0.97	0.82	0.90	0.83	0.94	0.89	0.87	0.84	0.92	0.88	0.93	0.82	0.85	0.96	0.93	0.89	0.94	0.89	0.84	0.00

Table A.19. Cosine dissimilarity matrix considering MarsaGram linear properties for all PUD corpora.

Annex 20.

Features
ADJ_amod_precedes_NOUN
ADV_advmod_precedes_ADJ
ADV_advmod_precedes_NOUN
ADV_advmod_precedes_VERB
CCONJ_cc_precedes_NOUN
CCONJ_cc_precedes_PROPN
CCONJ_cc_precedes_VERB
NOUN_appos_follows_NOUN
NOUN_appos_follows_PROPN
NOUN_nsubj_precedes_VERB
NUM_nummod_precedes_NOUN
PRON_nsubj_precedes_VERB
PROPN_appos_follows_NOUN
PROPN_appos_follows_PROPN
PROPN_nsubj_precedes_NOUN
PROPN_nsubj_precedes_VERB
PUNCT_punct_precedes_ADJ
PUNCT_punct_precedes_NOUN
PUNCT_punct_precedes_NUM
PUNCT_punct_precedes_PROPN
PUNCT_punct_precedes_VERB
PUNCT_punct_follows_ADJ
PUNCT_punct_follows_NOUN
PUNCT_punct_follows_NUM
PUNCT_punct_follows_PRON
PUNCT_punct_follows_PROPN
PUNCT_punct_follows_VERB
VERB_advcl_precedes_VERB

Table A.20. List of head directionality features attested in all PUD languages.

Annex 21.

	arb	cmn	ces	eng	fin	fra	deu	hin	isl	ind	ita	jpn	kor	pol	por	rus	spa	swe	tha	tur
arb	0.00	0.21	0.14	0.16	0.19	0.17	0.19	0.25	0.13	0.11	0.16	0.30	0.25	0.13	0.15	0.14	0.15	0.14	0.15	0.23
cmn	0.21	0.00	0.16	0.15	0.14	0.21	0.18	0.20	0.14	0.16	0.20	0.25	0.16	0.15	0.20	0.17	0.20	0.15	0.17	0.17
ces	0.14	0.16	0.00	0.10	0.13	0.16	0.13	0.20	0.09	0.12	0.15	0.28	0.22	0.08	0.15	0.05	0.15	0.08	0.16	0.17
eng	0.16	0.15	0.10	0.00	0.14	0.10	0.07	0.20	0.11	0.12	0.09	0.28	0.20	0.11	0.09	0.10	0.10	0.06	0.15	0.17
fin	0.19	0.14	0.13	0.14	0.00	0.21	0.17	0.17	0.11	0.14	0.20	0.26	0.19	0.13	0.20	0.13	0.20	0.12	0.16	0.14
fra	0.17	0.21	0.16	0.10	0.21	0.00	0.09	0.24	0.17	0.15	0.05	0.31	0.25	0.16	0.04	0.16	0.05	0.13	0.18	0.22
deu	0.19	0.18	0.13	0.07	0.17	0.09	0.00	0.20	0.15	0.15	0.09	0.29	0.22	0.14	0.09	0.13	0.09	0.10	0.18	0.18
hin	0.25	0.20	0.20	0.20	0.17	0.24	0.20	0.00	0.20	0.22	0.24	0.15	0.21	0.20	0.24	0.21	0.24	0.20	0.22	0.15
isl	0.13	0.14	0.09	0.11	0.11	0.17	0.15	0.20	0.00	0.09	0.16	0.27	0.20	0.08	0.16	0.10	0.16	0.07	0.12	0.17
ind	0.11	0.16	0.12	0.12	0.14	0.15	0.15	0.22	0.09	0.00	0.15	0.28	0.21	0.11	0.14	0.11	0.14	0.11	0.13	0.19
ita	0.16	0.20	0.15	0.09	0.20	0.05	0.09	0.24	0.16	0.15	0.00	0.31	0.24	0.15	0.04	0.15	0.05	0.12	0.17	0.21
jpn	0.30	0.25	0.28	0.28	0.26	0.31	0.29	0.15	0.27	0.28	0.31	0.00	0.26	0.27	0.31	0.29	0.31	0.28	0.28	0.24
kor	0.25	0.16	0.22	0.20	0.19	0.25	0.22	0.21	0.20	0.21	0.24	0.26	0.00	0.20	0.24	0.23	0.25	0.21	0.21	0.17
pol	0.13	0.15	0.08	0.11	0.13	0.16	0.14	0.20	0.08	0.11	0.15	0.27	0.20	0.00	0.15	0.10	0.15	0.09	0.13	0.17
por	0.15	0.20	0.15	0.09	0.20	0.04	0.09	0.24	0.16	0.14	0.04	0.31	0.24	0.15	0.00	0.15	0.03	0.12	0.17	0.21
rus	0.14	0.17	0.05	0.10	0.13	0.16	0.13	0.21	0.10	0.11	0.15	0.29	0.23	0.10	0.15	0.00	0.15	0.08	0.16	0.19
spa	0.15	0.20	0.15	0.10	0.20	0.05	0.09	0.24	0.16	0.14	0.05	0.31	0.25	0.15	0.03	0.15	0.00	0.12	0.17	0.22
swe	0.14	0.15	0.08	0.06	0.12	0.13	0.10	0.20	0.07	0.11	0.12	0.28	0.21	0.09	0.12	0.08	0.12	0.00	0.14	0.17
tha	0.15	0.17	0.16	0.15	0.16	0.18	0.18	0.22	0.12	0.13	0.17	0.28	0.21	0.13	0.17	0.16	0.17	0.14	0.00	0.20
tur	0.23	0.17	0.17	0.17	0.14	0.22	0.18	0.15	0.17	0.19	0.21	0.24	0.17	0.17	0.21	0.19	0.22	0.17	0.20	0.00

Table A.21. Euclidean dissimilarity matrix considering head directionality features for all PUD corpora.

Annex 22.

	arb	cmn	ces	eng	fin	fra	deu	hin	isl	ind	ita	jpn	kor	pol	por	rus	spa	swe	tha	tur
arb	0.00	0.68	0.31	0.39	0.62	0.35	0.48	0.88	0.28	0.17	0.32	0.92	0.91	0.25	0.30	0.27	0.29	0.32	0.39	0.85
cmn	0.68	0.00	0.48	0.42	0.42	0.58	0.52	0.67	0.42	0.48	0.57	0.66	0.48	0.47	0.58	0.49	0.58	0.44	0.60	0.59
ces	0.31	0.48	0.00	0.17	0.31	0.34	0.24	0.63	0.15	0.25	0.32	0.86	0.82	0.12	0.31	0.04	0.30	0.11	0.48	0.55
eng	0.39	0.42	0.17	0.00	0.39	0.11	0.07	0.64	0.22	0.28	0.10	0.84	0.71	0.23	0.10	0.16	0.10	0.06	0.46	0.53
fin	0.62	0.42	0.31	0.39	0.00	0.62	0.47	0.49	0.30	0.40	0.62	0.75	0.73	0.38	0.61	0.29	0.60	0.32	0.61	0.40
fra	0.35	0.58	0.34	0.11	0.62	0.00	0.09	0.75	0.38	0.30	0.03	0.90	0.82	0.35	0.02	0.31	0.02	0.19	0.45	0.65
deu	0.48	0.52	0.24	0.07	0.47	0.09	0.00	0.57	0.35	0.36	0.11	0.82	0.74	0.32	0.10	0.23	0.10	0.14	0.57	0.50
hin	0.88	0.67	0.63	0.64	0.49	0.75	0.57	0.00	0.71	0.79	0.75	0.21	0.71	0.70	0.76	0.65	0.76	0.64	0.89	0.36
isl	0.28	0.42	0.15	0.22	0.30	0.38	0.35	0.71	0.00	0.18	0.34	0.85	0.81	0.16	0.34	0.15	0.32	0.10	0.35	0.64
ind	0.17	0.48	0.25	0.28	0.40	0.30	0.36	0.79	0.18	0.00	0.29	0.87	0.82	0.25	0.27	0.20	0.26	0.22	0.35	0.75
ita	0.32	0.57	0.32	0.10	0.62	0.03	0.11	0.75	0.34	0.29	0.00	0.89	0.82	0.32	0.02	0.31	0.03	0.17	0.45	0.64
jpn	0.92	0.66	0.86	0.84	0.75	0.90	0.82	0.21	0.85	0.87	0.89	0.00	0.68	0.86	0.90	0.87	0.90	0.87	0.95	0.61
kor	0.91	0.48	0.82	0.71	0.73	0.82	0.74	0.71	0.81	0.82	0.82	0.68	0.00	0.83	0.82	0.83	0.82	0.78	0.91	0.55
pol	0.25	0.47	0.12	0.23	0.38	0.35	0.32	0.70	0.16	0.25	0.32	0.86	0.83	0.00	0.31	0.14	0.30	0.17	0.42	0.61
por	0.30	0.58	0.31	0.10	0.61	0.02	0.10	0.76	0.34	0.27	0.02	0.90	0.82	0.31	0.00	0.29	0.01	0.17	0.43	0.66
rus	0.27	0.49	0.04	0.16	0.29	0.31	0.23	0.65	0.15	0.20	0.31	0.87	0.83	0.14	0.29	0.00	0.28	0.10	0.46	0.59
spa	0.29	0.58	0.30	0.10	0.60	0.02	0.10	0.76	0.32	0.26	0.03	0.90	0.82	0.30	0.01	0.28	0.00	0.16	0.43	0.67
swe	0.32	0.44	0.11	0.06	0.32	0.19	0.14	0.64	0.10	0.22	0.17	0.87	0.78	0.17	0.17	0.10	0.16	0.00	0.38	0.55
tha	0.39	0.60	0.48	0.46	0.61	0.45	0.57	0.89	0.35	0.35	0.45	0.95	0.91	0.42	0.43	0.46	0.43	0.38	0.00	0.89
tur	0.85	0.59	0.55	0.53	0.40	0.65	0.50	0.36	0.64	0.75	0.64	0.61	0.55	0.61	0.66	0.59	0.67	0.55	0.89	0.00

Table A.22. Cosine dissimilarity matrix considering head directionality features for all PUD corpora.

Annex 23.

	arb	cmn	ces	eng	fin	fra	deu	hin	isl	ind	ita	jpn	kor	pol	por	rus	spa	swe	tha	tur
arb	0.00	0.04	0.01	0.01	0.02	0.01	0.03	0.06	0.02	0.02	0.01	0.04	0.04	0.01	0.01	0.01	0.00	0.02	0.05	0.03
cmn	0.04	0.00	0.04	0.03	0.03	0.03	0.06	0.09	0.03	0.02	0.04	0.08	0.08	0.04	0.03	0.03	0.04	0.02	0.01	0.07
ces	0.01	0.04	0.00	0.01	0.01	0.00	0.02	0.06	0.01	0.02	0.01	0.04	0.04	0.01	0.01	0.01	0.01	0.02	0.04	0.04
eng	0.01	0.03	0.01	0.00	0.02	0.01	0.03	0.07	0.00	0.00	0.01	0.05	0.05	0.01	0.01	0.01	0.01	0.01	0.03	0.05
fin	0.02	0.03	0.01	0.02	0.00	0.01	0.03	0.06	0.02	0.02	0.02	0.05	0.05	0.02	0.02	0.02	0.02	0.02	0.03	0.05
fra	0.01	0.03	0.00	0.01	0.01	0.00	0.03	0.06	0.01	0.01	0.01	0.04	0.05	0.01	0.01	0.01	0.01	0.01	0.04	0.04
deu	0.03	0.06	0.02	0.03	0.03	0.03	0.00	0.04	0.04	0.04	0.03	0.02	0.02	0.03	0.03	0.03	0.03	0.04	0.06	0.02
hin	0.06	0.09	0.06	0.07	0.06	0.06	0.04	0.00	0.07	0.07	0.06	0.03	0.02	0.06	0.06	0.06	0.06	0.07	0.09	0.03
isl	0.02	0.03	0.01	0.00	0.02	0.01	0.04	0.07	0.00	0.00	0.01	0.05	0.06	0.01	0.01	0.01	0.01	0.00	0.03	0.05
ind	0.02	0.02	0.02	0.00	0.02	0.01	0.04	0.07	0.00	0.00	0.01	0.05	0.06	0.01	0.01	0.01	0.01	0.00	0.03	0.05
ita	0.01	0.04	0.01	0.01	0.02	0.01	0.03	0.06	0.01	0.01	0.00	0.04	0.05	0.00	0.00	0.00	0.00	0.01	0.04	0.04
jpn	0.04	0.08	0.04	0.05	0.05	0.04	0.02	0.03	0.05	0.05	0.04	0.00	0.01	0.04	0.04	0.04	0.04	0.06	0.08	0.01
kor	0.04	0.08	0.04	0.05	0.05	0.05	0.02	0.02	0.06	0.06	0.05	0.01	0.00	0.05	0.05	0.05	0.05	0.06	0.08	0.01
pol	0.01	0.04	0.01	0.01	0.02	0.01	0.03	0.06	0.01	0.01	0.00	0.04	0.05	0.00	0.00	0.00	0.00	0.02	0.04	0.04
por	0.01	0.03	0.01	0.01	0.02	0.01	0.03	0.06	0.01	0.01	0.00	0.04	0.05	0.00	0.00	0.00	0.00	0.01	0.04	0.04
rus	0.01	0.03	0.01	0.01	0.02	0.01	0.03	0.06	0.01	0.01	0.00	0.04	0.05	0.00	0.00	0.00	0.01	0.01	0.04	0.04
spa	0.00	0.04	0.01	0.01	0.02	0.01	0.03	0.06	0.01	0.01	0.00	0.04	0.05	0.00	0.00	0.01	0.00	0.02	0.04	0.04
swe	0.02	0.02	0.02	0.01	0.02	0.01	0.04	0.07	0.00	0.00	0.01	0.06	0.06	0.02	0.01	0.01	0.02	0.00	0.03	0.05
tha	0.05	0.01	0.04	0.03	0.03	0.04	0.06	0.09	0.03	0.03	0.04	0.08	0.08	0.04	0.04	0.04	0.04	0.03	0.00	0.08
tur	0.03	0.07	0.04	0.05	0.05	0.04	0.02	0.03	0.05	0.05	0.04	0.01	0.01	0.04	0.04	0.04	0.04	0.05	0.08	0.00

Table A.23. Euclidean dissimilarity matrix considering OV and VO features for all PUD corpora.

Annex 24.

	arb	cmn	ces	eng	fin	fra	deu	hin	isl	ind	ita	jpn	kor	pol	por	rus	spa	swe	tha	tur
arb	0.00	0.00	0.02	0.00	0.04	0.02	0.38	0.94	0.00	0.00	0.00	0.94	0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.94
cmn	0.00	0.00	0.04	0.00	0.06	0.03	0.42	0.99	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	0.01	0.00	0.00	1.00
ces	0.02	0.04	0.00	0.03	0.00	0.00	0.23	0.73	0.03	0.04	0.02	0.73	0.74	0.02	0.02	0.01	0.01	0.03	0.02	0.74
eng	0.00	0.00	0.03	0.00	0.05	0.02	0.41	0.98	0.00	0.00	0.00	0.98	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.98
fin	0.04	0.06	0.00	0.05	0.00	0.01	0.18	0.66	0.05	0.06	0.04	0.66	0.66	0.03	0.04	0.03	0.03	0.05	0.03	0.66
fra	0.02	0.03	0.00	0.02	0.01	0.00	0.25	0.76	0.02	0.03	0.01	0.77	0.77	0.01	0.01	0.01	0.01	0.02	0.01	0.77
deu	0.38	0.42	0.23	0.41	0.18	0.25	0.00	0.18	0.41	0.42	0.37	0.18	0.18	0.35	0.37	0.35	0.34	0.40	0.36	0.18
hin	0.94	0.99	0.73	0.98	0.66	0.76	0.18	0.00	0.98	1.00	0.93	0.00	0.00	0.90	0.92	0.90	0.89	0.97	0.91	0.00
isl	0.00	0.00	0.03	0.00	0.05	0.02	0.41	0.98	0.00	0.00	0.00	0.98	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.98
ind	0.00	0.00	0.04	0.00	0.06	0.03	0.42	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.01	0.01	0.00	0.00	1.00
ita	0.00	0.00	0.02	0.00	0.04	0.01	0.37	0.93	0.00	0.00	0.00	0.93	0.93	0.00	0.00	0.00	0.00	0.00	0.00	0.93
jpn	0.94	1.00	0.73	0.98	0.66	0.77	0.18	0.00	0.98	1.00	0.93	0.00	0.00	0.91	0.93	0.90	0.89	0.97	0.91	0.00
kor	0.94	1.00	0.74	0.98	0.66	0.77	0.18	0.00	0.98	1.00	0.93	0.00	0.00	0.91	0.93	0.90	0.90	0.97	0.91	0.00
pol	0.00	0.00	0.02	0.00	0.03	0.01	0.35	0.90	0.00	0.00	0.00	0.91	0.91	0.00	0.00	0.00	0.00	0.00	0.00	0.91
por	0.00	0.00	0.02	0.00	0.04	0.01	0.37	0.92	0.00	0.00	0.00	0.93	0.93	0.00	0.00	0.00	0.00	0.00	0.00	0.93
rus	0.00	0.00	0.01	0.00	0.03	0.01	0.35	0.90	0.00	0.01	0.00	0.90	0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.90
spa	0.00	0.01	0.01	0.00	0.03	0.01	0.34	0.89	0.00	0.01	0.00	0.89	0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.90
swe	0.00	0.00	0.03	0.00	0.05	0.02	0.40	0.97	0.00	0.00	0.00	0.97	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.97
tha	0.00	0.00	0.02	0.00	0.03	0.01	0.36	0.91	0.00	0.00	0.00	0.91	0.91	0.00	0.00	0.00	0.00	0.00	0.00	0.91
tur	0.94	1.00	0.74	0.98	0.66	0.77	0.18	0.00	0.98	1.00	0.93	0.00	0.00	0.91	0.93	0.90	0.90	0.97	0.91	0.00

Table A.24. Cosine dissimilarity matrix considering OV and VO features for all PUD corpora.

Annex 25.

	LAS			MLAS		
Language	**PUD monolingual model**	**UDify standard**	**Delta**	**PUD monolingual model**	**UDify standard**	**Delta**
arb	83.34	67.07	16.27	57.66	10.67	46.99
ces	86.80	87.95	-1.15	57.83	77.39	-19.56
cmn	74.84	56.52	18.32	62.73	40.92	21.81
deu	89.55	84.46	5.09	67.00	2.10	64.90
eng	86.63	88.66	-2.03	74.99	75.61	-0.62
fin	82.46	86.58	-4.12	68.26	77.83	-9.57
fra	91.20	82.76	8.44	79.83	25.24	54.59
hin	77.46	58.42	19.04	54.00	3.32	50.68
ind	85.72	56.90	28.82	77.04	7.41	69.63
isl	78.90	-	-	48.79	-	-
ita	89.48	91.76	-2.28	76.27	25.55	50.72
jpn	91.57	93.62	-2.05	82.90	84.86	-1.96
kor	85.99	46.89	39.10	78.23	16.26	61.97
pol	86.88	-		61.31	-	-
por	89.65	80.17	9.48	78.05	17.51	60.54
rus	88.42	87.14	1.28	70.47	37.25	33.22
spa	91.24	83.08	8.16	79.84	18.06	61.78
swe	84.69	86.10	-1.41	69.89	57.12	12.77
tha	74.68	26.06	48.62	63.85	3.77	60.08
tur	76.68	46.07	30.61	56.02	2.61	53.41

Table A.25. Comparison between the LAS and MLAS values obtained in this thesis with UDify models trained with each monolingual PUD corpus and the scores published by Kondratyuk and Straka (2019) concerning the multilingual standard UDify model.

Annex 26.

	LAS				MLAS			
	mean	stdev	delta	p	mean	stdev	delta	p
arb_ces	83.29	0.38	-0.06	0.768	59.69	0.27	2.03	0.000
arb_cmn	83.26	0.18	-0.09	0.501	59.64	0.30	1.98	0.000
arb_deu	83.48	0.21	0.13	0.341	60.10	0.36	2.45	0.000
arb_eng	83.10	0.14	-0.24	0.055	60.07	0.32	2.41	0.000
arb_fin	83.06	0.32	-0.29	0.109	59.52	0.25	1.86	0.000
arb_fra	83.71	0.34	0.36	0.057	60.23	0.41	2.57	0.000
arb_hin	83.03	0.30	-0.32	0.074	59.26	0.20	1.61	0.000
arb_ind	83.69	0.25	0.34	0.049	59.66	0.32	2.00	0.000
arb_isl	83.49	0.18	0.14	0.279	59.55	0.30	1.89	0.000
arb_ita	83.91	0.30	0.57	0.005	60.57	0.43	2.91	0.000
arb_jpn	83.33	0.19	-0.01	0.907	60.08	0.40	2.42	0.000
arb_kor	82.92	0.30	-0.42	0.023	59.28	0.28	1.62	0.000
arb_pol	83.59	0.13	0.24	0.052	59.92	0.16	2.26	0.000
arb_por	83.58	0.17	0.23	0.086	60.07	0.42	2.42	0.000
arb_rus	83.54	0.14	0.19	0.115	60.45	0.23	2.79	0.000
arb_spa	83.83	0.28	0.48	0.010	60.04	0.28	2.38	0.000
arb_swe	83.48	0.27	0.13	0.398	59.75	0.53	2.09	0.000
arb_tha	83.55	0.19	0.21	0.130	59.65	0.31	1.99	0.000
arb_tur	83.50	0.20	0.16	0.246	59.83	0.47	2.18	0.000

Table A.26. UDify dependency parsing results for Arabic language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 27.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
cmn_arb	75.28	0.31	0.53	0.071	63.12	0.27	0.11	0.691
cmn_ces	75.63	0.41	0.87	0.018	63.39	0.43	0.38	0.278
cmn_deu	75.11	0.30	0.35	0.203	62.99	0.29	-0.02	0.933
cmn_eng	75.24	0.35	0.48	0.103	63.44	0.51	0.43	0.212
cmn_fin	74.99	0.28	0.23	0.390	62.50	0.45	-0.52	0.125
cmn_fra	75.94	0.20	1.18	0.001	63.72	0.22	0.71	0.022
cmn_hin	75.52	0.21	0.76	0.011	63.64	0.47	0.63	0.073
cmn_ind	75.76	0.13	1.00	0.004	63.57	0.32	0.56	0.096
cmn_isl	74.98	0.31	0.22	0.416	62.17	0.50	-0.84	0.026
cmn_ita	75.85	0.27	1.09	0.003	63.90	0.24	0.88	0.014
cmn_jpn	75.03	0.14	0.28	0.264	63.25	0.11	0.24	0.364
cmn_kor	75.52	0.27	0.77	0.012	63.22	0.37	0.21	0.477
cmn_pol	75.27	0.22	0.51	0.063	62.76	0.23	-0.25	0.367
cmn_por	75.57	0.34	0.81	0.012	63.59	0.26	0.58	0.057
cmn_rus	75.77	0.23	1.01	0.002	63.63	0.39	0.62	0.060
cmn_spa	75.49	0.18	0.74	0.012	63.65	0.30	0.64	0.043
cmn_swe	75.22	0.25	0.46	0.096	63.00	0.36	-0.02	0.959
cmn_tha	75.16	0.20	0.40	0.128	62.74	0.44	-0.27	0.393
cmn_tur	75.15	0.30	0.39	0.161	63.02	0.44	0.01	0.974

Table A.27. UDify dependency parsing results for Chinese language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 28.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
ces_arb	87.28	0.32	0.48	0.047	56.83	0.43	-1.00	0.003
ces_cmn	86.75	0.08	-0.05	0.775	57.32	0.34	-0.52	0.058
ces_deu	87.24	0.06	0.44	0.025	58.47	0.30	0.63	0.021
ces_eng	87.51	0.21	0.71	0.003	58.67	0.52	0.84	0.016
ces_fin	87.62	0.10	0.82	0.001	58.21	0.40	0.38	0.171
ces_fra	87.92	0.25	1.12	0.000	58.62	0.18	0.78	0.004
ces_hin	87.39	0.10	0.59	0.006	58.50	0.40	0.67	0.025
ces_ind	87.62	0.12	0.82	0.002	58.11	0.21	0.27	0.271
ces_isl	87.94	0.22	1.14	0.000	58.06	0.31	0.23	0.349
ces_ita	88.04	0.29	1.24	0.000	59.26	0.31	1.43	0.000
ces_jpn	86.82	0.30	0.02	0.912	58.60	0.30	0.77	0.008
ces_kor	86.99	0.23	0.19	0.349	58.31	0.51	0.48	0.125
ces_pol	87.76	0.18	0.96	0.000	59.36	0.17	1.53	0.000
ces_por	87.98	0.16	1.18	0.000	59.13	0.35	1.30	0.000
ces_rus	88.38	0.18	1.58	0.000	59.86	0.34	2.03	0.000
ces_spa	87.19	0.19	0.39	0.056	58.57	0.33	0.74	0.011
ces_swe	87.80	0.23	1.00	0.000	59.05	0.47	1.21	0.001
ces_tha	87.85	0.17	1.05	0.000	58.35	0.30	0.52	0.049
ces_tur	87.19	0.21	0.39	0.060	58.35	0.21	0.52	0.036

Table A.28. UDify dependency parsing results for Czech language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 29.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
deu_arb	89.74	0.19	0.19	0.097	66.73	0.61	-0.28	0.336
deu_ces	89.48	0.23	-0.08	0.535	67.56	0.55	0.56	0.053
deu_cmn	89.75	0.21	0.20	0.100	66.60	0.21	-0.41	0.021
deu_eng	89.95	0.13	0.39	0.001	68.41	0.25	1.41	0.000
deu_fin	89.54	0.21	-0.01	0.917	66.88	0.42	-0.12	0.574
deu_fra	89.85	0.18	0.30	0.015	68.21	0.45	1.21	0.000
deu_hin	89.45	0.13	-0.11	0.249	65.95	0.37	-1.06	0.000
deu_ind	89.95	0.20	0.39	0.005	67.62	0.37	0.62	0.010
deu_isl	89.61	0.30	0.05	0.719	65.92	0.68	-1.08	0.005
deu_ita	89.90	0.19	0.35	0.008	68.22	0.42	1.22	0.000
deu_jpn	89.76	0.21	0.21	0.088	67.23	0.44	0.23	0.325
deu_kor	89.92	0.09	0.36	0.001	66.96	0.25	-0.04	0.784
deu_pol	89.83	0.08	0.27	0.005	67.65	0.29	0.64	0.003
deu_por	89.87	0.12	0.31	0.004	68.10	0.36	1.10	0.000
deu_rus	89.86	0.18	0.30	0.014	67.55	0.22	0.54	0.005
deu_spa	89.59	0.09	0.04	0.662	67.42	0.46	0.41	0.095
deu_swe	89.79	0.18	0.23	0.044	67.12	0.29	0.12	0.508
deu_tha	89.72	0.32	0.16	0.294	66.95	0.40	-0.05	0.793
deu_tur	89.62	0.18	0.06	0.547	66.31	0.65	-0.69	0.038

Table A.29. UDify dependency parsing results for German language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 30.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
eng_arb	86.89	0.20	0.26	0.032	75.14	0.18	0.16	0.257
eng_ces	87.07	0.16	0.43	0.001	75.40	0.34	0.41	0.039
eng_cmn	86.90	0.17	0.26	0.015	75.04	0.19	0.06	0.678
eng_deu	87.15	0.12	0.52	0.000	75.57	0.36	0.59	0.009
eng_fin	86.93	0.19	0.30	0.013	75.11	0.29	0.12	0.465
eng_fra	87.57	0.15	0.94	0.000	76.71	0.21	1.73	0.000
eng_hin	87.04	0.18	0.41	0.001	75.49	0.28	0.50	0.009
eng_ind	87.07	0.14	0.44	0.000	74.88	0.39	-0.11	0.573
eng_isl	87.25	0.15	0.62	0.000	75.45	0.30	0.46	0.016
eng_ita	87.40	0.23	0.77	0.000	75.91	0.32	0.93	0.000
eng_jpn	86.99	0.11	0.35	0.001	75.08	0.19	0.10	0.475
eng_kor	87.04	0.14	0.41	0.001	75.31	0.35	0.33	0.096
eng_pol	86.99	0.25	0.36	0.013	75.21	0.46	0.22	0.334
eng_por	87.24	0.18	0.61	0.000	75.67	0.38	0.68	0.005
eng_rus	87.23	0.21	0.60	0.000	76.04	0.43	1.05	0.000
eng_spa	87.02	0.17	0.39	0.002	75.51	0.30	0.52	0.009
eng_swe	86.99	0.19	0.36	0.005	75.37	0.29	0.38	0.036
eng_tha	86.82	0.15	0.19	0.056	74.60	0.29	-0.39	0.035
eng_tur	87.02	0.18	0.39	0.002	75.11	0.23	0.12	0.403

Table A.30. UDify dependency parsing results for English language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 31.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
fin_arb	83.25	0.26	0.79	0.001	68.52	0.47	0.26	0.222
fin_ces	83.80	0.16	1.34	0.000	70.73	0.38	2.47	0.000
fin_cmn	82.87	0.44	0.41	0.081	68.60	0.44	0.34	0.113
fin_deu	83.36	0.25	0.90	0.000	69.16	0.39	0.90	0.000
fin_eng	83.61	0.26	1.15	0.000	69.60	0.29	1.34	0.000
fin_fra	84.04	0.28	1.58	0.000	69.72	0.51	1.46	0.000
fin_hin	83.28	0.47	0.82	0.005	69.97	0.58	1.71	0.000
fin_ind	84.24	0.28	1.78	0.000	69.95	0.21	1.70	0.000
fin_isl	83.67	0.40	1.20	0.000	69.81	0.57	1.55	0.000
fin_ita	83.51	0.46	1.04	0.001	70.29	0.44	2.03	0.000
fin_jpn	82.10	0.15	-0.36	0.019	68.53	0.27	0.28	0.061
fin_kor	83.03	0.13	0.56	0.001	69.54	0.42	1.28	0.000
fin_pol	83.89	0.26	1.43	0.000	70.41	0.34	2.15	0.000
fin_por	83.80	0.35	1.33	0.000	70.11	0.32	1.85	0.000
fin_rus	84.41	0.19	1.95	0.000	70.34	0.47	2.08	0.000
fin_spa	83.45	0.22	0.98	0.000	69.61	0.25	1.35	0.000
fin_swe	83.91	0.30	1.45	0.000	69.67	0.71	1.41	0.001
fin_tha	84.01	0.26	1.55	0.000	69.33	0.36	1.07	0.000
fin_tur	82.65	0.17	0.19	0.185	69.21	0.37	0.95	0.000

Table A.31. UDify dependency parsing results for Finnish language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 32.

	LAS				**MLAS**			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
fra_arb	91.32	0.15	0.13	0.248	80.18	0.18	0.35	0.050
fra_ces	91.22	0.14	0.03	0.801	79.84	0.18	0.01	0.948
fra_cmn	90.99	0.17	-0.21	0.081	79.35	0.24	-0.48	0.019
fra_deu	91.29	0.13	0.09	0.389	80.88	0.32	1.05	0.000
fra_eng	91.41	0.17	0.21	0.080	79.88	0.28	0.05	0.769
fra_fin	90.95	0.13	-0.25	0.033	79.17	0.20	-0.65	0.002
fra_hin	91.20	0.13	0.00	0.982	79.75	0.24	-0.08	0.652
fra_ind	91.51	0.19	0.31	0.023	79.26	0.29	-0.57	0.011
fra_isl	91.11	0.17	-0.08	0.457	78.84	0.23	-0.98	0.000
fra_ita	91.88	0.15	0.68	0.000	81.21	0.30	1.39	0.000
fra_jpn	91.03	0.10	-0.17	0.105	79.60	0.35	-0.23	0.277
fra_kor	91.24	0.11	0.04	0.656	79.30	0.17	-0.53	0.007
fra_pol	91.54	0.15	0.34	0.009	80.18	0.11	0.35	0.039
fra_por	91.85	0.14	0.65	0.000	80.85	0.27	1.03	0.000
fra_rus	91.42	0.18	0.22	0.075	80.41	0.30	0.58	0.010
fra_spa	91.31	0.11	0.12	0.254	80.73	0.30	0.90	0.001
fra_swe	90.96	0.07	-0.24	0.024	78.67	0.29	-1.15	0.000
fra_tha	91.45	0.17	0.26	0.043	79.45	0.25	-0.38	0.054
fra_tur	91.16	0.22	-0.04	0.779	79.37	0.40	-0.46	0.060

Table A.32. UDify dependency parsing results for French language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 33.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
hin_arb	77.28	0.22	-0.17	0.325	54.02	0.25	0.02	0.906
hin_ces	77.40	0.22	-0.05	0.771	54.04	0.52	0.04	0.857
hin_cmn	77.47	0.14	0.01	0.952	53.98	0.32	-0.02	0.921
hin_deu	77.16	0.24	-0.29	0.120	54.14	0.34	0.14	0.444
hin_eng	77.73	0.23	0.27	0.149	54.83	0.30	0.83	0.000
hin_fin	77.22	0.24	-0.23	0.213	53.80	0.22	-0.20	0.173
hin_fra	77.45	0.27	0.00	0.979	54.31	0.34	0.31	0.095
hin_ind	77.62	0.32	0.16	0.452	54.63	0.27	0.63	0.003
hin_isl	77.65	0.23	0.19	0.296	54.00	0.39	0.00	0.993
hin_ita	77.42	0.23	-0.04	0.833	54.47	0.83	0.47	0.210
hin_jpn	76.97	0.35	-0.48	0.037	54.05	0.21	0.05	0.733
hin_kor	77.35	0.29	-0.11	0.571	53.88	0.37	-0.12	0.514
hin_pol	77.59	0.25	0.14	0.459	54.27	0.27	0.27	0.094
hin_por	77.50	0.27	0.05	0.806	54.21	0.68	0.21	0.494
hin_rus	77.21	0.12	-0.25	0.130	54.09	0.28	0.09	0.576
hin_spa	77.47	0.22	0.01	0.956	53.90	0.26	-0.10	0.520
hin_swe	77.94	0.28	0.48	0.026	54.96	0.11	0.96	0.000
hin_tha	77.73	0.38	0.27	0.227	53.97	0.48	-0.03	0.911
hin_tur	77.61	0.13	0.15	0.337	54.54	0.21	0.54	0.002

Table A.33. UDify dependency parsing results for Hindi language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 34.

	LAS				**MLAS**			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
isl_arb	79.51	0.24	0.61	0.000	49.62	0.44	0.83	0.014
isl_ces	79.26	0.20	0.36	0.007	49.69	0.29	0.90	0.004
isl_cmn	79.27	0.23	0.37	0.010	49.08	0.31	0.29	0.266
isl_deu	79.55	0.18	0.65	0.000	50.12	0.39	1.33	0.001
isl_eng	79.45	0.23	0.54	0.001	48.92	0.19	0.14	0.564
isl_fin	79.13	0.17	0.23	0.038	48.18	0.42	-0.61	0.048
isl_fra	80.48	0.23	1.58	0.000	48.20	0.31	-0.59	0.037
isl_hin	79.22	0.23	0.32	0.018	49.95	0.42	1.16	0.002
isl_ind	80.24	0.28	1.34	0.000	49.22	0.41	0.43	0.146
isl_ita	80.42	0.10	1.52	0.000	50.71	0.22	1.92	0.000
isl_jpn	78.83	0.32	-0.07	0.654	49.29	0.53	0.50	0.132
isl_kor	79.16	0.34	0.25	0.136	49.36	0.49	0.57	0.080
isl_pol	79.57	0.35	0.67	0.002	50.27	0.43	1.48	0.000
isl_por	80.44	0.09	1.53	0.000	49.53	0.38	0.74	0.019
isl_rus	79.86	0.18	0.96	0.000	50.65	0.29	1.86	0.000
isl_spa	79.82	0.22	0.92	0.000	49.62	0.47	0.83	0.016
isl_swe	78.91	0.20	0.00	0.972	49.67	0.29	0.88	0.005
isl_tha	79.43	0.25	0.53	0.002	49.19	0.23	0.40	0.112
isl_tur	79.45	0.22	0.55	0.001	49.01	0.30	0.22	0.390

Table A.34. UDify dependency parsing results for Icelandic language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 35.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
ind_arb	85.83	0.20	0.10	0.374	77.25	0.26	0.22	0.246
ind_ces	85.47	0.08	-0.26	0.010	76.76	0.15	-0.27	0.105
ind_cmn	85.48	0.14	-0.24	0.029	77.09	0.38	0.06	0.785
ind_deu	85.34	0.15	-0.39	0.003	76.95	0.27	-0.08	0.655
ind_eng	86.02	0.42	0.30	0.141	77.67	0.57	0.63	0.042
ind_fin	85.66	0.17	-0.06	0.557	76.58	0.28	-0.45	0.031
ind_fra	85.57	0.22	-0.15	0.220	76.64	0.42	-0.39	0.106
ind_hin	86.13	0.25	0.41	0.009	77.74	0.23	0.70	0.002
ind_isl	85.90	0.22	0.18	0.168	76.64	0.36	-0.40	0.075
ind_ita	86.05	0.16	0.33	0.013	77.61	0.28	0.58	0.015
ind_jpn	85.93	0.14	0.21	0.054	77.69	0.21	0.65	0.003
ind_kor	85.52	0.09	-0.21	0.035	76.66	0.18	-0.37	0.039
ind_pol	85.37	0.21	-0.35	0.013	76.89	0.24	-0.14	0.420
ind_por	86.08	0.25	0.35	0.019	77.62	0.46	0.59	0.030
ind_rus	85.73	0.29	0.00	0.985	77.21	0.29	0.17	0.368
ind_spa	85.79	0.16	0.06	0.539	77.42	0.32	0.38	0.076
ind_swe	85.88	0.14	0.16	0.124	77.27	0.36	0.23	0.285
ind_tha	85.95	0.28	0.22	0.134	77.32	0.24	0.28	0.129
ind_tur	85.75	0.24	0.02	0.846	76.53	0.37	-0.51	0.033

Table A.35. UDify dependency parsing results for Indonesian language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 36.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
ita_arb	89.52	0.12	0.05	0.528	76.01	0.24	-0.26	0.171
ita_ces	90.02	0.24	0.54	0.001	77.33	0.30	1.06	0.000
ita_cmn	89.67	0.16	0.19	0.050	76.37	0.30	0.10	0.599
ita_deu	89.51	0.16	0.03	0.730	76.98	0.25	0.71	0.002
ita_eng	90.19	0.13	0.71	0.000	77.18	0.25	0.91	0.000
ita_fin	89.75	0.21	0.28	0.023	76.46	0.17	0.19	0.264
ita_fra	90.25	0.14	0.78	0.000	77.21	0.31	0.94	0.001
ita_hin	89.76	0.19	0.29	0.013	76.96	0.33	0.69	0.006
ita_ind	90.02	0.12	0.54	0.000	76.72	0.29	0.45	0.036
ita_isl	90.05	0.26	0.57	0.001	76.52	0.36	0.25	0.255
ita_jpn	89.66	0.16	0.18	0.060	76.58	0.46	0.31	0.224
ita_kor	89.83	0.20	0.35	0.005	76.91	0.34	0.65	0.009
ita_pol	89.71	0.21	0.23	0.048	77.33	0.36	1.06	0.000
ita_por	89.77	0.17	0.29	0.008	77.33	0.34	1.06	0.000
ita_rus	90.08	0.15	0.60	0.000	77.02	0.30	0.76	0.003
ita_spa	89.38	0.19	-0.10	0.337	76.84	0.51	0.57	0.046
ita_swe	89.78	0.17	0.31	0.006	76.48	0.24	0.21	0.253
ita_tha	89.52	0.22	0.05	0.652	75.96	0.33	-0.31	0.147
ita_tur	89.77	0.15	0.29	0.005	76.51	0.35	0.24	0.259

Table A.36. UDify dependency parsing results for Italian language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 37.

	LAS				**MLAS**			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
jpn_arb	91.22	0.15	-0.35	0.007	81.92	0.23	-0.97	0.000
jpn_ces	91.25	0.28	-0.32	0.053	82.02	0.50	-0.88	0.008
jpn_cmn	91.09	0.15	-0.48	0.001	81.99	0.19	-0.91	0.000
jpn_deu	91.34	0.17	-0.23	0.060	82.22	0.46	-0.68	0.017
jpn_eng	91.25	0.29	-0.31	0.055	82.30	0.41	-0.60	0.023
jpn_fin	91.14	0.16	-0.43	0.002	81.86	0.35	-1.04	0.000
jpn_fra	91.47	0.11	-0.10	0.321	82.29	0.30	-0.61	0.010
jpn_hin	91.27	0.09	-0.30	0.008	82.23	0.22	-0.67	0.003
jpn_ind	91.55	0.15	-0.01	0.926	82.54	0.49	-0.35	0.202
jpn_isl	91.17	0.04	-0.40	0.001	81.83	0.25	-1.07	0.000
jpn_ita	91.44	0.10	-0.13	0.193	82.41	0.29	-0.48	0.028
jpn_kor	91.41	0.16	-0.15	0.175	82.25	0.30	-0.65	0.008
jpn_pol	91.35	0.18	-0.21	0.082	82.25	0.28	-0.65	0.006
jpn_por	91.40	0.17	-0.17	0.154	82.27	0.30	-0.63	0.008
jpn_rus	91.30	0.17	-0.26	0.033	82.24	0.26	-0.66	0.005
jpn_spa	91.57	0.11	0.00	0.973	82.65	0.27	-0.25	0.207
jpn_swe	91.22	0.29	-0.34	0.040	81.87	0.48	-1.02	0.002
jpn_tha	91.62	0.21	0.06	0.645	82.55	0.45	-0.35	0.169
jpn_tur	91.25	0.11	-0.31	0.007	82.54	0.25	-0.36	0.073

Table A.37. UDify dependency parsing results for Italian language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 38.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
kor_arb	85.44	0.22	-0.55	0.001	77.77	0.30	-0.45	0.006
kor_ces	85.51	0.15	-0.48	0.001	77.56	0.29	-0.67	0.000
kor_cmn	85.26	0.19	-0.73	0.000	77.11	0.30	-1.12	0.000
kor_deu	85.21	0.20	-0.78	0.000	77.03	0.25	-1.19	0.000
kor_eng	85.32	0.26	-0.68	0.000	77.36	0.32	-0.87	0.000
kor_fin	85.34	0.18	-0.65	0.000	77.33	0.22	-0.90	0.000
kor_fra	85.42	0.16	-0.57	0.000	77.71	0.28	-0.51	0.002
kor_hin	85.76	0.32	-0.23	0.163	77.87	0.49	-0.36	0.111
kor_ind	85.25	0.23	-0.74	0.000	77.63	0.35	-0.59	0.003
kor_isl	85.66	0.19	-0.33	0.020	77.65	0.34	-0.58	0.004
kor_ita	85.79	0.30	-0.20	0.189	77.79	0.26	-0.44	0.004
kor_jpn	85.17	0.36	-0.82	0.001	77.90	0.33	-0.32	0.048
kor_pol	85.39	0.19	-0.60	0.000	77.42	0.14	-0.81	0.000
kor_por	85.35	0.29	-0.64	0.001	77.67	0.35	-0.56	0.004
kor_rus	85.77	0.25	-0.22	0.123	77.78	0.31	-0.44	0.008
kor_spa	85.43	0.12	-0.56	0.000	77.59	0.27	-0.63	0.000
kor_swe	85.49	0.26	-0.50	0.004	77.79	0.21	-0.44	0.001
kor_tha	85.74	0.17	-0.26	0.037	77.49	0.27	-0.73	0.000
kor_tur	34.02	0.44	-51.97	0.000	12.25	0.68	-65.97	0.000

Table A.38. UDify dependency parsing results for Korean language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 39.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
pol_arb	87.46	0.23	0.58	0.001	62.39	0.33	1.08	0.001
pol_ces	87.62	0.27	0.74	0.000	62.66	0.39	1.36	0.000
pol_cmn	86.80	0.13	-0.08	0.458	61.31	0.30	0.00	0.992
pol_deu	87.07	0.24	0.19	0.173	62.59	0.67	1.29	0.004
pol_eng	87.08	0.31	0.20	0.219	62.21	0.45	0.90	0.008
pol_fin	87.11	0.17	0.24	0.058	60.83	0.36	-0.48	0.089
pol_fra	88.54	0.21	1.66	0.000	63.55	0.21	2.24	0.000
pol_hin	86.97	0.16	0.09	0.428	61.50	0.23	0.20	0.411
pol_ind	87.52	0.14	0.64	0.000	62.18	0.35	0.87	0.010
pol_isl	86.87	0.30	-0.01	0.958	61.10	0.46	-0.21	0.466
pol_ita	87.28	0.16	0.40	0.004	62.22	0.28	0.91	0.003
pol_jpn	87.22	0.37	0.35	0.072	62.21	0.34	0.91	0.005
pol_kor	87.12	0.12	0.24	0.032	61.63	0.24	0.33	0.185
pol_por	87.66	0.34	0.78	0.001	62.64	0.64	1.33	0.003
pol_rus	87.91	0.18	1.03	0.000	62.52	0.48	1.21	0.002
pol_spa	87.29	0.19	0.42	0.005	62.58	0.36	1.27	0.001
pol_swe	87.31	0.27	0.43	0.010	62.24	0.39	0.94	0.005
pol_tha	87.42	0.14	0.54	0.000	61.53	0.31	0.23	0.371
pol_tur	86.81	0.12	-0.07	0.476	61.79	0.34	0.48	0.085

Table A.39. UDify dependency parsing results for Polish language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 40.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
por_arb	90.30	0.14	0.64	0.000	78.40	0.21	0.35	0.090
por_ces	90.13	0.23	0.47	0.002	78.54	0.35	0.49	0.047
por_cmn	89.98	0.15	0.33	0.005	78.30	0.34	0.25	0.274
por_deu	89.61	0.14	-0.04	0.667	78.74	0.18	0.69	0.003
por_eng	89.90	0.24	0.24	0.065	78.25	0.40	0.20	0.404
por_fin	89.80	0.19	0.14	0.179	78.06	0.27	0.01	0.955
por_fra	90.78	0.14	1.12	0.000	79.39	0.29	1.34	0.000
por_hin	89.85	0.20	0.20	0.092	78.09	0.40	0.04	0.850
por_ind	90.27	0.08	0.62	0.000	78.65	0.35	0.60	0.021
por_isl	90.18	0.17	0.53	0.000	78.16	0.19	0.11	0.550
por_ita	90.44	0.16	0.78	0.000	79.26	0.21	1.21	0.000
por_jpn	89.63	0.14	-0.03	0.750	77.76	0.32	-0.29	0.200
por_kor	89.96	0.29	0.31	0.043	77.95	0.36	-0.10	0.668
por_pol	90.01	0.19	0.35	0.006	78.14	0.21	0.09	0.651
por_rus	89.84	0.07	0.19	0.024	78.52	0.34	0.47	0.052
por_spa	91.00	0.11	1.35	0.000	80.34	0.09	2.29	0.000
por_swe	89.71	0.10	0.06	0.458	78.31	0.17	0.26	0.183
por_tha	90.14	0.18	0.48	0.001	78.67	0.32	0.62	0.015
por_tur	89.96	0.17	0.30	0.010	77.87	0.25	-0.18	0.378

Table A.40. UDify dependency parsing results for Portuguese language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 41.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
rus_arb	89.03	0.21	0.61	0.000	73.47	0.42	3.00	0.000
rus_ces	89.08	0.10	0.66	0.000	73.71	0.29	3.23	0.000
rus_cmn	88.94	0.32	0.52	0.005	73.26	0.25	2.79	0.000
rus_deu	88.24	0.16	-0.18	0.076	70.44	0.39	-0.04	0.845
rus_eng	89.07	0.18	0.65	0.000	73.82	0.18	3.35	0.000
rus_fin	88.75	0.28	0.33	0.027	72.45	0.49	1.98	0.000
rus_fra	89.21	0.24	0.80	0.000	74.59	0.41	4.11	0.000
rus_hin	88.50	0.14	0.09	0.308	73.58	0.27	3.10	0.000
rus_ind	89.08	0.16	0.67	0.000	73.83	0.33	3.35	0.000
rus_isl	89.61	0.15	1.19	0.000	73.76	0.40	3.28	0.000
rus_ita	89.26	0.21	0.84	0.000	74.58	0.37	4.10	0.000
rus_jpn	88.55	0.17	0.14	0.159	73.39	0.26	2.92	0.000
rus_kor	89.11	0.18	0.69	0.000	73.30	0.22	2.83	0.000
rus_pol	88.99	0.21	0.58	0.000	74.39	0.37	3.92	0.000
rus_por	89.21	0.18	0.79	0.000	74.30	0.26	3.82	0.000
rus_spa	88.91	0.16	0.50	0.000	74.48	0.32	4.01	0.000
rus_swe	88.88	0.26	0.47	0.003	73.77	0.35	3.29	0.000
rus_tha	88.85	0.12	0.43	0.000	73.76	0.11	3.29	0.000
rus_tur	89.37	0.22	0.95	0.000	73.75	0.35	3.28	0.000

Table A.41. UDify dependency parsing results for Russian language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 42.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
spa_arb	91.42	0.12	0.18	0.016	79.79	0.22	-0.05	0.669
spa_ces	91.55	0.16	0.31	0.002	80.83	0.36	0.99	0.000
spa_cmn	91.39	0.16	0.14	0.083	79.97	0.28	0.13	0.377
spa_deu	91.25	0.13	0.00	0.947	79.95	0.21	0.11	0.351
spa_eng	91.71	0.12	0.47	0.000	80.68	0.20	0.84	0.000
spa_fin	91.19	0.17	-0.05	0.533	80.45	0.46	0.61	0.013
spa_fra	91.58	0.14	0.34	0.001	80.98	0.22	1.14	0.000
spa_hin	91.45	0.12	0.21	0.007	80.41	0.24	0.57	0.001
spa_ind	91.77	0.20	0.53	0.000	80.29	0.47	0.45	0.056
spa_isl	91.31	0.09	0.07	0.214	80.11	0.07	0.27	0.009
spa_ita	91.46	0.24	0.22	0.061	80.68	0.29	0.84	0.000
spa_jpn	91.15	0.14	-0.09	0.238	79.72	0.35	-0.12	0.482
spa_kor	91.57	0.05	0.33	0.000	80.57	0.15	0.74	0.000
spa_pol	91.84	0.07	0.60	0.000	80.89	0.20	1.05	0.000
spa_por	92.10	0.15	0.86	0.000	81.24	0.22	1.40	0.000
spa_rus	91.67	0.18	0.43	0.000	80.44	0.34	0.60	0.003
spa_swe	91.58	0.17	0.34	0.002	80.03	0.34	0.19	0.259
spa_tha	91.54	0.13	0.29	0.001	79.68	0.16	-0.16	0.147
spa_tur	91.35	0.20	0.11	0.249	80.20	0.35	0.36	0.048

Table A.42. UDify dependency parsing results for Spanish language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 43.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
swe_arb	84.70	0.25	0.00	0.980	69.06	0.29	-0.82	0.003
swe_ces	85.27	0.22	0.58	0.002	70.93	0.44	1.04	0.002
swe_cmn	84.66	0.16	-0.04	0.780	69.76	0.36	-0.13	0.582
swe_deu	85.28	0.14	0.59	0.001	69.68	0.33	-0.20	0.368
swe_eng	85.99	0.17	1.30	0.000	71.25	0.20	1.37	0.000
swe_fin	85.24	0.16	0.55	0.001	70.44	0.22	0.55	0.017
swe_fra	85.68	0.24	0.99	0.000	71.50	0.37	1.61	0.000
swe_hin	84.89	0.25	0.20	0.214	69.82	0.34	-0.06	0.776
swe_ind	85.56	0.27	0.87	0.000	71.06	0.22	1.18	0.000
swe_isl	85.35	0.13	0.66	0.000	70.29	0.39	0.40	0.116
swe_ita	85.63	0.15	0.94	0.000	70.43	0.33	0.54	0.032
swe_jpn	85.02	0.22	0.32	0.041	70.07	0.51	0.18	0.514
swe_kor	85.21	0.18	0.52	0.003	70.33	0.36	0.44	0.079
swe_pol	85.28	0.16	0.59	0.001	70.80	0.34	0.91	0.002
swe_por	85.55	0.38	0.86	0.001	70.57	0.57	0.68	0.040
swe_rus	85.98	0.08	1.29	0.000	71.21	0.17	1.32	0.000
swe_spa	85.51	0.15	0.82	0.000	70.26	0.39	0.37	0.139
swe_tha	85.14	0.21	0.45	0.007	70.62	0.35	0.73	0.009
swe_tur	84.50	0.20	-0.19	0.184	69.53	0.36	-0.36	0.143

Table A.43. UDify dependency parsing results for Swedish language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 44.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
tha_arb	74.59	0.36	0.03	0.833	63.95	0.41	0.09	0.636
tha_ces	74.41	0.26	-0.15	0.255	63.54	0.42	-0.31	0.157
tha_cmn	74.15	0.39	-0.41	0.033	62.95	0.47	-0.90	0.002
tha_deu	74.24	0.31	-0.32	0.041	63.40	0.35	-0.46	0.025
tha_eng	74.43	0.23	-0.13	0.248	63.39	0.49	-0.46	0.062
tha_fin	73.91	0.41	-0.65	0.004	63.02	0.59	-0.83	0.009
tha_fra	74.84	0.18	0.29	0.009	64.01	0.41	0.16	0.440
tha_hin	73.89	0.28	-0.67	0.000	63.28	0.29	-0.57	0.004
tha_ind	74.57	0.15	0.01	0.891	63.71	0.25	-0.14	0.352
tha_isl	74.46	0.26	-0.09	0.446	63.68	0.32	-0.18	0.304
tha_ita	74.73	0.28	0.18	0.196	63.79	0.36	-0.06	0.755
tha_jpn	73.76	0.35	-0.79	0.000	62.78	0.45	-1.07	0.000
tha_kor	73.64	0.27	-0.92	0.000	62.47	0.39	-1.38	0.000
tha_pol	74.64	0.19	0.08	0.400	63.80	0.18	-0.05	0.697
tha_por	74.32	0.44	-0.24	0.235	63.21	0.39	-0.64	0.007
tha_rus	74.27	0.16	-0.29	0.006	63.54	0.22	-0.31	0.040
tha_spa	74.83	0.18	0.27	0.012	64.01	0.24	0.16	0.279
tha_swe	74.58	0.27	0.02	0.879	63.60	0.43	-0.25	0.242
tha_tur	73.83	0.23	-0.72	0.000	62.65	0.29	-1.21	0.000

Table A.44. UDify dependency parsing results for Thai language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 45.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
tur_arb	76.36	0.11	-0.31	0.010	57.03	0.49	1.00	0.003
tur_ces	76.54	0.12	-0.14	0.233	57.74	0.07	1.71	0.000
tur_cmn	76.62	0.16	-0.05	0.625	57.63	0.33	1.61	0.000
tur_deu	77.14	0.20	0.46	0.003	58.42	0.37	2.40	0.000
tur_eng	76.74	0.27	0.07	0.647	57.93	0.37	1.90	0.000
tur_fin	76.67	0.20	-0.01	0.940	58.10	0.42	2.08	0.000
tur_fra	76.37	0.16	-0.30	0.019	57.91	0.23	1.89	0.000
tur_hin	77.11	0.19	0.44	0.004	57.96	0.30	1.93	0.000
tur_ind	76.40	0.28	-0.28	0.096	56.96	0.52	0.94	0.007
tur_isl	76.57	0.41	-0.11	0.591	57.23	0.39	1.21	0.000
tur_ita	76.50	0.15	-0.17	0.158	57.96	0.35	1.93	0.000
tur_jpn	76.46	0.29	-0.22	0.169	57.68	0.57	1.66	0.000
tur_kor	77.76	0.34	1.08	0.000	58.68	0.43	2.65	0.000
tur_pol	76.73	0.18	0.06	0.630	58.35	0.28	2.32	0.000
tur_por	76.67	0.31	0.00	0.978	57.79	0.38	1.77	0.000
tur_rus	77.25	0.22	0.58	0.001	58.73	0.38	2.70	0.000
tur_spa	76.62	0.17	-0.06	0.601	57.86	0.29	1.84	0.000
tur_swe	76.71	0.27	0.03	0.831	58.46	0.44	2.44	0.000
tur_tha	76.39	0.31	-0.28	0.098	57.10	0.50	1.08	0.002
tur_tur	77.67	0.46	1.00	0.001	58.58	0.55	2.55	0.000

Table A.45. UDify dependency parsing results for Turkish language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 46.

Language	LAS	Std. Dev.	Delta (UDify - UDpipe)	MLAS	Std. Dev.	Delta (UDify - UDpipe)
arb	83.34	0.24	29.97	57.66	0.72	33.29
ces	86.80	0.40	26.90	57.83	0.48	26.00
cmn	74.84	0.56	22.60	62.73	0.73	24.09
deu	89.55	0.17	27.04	67.00	0.29	38.16
eng	86.63	0.15	21.35	74.99	0.26	28.13
fin	82.46	0.28	29.52	68.26	0.17	26.72
fra	91.20	0.21	17.68	79.83	0.34	27.78
hin	77.46	0.35	15.43	54.00	0.25	15.05
ind	85.72	0.19	25.40	77.04	0.34	30.37
isl	78.90	0.16	22.32	48.79	0.52	14.52
ita	89.48	0.14	20.00	76.27	0.35	30.35
jpn	91.57	0.20	5.69	82.90	0.36	11.69
kor	85.99	0.20	20.06	78.23	0.13	28.88
pol	86.88	0.21	28.64	61.31	0.51	29.70
por	89.65	0.16	18.89	78.05	0.40	26.39
rus	88.42	0.15	23.78	70.47	0.20	28.48
spa	91.24	0.09	21.44	79.84	0.19	30.14
swe	84.69	0.26	20.45	69.89	0.00	22.74
tha	74.68	0.13	16.61	63.85	0.00	18.54
tur	76.68	0.21	25.22	56.02	0.38	23.51

Table A.46. UDPipe 1.0 LAS and MLAS scores regarding monolingual parsing models in comparison with UDify tool.

Annex 47.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
arb_ces	52.09	0.98	-1.28	0.059	22.32	0.39	-2.05	0.000
arb_cmn	43.66	0.94	-9.71	0.000	19.08	1.35	-5.28	0.000
arb_deu	51.96	0.63	-1.41	0.013	22.69	0.71	-1.68	0.004
arb_eng	46.61	0.55	-6.76	0.000	20.05	0.21	-4.32	0.000
arb_fin	51.83	0.92	-1.53	0.027	22.82	0.52	-1.54	0.002
arb_fra	53.28	0.33	-0.09	0.777	24.00	0.11	-0.37	0.021
arb_hin	45.06	1.05	-8.31	0.000	19.01	0.38	-5.36	0.000
arb_ind	51.67	0.63	-1.70	0.006	20.84	0.28	-3.52	0.000
arb_isl	52.31	0.40	-1.05	0.017	23.44	0.10	-0.93	0.000
arb_ita	52.14	0.61	-1.23	0.021	21.89	0.52	-2.48	0.000
arb_jpn	41.45	1.36	-11.92	0.000	19.65	1.18	-4.72	0.000
arb_kor	45.73	0.75	-7.64	0.000	19.90	0.80	-4.47	0.000
arb_pol	52.50	0.71	-0.87	0.093	23.77	0.46	-0.60	0.056
arb_por	52.71	0.81	-0.66	0.216	23.98	0.52	-0.39	0.220
arb_rus	51.74	0.65	-1.63	0.008	23.15	0.18	-1.21	0.000
arb_spa	52.02	0.59	-1.35	0.013	20.95	0.60	-3.42	0.000
arb_swe	51.75	0.47	-1.62	0.003	22.63	0.12	-1.73	0.000
arb_tha	48.93	0.61	-4.44	0.000	21.38	0.13	-2.98	0.000
arb_tur	41.00	1.08	-12.37	0.000	19.05	1.12	-5.32	0.000

Table A.47. UDPipe 1.0 dependency parsing results for Arabic language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 48.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
cmn_arb	57.34	0.28	5.09	0.000	46.77	0.44	8.13	0.000
cmn_ces	57.25	0.91	5.01	0.000	46.93	1.11	8.29	0.000
cmn_deu	56.38	0.15	4.13	0.000	46.54	0.15	7.90	0.000
cmn_eng	57.04	0.85	4.80	0.000	46.55	0.87	7.91	0.000
cmn_fin	58.07	0.78	5.83	0.000	47.79	0.68	9.15	0.000
cmn_fra	57.13	0.57	4.89	0.000	45.70	0.61	7.06	0.000
cmn_hin	57.25	0.28	5.01	0.000	47.19	0.15	8.55	0.000
cmn_ind	56.94	0.44	4.70	0.000	47.09	0.51	8.45	0.000
cmn_isl	56.92	0.40	4.68	0.000	46.29	0.48	7.65	0.000
cmn_ita	56.93	0.56	4.68	0.000	46.79	0.58	8.15	0.000
cmn_jpn	57.12	0.14	4.88	0.000	47.14	0.39	8.50	0.000
cmn_kor	57.91	0.58	5.66	0.000	48.04	0.57	9.40	0.000
cmn_pol	19.44	1.77	-32.80	0.000	11.59	1.19	-27.05	0.000
cmn_por	50.56	0.53	-1.68	0.003	37.44	0.72	-1.20	0.031
cmn_rus	58.31	0.13	6.06	0.000	48.60	0.32	9.96	0.000
cmn_spa	56.61	0.38	4.36	0.000	46.61	0.23	7.97	0.000
cmn_swe	55.92	0.48	3.68	0.000	45.51	0.41	6.87	0.000
cmn_tha	57.21	0.71	4.97	0.000	47.30	0.92	8.66	0.000
cmn_tur	57.42	0.68	5.18	0.000	46.64	0.61	8.00	0.000

Table A.48. UDPipe 1.0 dependency parsing results for Chinese language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 49.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
ces_arb	63.44	0.54	3.53	0.000	38.02	0.64	6.19	0.000
ces_cmn	64.36	0.32	4.45	0.000	37.92	0.24	6.08	0.000
ces_deu	61.80	0.48	1.89	0.002	35.35	0.50	3.52	0.000
ces_eng	59.90	0.35	-0.01	0.983	35.63	0.17	3.79	0.000
ces_fin	61.36	0.69	1.45	0.018	36.28	0.62	4.45	0.000
ces_fra	59.75	0.88	-0.16	0.778	32.97	0.82	1.13	0.074
ces_hin	62.81	0.27	2.90	0.000	37.09	0.15	5.26	0.000
ces_ind	62.47	0.63	2.56	0.001	35.76	0.39	3.93	0.000
ces_isl	62.38	0.67	2.48	0.001	33.96	0.37	2.13	0.001
ces_ita	61.86	0.58	1.95	0.003	34.43	0.50	2.60	0.001
ces_jpn	62.55	0.64	2.64	0.001	36.10	0.43	4.27	0.000
ces_kor	61.75	0.28	1.85	0.001	36.75	0.18	4.91	0.000
ces_pol	64.24	0.44	4.33	0.000	36.85	0.23	5.01	0.000
ces_por	59.84	0.65	-0.07	0.878	34.29	0.51	2.46	0.001
ces_rus	63.67	0.41	3.77	0.000	38.01	0.41	6.17	0.000
ces_spa	59.37	0.37	-0.53	0.172	33.89	0.26	2.05	0.001
ces_swe	62.66	0.53	2.76	0.000	35.69	0.26	3.85	0.000
ces_tha	63.54	0.75	3.64	0.000	36.36	0.67	4.53	0.000
ces_tur	60.21	0.78	0.31	0.550	35.95	0.23	4.11	0.000

Table A.49. UDPipe 1.0 dependency parsing results for Czech language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 50.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
deu_arb	61.81	0.82	-0.70	0.258	28.53	0.90	-0.32	0.533
deu_cmn	61.37	0.39	-1.14	0.037	28.23	0.64	-0.62	0.143
deu_ces	60.70	0.90	-1.81	0.022	27.79	0.38	-1.06	0.007
deu_eng	59.26	0.52	-3.25	0.000	25.87	0.91	-2.97	0.001
deu_fin	60.80	0.59	-1.72	0.012	27.53	0.46	-1.32	0.004
deu_fra	59.37	0.47	-3.14	0.000	26.01	0.23	-2.84	0.000
deu_hin	61.45	1.09	-1.07	0.160	27.70	0.74	-1.15	0.032
deu_ind	60.03	0.60	-2.48	0.002	27.09	0.43	-1.75	0.001
deu_isl	61.73	0.81	-0.78	0.210	28.43	0.50	-0.41	0.228
deu_ita	60.07	0.90	-2.45	0.006	26.26	0.33	-2.59	0.000
deu_jpn	61.07	1.17	-1.44	0.084	26.32	0.91	-2.52	0.002
deu_kor	61.22	0.70	-1.30	0.046	27.94	0.76	-0.91	0.075
deu_pol	61.71	0.22	-0.81	0.088	28.53	0.20	-0.31	0.179
deu_por	59.99	0.48	-2.52	0.001	27.28	0.59	-1.57	0.004
deu_rus	61.06	0.42	-1.45	0.016	28.68	0.48	-0.16	0.606
deu_spa	59.95	0.47	-2.56	0.001	26.99	0.50	-1.85	0.001
deu_swe	60.26	1.06	-2.26	0.014	26.50	0.32	-2.34	0.000
deu_tha	61.31	0.85	-1.20	0.080	27.30	0.48	-1.55	0.002
deu_tur	59.87	1.01	-2.64	0.006	26.71	0.64	-2.13	0.001

Table A.50. UDPipe 1.0 dependency parsing results for German language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 51.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
eng_arb	70.76	0.11	5.49	0.000	55.96	0.22	9.11	0.000
eng_cmn	70.88	0.84	5.61	0.000	56.72	0.85	9.87	0.000
eng_ces	69.82	0.37	4.55	0.000	55.39	0.36	8.54	0.000
eng_deu	68.28	0.21	3.00	0.000	53.27	0.36	6.42	0.000
eng_fin	69.02	0.26	3.74	0.000	54.36	0.58	7.50	0.000
eng_fra	69.02	0.24	3.74	0.000	54.07	0.55	7.22	0.000
eng_hin	70.99	0.56	5.71	0.000	56.50	0.55	9.65	0.000
eng_ind	70.66	0.74	5.38	0.000	56.14	0.87	9.29	0.000
eng_isl	71.59	0.46	6.31	0.000	57.02	0.57	10.17	0.000
eng_ita	69.88	0.21	4.60	0.000	55.03	0.22	8.17	0.000
eng_jpn	70.78	0.81	5.50	0.000	55.55	0.68	8.69	0.000
eng_kor	70.62	0.61	5.34	0.000	56.04	0.70	9.19	0.000
eng_pol	69.91	0.23	4.64	0.000	55.50	0.19	8.65	0.000
eng_por	69.20	0.46	3.92	0.000	53.59	0.46	6.73	0.000
eng_rus	71.36	0.20	6.08	0.000	57.51	0.36	10.66	0.000
eng_spa	69.32	0.65	4.05	0.000	54.54	0.66	7.69	0.000
eng_swe	69.72	0.52	4.44	0.000	54.79	0.57	7.93	0.000
eng_tha	69.84	0.34	4.56	0.000	55.17	0.34	8.32	0.000
eng_tur	68.76	0.36	3.48	0.000	53.77	0.60	6.92	0.000

Table A.51. UDPipe 1.0 dependency parsing results for English language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 52.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
fin_arb	53.53	1.09	0.59	0.410	42.35	0.85	0.81	0.263
fin_cmn	52.60	0.85	-0.34	0.571	41.20	0.82	-0.33	0.629
fin_ces	52.13	0.94	-0.81	0.227	40.93	0.90	-0.61	0.404
fin_deu	51.63	0.40	-1.31	0.022	39.80	0.67	-1.73	0.029
fin_eng	51.79	0.53	-1.15	0.047	40.32	0.46	-1.22	0.071
fin_fra	52.31	0.94	-0.63	0.336	40.14	1.46	-1.40	0.167
fin_hin	52.92	1.42	-0.02	0.983	41.96	1.07	0.42	0.587
fin_ind	50.49	0.68	-2.45	0.003	39.33	0.86	-2.20	0.016
fin_isl	53.46	0.67	0.52	0.340	40.45	0.20	-1.09	0.080
fin_ita	52.09	1.56	-0.85	0.365	39.33	1.23	-2.21	0.032
fin_jpn	52.82	0.75	-0.12	0.830	40.69	0.21	-0.84	0.155
fin_kor	53.32	0.98	0.38	0.566	42.30	0.77	0.77	0.274
fin_pol	52.97	1.40	0.03	0.971	41.24	1.13	-0.30	0.709
fin_por	52.50	0.87	-0.44	0.479	40.73	0.62	-0.80	0.226
fin_rus	53.56	0.71	0.62	0.281	42.22	0.79	0.69	0.328
fin_spa	51.18	0.41	-1.76	0.007	39.24	0.21	-2.30	0.004
fin_swe	51.52	0.62	-1.42	0.028	39.31	1.18	-2.23	0.028
fin_tha	53.02	0.64	0.08	0.874	40.93	0.49	-0.60	0.327
fin_tur	51.73	0.62	-1.21	0.049	40.42	0.71	-1.12	0.121

Table A.52. UDPipe 1.0 dependency parsing results for Finnish language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 53.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
fra_arb	72.76	0.57	-0.75	0.116	50.86	0.88	-1.19	0.115
fra_cmn	71.82	1.01	-1.70	0.028	48.42	1.30	-3.63	0.004
fra_ces	70.58	1.07	-2.94	0.003	48.44	0.71	-3.61	0.001
fra_deu	71.12	0.69	-2.40	0.002	49.14	0.69	-2.91	0.002
fra_eng	69.82	0.65	-3.70	0.000	46.50	1.72	-5.55	0.001
fra_fin	71.97	0.84	-1.55	0.024	49.24	0.94	-2.81	0.005
fra_hin	71.70	0.71	-1.82	0.008	47.99	0.87	-4.05	0.001
fra_ind	71.13	0.54	-2.38	0.001	47.58	0.38	-4.46	0.000
fra_isl	72.94	0.64	-0.58	0.235	50.07	0.64	-1.98	0.013
fra_ita	71.15	1.38	-2.36	0.020	48.50	1.57	-3.54	0.008
fra_jpn	71.43	0.88	-2.09	0.008	49.73	1.18	-2.31	0.022
fra_kor	72.47	0.86	-1.05	0.093	49.63	1.18	-2.42	0.018
fra_pol	70.98	0.81	-2.54	0.002	48.66	0.70	-3.38	0.001
fra_por	72.22	0.52	-1.30	0.016	50.07	0.75	-1.98	0.017
fra_rus	72.35	0.43	-1.16	0.019	50.56	0.54	-1.48	0.034
fra_spa	71.63	0.50	-1.88	0.003	49.35	0.80	-2.69	0.005
fra_swe	71.57	0.63	-1.95	0.004	49.40	1.05	-2.65	0.009
fra_tha	72.40	0.39	-1.12	0.019	49.68	0.27	-2.37	0.003
fra_tur	70.25	0.63	-3.27	0.000	48.19	0.36	-3.86	0.000

Table A.53. UDPipe 1.0 dependency parsing results for French language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 54.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
hin_arb	65.23	0.36	3.19	0.000	41.82	0.42	2.87	0.000
hin_cmn	65.36	0.37	3.33	0.000	42.98	0.38	4.03	0.000
hin_ces	64.75	0.63	2.72	0.000	41.96	0.57	3.01	0.000
hin_deu	64.79	0.29	2.76	0.000	42.04	0.22	3.09	0.000
hin_eng	63.74	0.24	1.71	0.000	42.04	0.13	3.09	0.000
hin_fin	65.18	0.73	3.15	0.000	42.77	0.63	3.82	0.000
hin_fra	63.95	0.36	1.92	0.000	41.57	0.20	2.62	0.000
hin_ind	64.18	0.19	2.15	0.000	41.78	0.13	2.82	0.000
hin_isl	64.45	0.33	2.42	0.000	41.81	0.30	2.86	0.000
hin_ita	64.70	0.45	2.67	0.000	42.78	0.54	3.83	0.000
hin_jpn	64.37	0.64	2.34	0.001	40.46	0.20	1.51	0.000
hin_kor	63.90	0.53	1.87	0.001	41.75	0.48	2.80	0.000
hin_pol	65.03	0.81	3.00	0.000	42.11	0.62	3.16	0.000
hin_por	63.37	0.16	1.34	0.000	41.75	0.43	2.80	0.000
hin_rus	63.62	0.64	1.59	0.004	41.63	0.38	2.68	0.000
hin_spa	64.36	0.46	2.33	0.000	41.46	0.54	2.51	0.000
hin_swe	64.03	0.02	2.00	0.000	41.74	0.14	2.78	0.000
hin_tha	62.78	0.35	0.75	0.020	39.75	0.51	0.80	0.038
hin_tur	64.11	0.52	2.08	0.000	41.69	0.51	2.74	0.000

Table A.54. UDPipe 1.0 dependency parsing results for Hindi language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 55.

	LAS				**MLAS**			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
isl_arb	59.47	0.24	2.88	0.003	36.97	0.31	2.70	0.001
isl_cmn	59.32	0.57	2.74	0.005	35.82	0.31	1.55	0.011
isl_ces	59.55	0.57	2.97	0.004	36.54	0.36	2.27	0.002
isl_deu	59.65	0.66	3.07	0.004	34.65	0.51	0.39	0.444
isl_eng	57.89	0.39	1.30	0.074	33.90	0.23	-0.37	0.406
isl_fin	60.38	0.19	3.80	0.001	37.14	0.29	2.87	0.001
isl_fra	58.81	0.66	2.23	0.015	35.40	0.35	1.13	0.041
isl_hin	59.62	0.66	3.04	0.004	36.25	0.29	1.98	0.003
isl_ind	58.80	0.54	2.21	0.013	34.93	0.34	0.66	0.176
isl_ita	58.09	0.66	1.51	0.062	34.95	0.53	0.69	0.201
isl_jpn	58.17	0.70	1.59	0.056	36.72	0.62	2.45	0.003
isl_kor	59.03	0.41	2.45	0.007	35.98	0.09	1.71	0.005
isl_pol	60.98	0.75	4.39	0.001	36.83	0.45	2.56	0.001
isl_por	58.05	0.41	1.46	0.053	34.85	0.40	0.58	0.240
isl_rus	61.94	0.48	5.36	0.000	39.39	0.31	5.13	0.000
isl_spa	58.78	0.24	2.19	0.009	34.80	0.20	0.54	0.238
isl_swe	60.05	0.82	3.46	0.003	34.99	0.50	0.73	0.174
isl_tha	59.44	0.17	2.86	0.003	37.49	0.11	3.22	0.000
isl_tur	56.12	0.66	-0.47	0.507	33.42	0.30	-0.85	0.093

Table A.55. UDPipe 1.0 dependency parsing results for Icelandic language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 56.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
ind_arb	68.33	0.42	8.00	0.000	58.75	0.45	12.08	0.000
ind_cmn	69.48	1.02	9.16	0.000	60.12	1.05	13.45	0.000
ind_ces	68.73	0.56	8.40	0.000	58.97	0.92	12.30	0.000
ind_deu	66.36	0.88	6.04	0.000	56.36	1.02	9.69	0.000
ind_eng	66.69	0.41	6.36	0.000	57.08	0.70	10.41	0.000
ind_fin	66.71	0.57	6.38	0.000	56.42	0.92	9.75	0.000
ind_fra	67.75	0.64	7.42	0.000	57.60	1.00	10.94	0.000
ind_hin	67.85	0.79	7.52	0.000	58.22	1.16	11.55	0.000
ind_isl	66.43	0.75	6.10	0.000	56.50	0.90	9.84	0.000
ind_ita	68.09	0.72	7.77	0.000	58.13	0.94	11.46	0.000
ind_jpn	67.18	0.49	6.85	0.000	56.70	0.47	10.03	0.000
ind_kor	68.43	0.32	8.11	0.000	58.68	0.47	12.02	0.000
ind_pol	67.69	0.37	7.36	0.000	57.04	0.48	10.38	0.000
ind_por	67.90	0.21	7.58	0.000	57.64	0.37	10.97	0.000
ind_rus	68.57	0.61	8.25	0.000	58.97	0.84	12.30	0.000
ind_spa	68.59	0.39	8.27	0.000	58.67	0.31	12.00	0.000
ind_swe	67.40	0.42	7.07	0.000	57.37	0.76	10.70	0.000
ind_tha	67.60	0.84	7.27	0.000	56.58	0.92	9.92	0.000
ind_tur	65.52	0.64	5.20	0.000	55.08	0.68	8.41	0.000

Table A.56. UDPipe 1.0 dependency parsing results for Indonesian language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 57.

	LAS				**MLAS**			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
ita_arb	68.60	0.30	-0.88	0.042	45.30	0.35	-0.62	0.269
ita_cmn	68.50	0.62	-0.98	0.067	43.78	0.39	-2.14	0.006
ita_ces	68.44	0.69	-1.04	0.065	44.89	1.12	-1.04	0.209
ita_deu	67.59	0.85	-1.89	0.011	43.40	0.85	-2.52	0.007
ita_eng	67.56	0.36	-1.92	0.002	42.35	0.75	-3.57	0.001
ita_fin	67.91	0.56	-1.56	0.009	44.11	0.75	-1.81	0.025
ita_fra	67.80	0.70	-1.68	0.011	43.32	0.41	-2.60	0.002
ita_hin	68.31	0.71	-1.17	0.047	43.20	0.83	-2.72	0.005
ita_ind	68.17	0.57	-1.31	0.021	42.77	0.65	-3.15	0.002
ita_isl	69.47	0.14	-0.01	0.970	45.10	0.37	-0.82	0.159
ita_jpn	69.12	0.58	-0.36	0.432	45.53	0.28	-0.39	0.457
ita_kor	68.80	0.37	-0.68	0.107	43.96	0.37	-1.96	0.009
ita_pol	68.49	0.70	-0.99	0.077	44.82	0.76	-1.11	0.119
ita_por	68.04	0.44	-1.44	0.009	44.27	0.56	-1.66	0.024
ita_rus	69.74	0.61	0.26	0.573	46.54	0.53	0.62	0.297
ita_spa	68.09	0.34	-1.39	0.007	43.58	0.85	-2.34	0.011
ita_swe	68.33	0.35	-1.15	0.017	44.52	0.83	-1.40	0.069
ita_tha	68.53	0.65	-0.95	0.078	44.72	0.74	-1.20	0.093
ita_tur	69.02	0.78	-0.46	0.389	43.88	0.75	-2.04	0.015

Table A.57. UDPipe 1.0 dependency parsing results for Italian language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 58.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
jpn_arb	85.65	0.71	-0.23	0.606	70.93	1.07	-0.28	0.680
jpn_cmn	84.89	0.51	-0.99	0.026	66.90	0.54	-4.31	0.000
jpn_ces	85.50	0.52	-0.38	0.312	71.12	0.54	-0.09	0.846
jpn_deu	85.55	0.41	-0.33	0.313	70.43	0.48	-0.77	0.120
jpn_eng	85.55	1.15	-0.33	0.609	69.61	1.83	-1.60	0.154
jpn_fin	85.21	0.39	-0.67	0.064	70.37	1.02	-0.84	0.224
jpn_fra	85.77	0.54	-0.11	0.773	70.90	0.78	-0.30	0.583
jpn_hin	85.22	0.26	-0.66	0.041	70.80	0.39	-0.41	0.351
jpn_ind	85.26	0.58	-0.62	0.141	70.29	0.52	-0.92	0.080
jpn_isl	85.66	0.62	-0.22	0.591	70.45	0.69	-0.76	0.172
jpn_ita	85.55	0.40	-0.33	0.307	70.43	0.72	-0.78	0.176
jpn_kor	85.60	0.52	-0.27	0.450	70.63	0.74	-0.58	0.299
jpn_pol	85.21	0.66	-0.67	0.141	70.87	0.88	-0.34	0.569
jpn_por	85.49	0.55	-0.39	0.314	70.63	0.66	-0.58	0.273
jpn_rus	85.22	0.55	-0.66	0.111	70.23	0.69	-0.98	0.095
jpn_spa	85.75	1.07	-0.13	0.835	70.83	1.49	-0.38	0.660
jpn_swe	85.83	0.38	-0.05	0.862	70.72	0.37	-0.49	0.265
jpn_tha	85.82	0.22	-0.06	0.808	71.22	0.14	0.01	0.984
jpn_tur	85.78	0.47	-0.10	0.773	70.79	0.45	-0.42	0.351

Table A.58. UDPipe 1.0 dependency parsing results for Japanese language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 59.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
kor_arb	63.66	0.59	-2.26	0.019	47.26	0.39	-2.09	0.020
kor_cmn	63.65	0.68	-2.28	0.021	46.72	0.68	-2.63	0.011
kor_ces	62.62	0.86	-3.31	0.005	46.11	1.23	-3.24	0.011
kor_deu	64.35	0.77	-1.58	0.081	47.23	0.44	-2.11	0.020
kor_eng	63.68	0.48	-2.25	0.017	46.28	0.49	-3.06	0.004
kor_fin	64.24	0.81	-1.69	0.069	47.17	0.85	-2.18	0.030
kor_fra	63.81	0.61	-2.12	0.025	47.07	0.69	-2.27	0.020
kor_hin	62.87	0.83	-3.06	0.007	46.39	0.77	-2.95	0.007
kor_ind	63.72	0.71	-2.21	0.024	47.02	0.09	-2.32	0.011
kor_isl	63.49	0.61	-2.44	0.014	46.81	0.57	-2.54	0.011
kor_ita	64.40	0.50	-1.53	0.070	46.93	0.82	-2.42	0.019
kor_jpn	63.35	0.30	-2.58	0.008	45.76	0.50	-3.58	0.002
kor_pol	63.60	1.14	-2.33	0.036	47.07	1.17	-2.27	0.039
kor_por	63.79	0.55	-2.14	0.023	47.20	0.57	-2.15	0.022
kor_rus	63.71	0.61	-2.22	0.021	46.60	0.90	-2.75	0.013
kor_spa	64.21	1.29	-1.72	0.110	47.37	1.23	-1.97	0.067
kor_swe	63.00	1.43	-2.92	0.023	46.35	1.09	-3.00	0.012
kor_tha	62.38	1.29	-3.55	0.008	45.07	0.94	-4.28	0.002
kor_tur	64.24	0.78	-1.69	0.067	47.45	0.16	-1.90	0.026

Table A.59. UDPipe 1.0 dependency parsing results for Korean language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 60.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
pol_arb	62.08	0.92	3.84	0.002	35.20	0.54	3.60	0.000
pol_cmn	61.57	0.51	3.33	0.002	36.13	0.56	4.52	0.000
pol_ces	61.28	0.57	3.04	0.003	34.33	0.35	2.73	0.001
pol_deu	61.21	0.71	2.96	0.004	35.92	0.40	4.32	0.000
pol_eng	58.84	0.13	0.59	0.324	33.77	0.38	2.17	0.002
pol_fin	60.61	1.11	2.37	0.023	34.65	0.80	3.05	0.001
pol_fra	59.34	0.79	1.10	0.155	33.96	0.38	2.35	0.001
pol_hin	61.71	0.57	3.47	0.001	36.17	0.49	4.57	0.000
pol_ind	58.03	0.89	-0.22	0.769	33.09	0.74	1.48	0.031
pol_isl	60.77	0.34	2.53	0.005	34.92	0.54	3.32	0.000
pol_ita	57.10	0.17	-1.14	0.085	33.49	0.32	1.88	0.004
pol_jpn	61.23	0.80	2.98	0.005	36.09	0.80	4.48	0.000
pol_kor	60.01	0.45	1.76	0.025	35.95	0.33	4.35	0.000
pol_por	58.62	0.58	0.38	0.562	33.82	0.52	2.21	0.003
pol_rus	61.84	0.48	3.60	0.001	36.69	0.30	5.08	0.000
pol_spa	59.20	0.69	0.96	0.189	33.98	0.37	2.38	0.001
pol_swe	58.86	0.71	0.62	0.382	33.49	0.36	1.89	0.004
pol_tha	61.08	0.36	2.84	0.003	36.17	0.15	4.57	0.000
pol_tur	56.55	0.68	-1.70	0.039	31.94	0.54	0.33	0.503

Table A.60. UDPipe 1.0 dependency parsing results for Polish language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 61.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
por_arb	74.99	0.36	4.23	0.000	57.78	0.48	6.12	0.000
por_cmn	74.71	0.47	3.95	0.000	57.09	0.39	5.43	0.000
por_ces	73.98	0.61	3.22	0.000	56.00	0.75	4.34	0.000
por_deu	73.17	0.50	2.41	0.000	54.36	0.44	2.70	0.000
por_eng	72.72	0.88	1.95	0.006	55.33	1.01	3.67	0.000
por_fin	74.27	0.57	3.50	0.000	55.76	0.41	4.10	0.000
por_fra	75.41	0.38	4.65	0.000	58.61	0.50	6.95	0.000
por_hin	74.46	0.86	3.70	0.000	57.99	1.14	6.32	0.000
por_ind	74.22	0.44	3.46	0.000	55.96	0.26	4.30	0.000
por_isl	74.87	0.32	4.10	0.000	57.02	0.38	5.36	0.000
por_ita	74.82	0.25	4.06	0.000	57.71	0.24	6.04	0.000
por_jpn	75.32	0.22	4.56	0.000	58.89	0.35	7.22	0.000
por_kor	75.44	0.46	4.67	0.000	58.81	0.60	7.15	0.000
por_pol	74.78	0.21	4.02	0.000	58.17	0.32	6.51	0.000
por_rus	74.86	0.47	4.09	0.000	58.24	0.53	6.57	0.000
por_spa	76.58	0.41	5.81	0.000	58.46	0.53	6.79	0.000
por_swe	74.89	0.41	4.12	0.000	56.80	0.53	5.13	0.000
por_tha	70.08	0.34	-0.68	0.022	50.08	0.89	-1.59	0.013
por_tur	73.87	0.78	3.11	0.000	56.79	0.69	5.13	0.000

Table A.61. UDPipe 1.0 dependency parsing results for Portuguese language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 62.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
rus_arb	66.19	0.16	1.54	0.001	45.16	0.24	3.17	0.000
rus_cmn	67.65	0.69	3.00	0.000	46.81	0.67	4.81	0.000
rus_ces	67.10	0.58	2.46	0.001	45.87	0.36	3.87	0.000
rus_deu	65.49	0.20	0.84	0.014	44.89	0.25	2.90	0.000
rus_eng	65.94	0.54	1.30	0.010	44.96	0.32	2.96	0.000
rus_fin	67.11	0.32	2.47	0.000	45.18	0.25	3.19	0.000
rus_fra	65.97	0.46	1.32	0.006	45.06	0.26	3.06	0.000
rus_hin	66.15	0.40	1.51	0.002	44.62	0.45	2.63	0.000
rus_ind	67.68	0.46	3.04	0.000	46.37	0.23	4.38	0.000
rus_isl	66.49	0.40	1.85	0.001	45.02	0.27	3.03	0.000
rus_ita	66.18	0.57	1.54	0.005	45.62	0.58	3.63	0.000
rus_jpn	66.17	1.05	1.52	0.037	44.86	0.63	2.87	0.000
rus_kor	65.87	0.66	1.22	0.022	45.28	0.52	3.29	0.000
rus_pol	65.06	0.27	0.42	0.161	44.88	0.25	2.88	0.000
rus_por	66.38	0.43	1.74	0.001	44.43	0.33	2.43	0.000
rus_spa	66.56	0.69	1.92	0.003	45.18	0.37	3.19	0.000
rus_swe	66.64	0.23	2.00	0.000	44.71	0.33	2.72	0.000
rus_tha	67.37	0.37	2.73	0.000	46.30	0.15	4.31	0.000
rus_tur	66.25	0.40	1.61	0.002	45.83	0.32	3.83	0.000

Table A.62. UDPipe 1.0 dependency parsing results for Russian language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 63.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
spa_arb	69.49	0.77	-0.31	0.521	48.91	0.94	-0.79	0.223
spa_cmn	69.61	1.08	-0.19	0.765	46.90	1.63	-2.79	0.019
spa_ces	69.57	0.67	-0.23	0.594	48.21	1.13	-1.48	0.065
spa_deu	69.56	0.29	-0.23	0.437	46.91	0.75	-2.79	0.001
spa_eng	69.02	0.42	-0.78	0.052	46.24	0.33	-3.45	0.000
spa_fin	68.94	0.77	-0.86	0.110	47.13	0.79	-2.57	0.003
spa_fra	69.94	0.22	0.14	0.624	47.34	0.28	-2.36	0.001
spa_hin	69.48	0.98	-0.32	0.577	47.27	1.21	-2.42	0.013
spa_ind	69.30	0.47	-0.50	0.187	45.75	0.67	-3.95	0.000
spa_isl	69.76	0.83	-0.04	0.937	48.24	0.81	-1.46	0.032
spa_ita	69.99	0.42	0.19	0.575	46.67	0.47	-3.03	0.000
spa_jpn	69.05	0.60	-0.75	0.098	47.46	0.39	-2.23	0.001
spa_kor	68.75	0.52	-1.05	0.026	45.94	0.65	-3.75	0.000
spa_pol	70.06	0.61	0.26	0.527	49.61	1.01	-0.09	0.893
spa_por	71.00	0.85	1.20	0.049	49.17	0.91	-0.53	0.386
spa_rus	69.16	1.00	-0.64	0.295	49.13	1.26	-0.57	0.456
spa_swe	67.97	0.32	-1.83	0.001	45.89	0.22	-3.80	0.000
spa_tha	69.38	0.45	-0.41	0.258	47.10	0.15	-2.59	0.000
spa_tur	68.76	0.96	-1.04	0.101	47.66	1.47	-2.04	0.044

Table A.63. UDPipe 1.0 dependency parsing results for Spanish language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 64.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
swe_arb	63.00	1.15	-1.23	0.145	46.29	1.24	-0.86	0.276
swe_cmn	62.62	1.58	-1.62	0.127	44.52	1.58	-2.63	0.023
swe_ces	62.88	0.67	-1.35	0.055	45.74	0.34	-1.41	0.012
swe_deu	63.31	0.52	-0.93	0.131	46.11	0.69	-1.04	0.083
swe_eng	62.46	1.03	-1.78	0.042	44.28	0.98	-2.87	0.003
swe_fin	62.67	0.65	-1.57	0.032	45.20	0.56	-1.95	0.005
swe_fra	62.30	0.62	-1.94	0.013	43.41	0.47	-3.74	0.000
swe_hin	62.84	0.33	-1.40	0.029	45.08	0.24	-2.07	0.002
swe_ind	61.72	0.68	-2.51	0.005	42.26	0.97	-4.89	0.000
swe_isl	63.87	0.84	-0.36	0.579	45.82	0.45	-1.32	0.021
swe_ita	61.83	0.37	-2.41	0.003	43.43	0.31	-3.72	0.000
swe_jpn	62.62	0.94	-1.62	0.049	44.52	0.85	-2.63	0.003
swe_kor	63.36	0.41	-0.87	0.135	45.04	0.74	-2.11	0.007
swe_pol	63.52	1.43	-0.72	0.431	46.46	0.97	-0.69	0.296
swe_por	61.25	0.68	-2.99	0.002	43.75	0.62	-3.39	0.000
swe_rus	63.62	1.02	-0.62	0.402	46.63	1.23	-0.52	0.495
swe_spa	62.15	1.10	-2.09	0.027	44.89	1.29	-2.26	0.022
swe_tha	62.59	1.12	-1.65	0.063	44.32	0.91	-2.83	0.003
swe_tur	61.35	1.22	-2.89	0.009	44.57	1.36	-2.58	0.015

Table A.64. UDPipe 1.0 dependency parsing results for Swedish language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 65.

	LAS				MLAS			
	mean	**stdev**	**delta**	**p**	**mean**	**stdev**	**delta**	**p**
tha_arb	61.58	1.17	3.51	0.002	50.76	1.43	5.45	0.000
tha_cmn	59.61	0.70	1.54	0.013	47.32	0.65	2.01	0.006
tha_ces	62.32	0.39	4.25	0.000	51.16	0.64	5.85	0.000
tha_deu	62.10	0.80	4.03	0.000	51.03	0.76	5.72	0.000
tha_eng	62.32	0.39	4.25	0.000	51.16	0.64	5.85	0.000
tha_fin	61.83	0.95	3.75	0.000	50.39	1.04	5.08	0.000
tha_fra	62.64	0.66	4.57	0.000	51.71	0.85	6.40	0.000
tha_hin	61.68	0.41	3.61	0.000	50.58	0.42	5.27	0.000
tha_ind	62.17	0.75	4.10	0.000	50.89	0.85	5.59	0.000
tha_isl	61.86	0.53	3.79	0.000	50.64	0.55	5.34	0.000
tha_ita	62.48	0.64	4.40	0.000	51.55	0.89	6.25	0.000
tha_jpn	60.60	0.45	2.53	0.000	49.35	0.67	4.04	0.000
tha_kor	61.32	0.84	3.25	0.001	50.28	1.20	4.98	0.000
tha_pol	61.90	0.86	3.83	0.000	50.74	1.05	5.43	0.000
tha_por	62.17	0.57	4.10	0.000	50.86	0.69	5.55	0.000
tha_rus	61.58	0.26	3.51	0.000	50.32	0.18	5.02	0.000
tha_spa	61.63	0.59	3.56	0.000	50.41	0.68	5.10	0.000
tha_swe	62.15	0.70	4.07	0.000	51.26	0.84	5.95	0.000
tha_tur	61.59	0.75	3.52	0.000	50.67	0.73	5.36	0.000

Table A.65. UDPipe 1.0 dependency parsing results for Thai language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 66.

	LAS				MLAS			
	mean	stdev	delta	p	mean	stdev	delta	p
tur_arb	49.49	1.05	-1.97	0.042	30.60	1.20	-1.91	0.042
tur_cmn	50.27	0.84	-1.19	0.138	30.89	0.63	-1.62	0.023
tur_ces	48.08	0.54	-3.38	0.002	29.99	0.70	-2.52	0.004
tur_deu	47.07	0.39	-4.39	0.000	27.99	0.69	-4.52	0.000
tur_eng	48.94	0.27	-2.52	0.005	29.53	0.51	-2.98	0.001
tur_fin	48.88	0.80	-2.59	0.009	29.57	0.84	-2.94	0.003
tur_fra	48.48	0.37	-2.98	0.002	29.29	0.55	-3.22	0.001
tur_hin	48.78	1.04	-2.69	0.012	29.70	0.77	-2.81	0.003
tur_ind	48.39	0.17	-3.07	0.002	29.17	0.36	-3.34	0.000
tur_isl	48.36	0.95	-3.10	0.005	28.68	0.76	-3.83	0.001
tur_ita	48.82	0.13	-2.64	0.003	28.92	0.38	-3.59	0.000
tur_jpn	49.93	0.73	-1.53	0.060	30.80	0.65	-1.71	0.020
tur_kor	50.22	0.38	-1.24	0.079	30.40	0.19	-2.11	0.003
tur_pol	47.84	0.81	-3.62	0.002	29.83	1.02	-2.68	0.007
tur_por	48.35	0.78	-3.12	0.004	29.03	1.09	-3.48	0.002
tur_rus	48.78	0.75	-2.68	0.007	30.19	1.05	-2.32	0.015
tur_spa	47.96	0.84	-3.50	0.002	28.52	0.73	-3.99	0.000
tur_swe	50.01	0.86	-1.45	0.084	30.06	0.85	-2.45	0.007
tur_tha	49.50	0.79	-1.96	0.028	30.71	0.97	-1.80	0.033

Table A.66. UDPipe 1.0 dependency parsing results for Turkish language concerning corpora association experiments. Positive deltas appear in green, while negative ones, in red. P-values equal or lower to 0.01 are also coloured in green.

Annex 67.

	MarsaGram All		MarsaGram Linear		Head/Dependent		VO_OV		Lang2vec	
	Euc.	**cos**	**Euc.**	**cos**	**Euc.**	**cos**	**Euc.**	**cos**	**Euc.**	**cos**
arb	0.23	-0.45	-0.25	-0.61	-0.59	-0.66	-0.75	-0.52	-0.82	-0.85
cmn	-0.33	0.09	-0.15	0.03	0.02	0.16	0.22	0.45	0.18	0.15
ces	-0.15	0.05	0.05	0.18	-0.19	-0.12	0.29	0.11	0.01	0.03
eng	0.43	0.20	0.35	0.30	0.29	0.31	-0.15	-0.19	0.20	0.20
fin	0.04	0.29	-0.22	0.09	-0.10	0.06	0.13	-0.11	-0.08	-0.11
fra	0.08	0.06	0.04	0.07	0.17	0.14	-0.08	-0.26	-0.04	0.00
deu	0.80	0.59	0.20	0.29	0.60	0.55	0.02	-0.11	0.00	0.01
hin	0.46	0.28	0.36	0.32	-0.25	-0.26	-0.26	-0.06	-0.25	-0.29
isl	-0.02	-0.08	-0.34	-0.22	-0.48	-0.36	0.07	-0.01	-0.37	-0.38
ind	0.26	0.15	-0.07	-0.12	0.00	-0.18	0.02	-0.12	-0.16	-0.15
ita	0.58	0.58	0.57	0.55	0.61	0.58	0.18	0.04	0.47	0.51
jpn	-0.28	-0.14	-0.23	-0.19	0.19	0.31	0.11	0.09	0.22	0.21
kor	0.25	0.19	0.08	0.04	0.22	0.02	-0.33	-0.26	0.03	0.01
pol	-0.06	-0.02	-0.18	-0.02	-0.14	-0.03	0.18	0.01	0.10	0.09
por	-0.09	-0.07	-0.15	-0.12	-0.08	-0.11	-0.22	-0.02	-0.11	-0.09
rus	-0.31	-0.05	0.17	0.02	-0.21	-0.20	0.03	-0.17	0.05	0.10
spa	-0.23	-0.43	-0.35	-0.51	-0.47	-0.46	-0.59	-0.42	-0.43	-0.49
swe	-0.03	0.02	0.02	0.21	-0.15	-0.16	0.14	0.01	0.01	-0.08
tha	0.13	-0.50	0.01	-0.22	-0.24	-0.39	-0.60	-0.32	-0.69	-0.70
tur	-0.43	-0.11	-0.30	-0.20	0.00	-0.16	-0.09	-0.21	-0.20	-0.18

Table A.67. UDPipe Spearman’s correlation coefficient for each PUD language and each typological strategy calculated between the language distances and the LAS deltas. Values in green indicate a strong correlation, in yellow, a moderate one. Correlation coefficients between -0.50 and -0.40 are presented in red.

Annex 68.

	MarsaGram All		**MarsaGram Linear**		**Head/Dependent**		**VO_OV**		**Lang2vec**	
	Euc.	**cos**	**Euc.**	**cos**	**Euc.**	**cos**	**Euc.**	**cos**	**Euc.**	**cos**
arb	0.01	-0.39	-0.19	-0.55	-0.63	-0.71	-0.69	-0.63	-0.71	-0.72
cmn	-0.16	-0.29	0.04	-0.22	0.19	0.15	0.09	0.19	0.00	0.06
ces	-0.10	0.28	-0.12	0.28	-0.02	0.17	0.42	0.22	0.04	0.09
eng	0.41	0.27	0.37	0.33	0.22	0.23	-0.03	-0.03	0.17	0.14
fin	0.10	0.46	-0.22	0.11	-0.02	0.09	0.40	0.31	-0.12	-0.10
fra	0.01	0.09	0.07	-0.02	0.06	0.05	-0.11	-0.09	0.00	0.01
deu	0.45	0.37	0.04	0.10	0.25	0.26	0.05	0.05	-0.23	-0.20
hin	0.41	-0.05	0.09	0.56	0.00	-0.01	-0.05	0.16	-0.35	-0.45
isl	0.09	0.21	-0.03	-0.06	-0.08	-0.07	0.05	-0.14	-0.02	0.02
ind	0.21	0.03	-0.22	-0.01	-0.16	-0.21	-0.13	-0.25	-0.16	-0.19
ita	0.48	0.37	0.39	0.32	0.33	0.25	-0.08	-0.08	0.16	0.16
jpn	-0.09	0.17	-0.11	-0.26	-0.10	-0.12	0.10	-0.04	0.03	0.10
kor	0.30	0.19	-0.02	0.22	0.09	0.06	-0.33	-0.01	0.00	-0.10
pol	0.02	0.10	-0.03	-0.03	0.14	0.16	0.30	0.17	0.09	0.13
por	0.11	-0.03	0.08	-0.04	0.04	0.09	-0.15	0.27	-0.01	0.04
rus	-0.36	0.09	-0.10	-0.03	-0.22	-0.10	-0.01	-0.17	0.20	0.23
spa	0.10	-0.10	0.04	-0.07	-0.05	-0.06	-0.30	-0.20	-0.16	-0.16
swe	0.25	0.12	0.06	0.16	-0.07	-0.01	0.15	0.03	0.00	-0.04
tha	-0.04	-0.38	-0.16	-0.15	-0.22	-0.23	-0.24	-0.19	-0.40	-0.41
tur	-0.39	0.33	-0.31	-0.26	0.06	0.17	0.03	-0.12	-0.24	-0.15

Table A.68. UDPipe Pearson’s correlation coefficient for each PUD language and each typological strategy calculated between the language distances and the MLAS deltas. Values in green indicate a strong correlation, in yellow, a moderate one. Correlation coefficients between -0.50 and -0.40 are presented in red.

Annex 69.

	MarsaGram All		MarsaGram Linear		Head/Dependent		VO_OV		Lang2vec	
	Euc.	**cos**	**Euc.**	**cos**	**Euc.**	**cos**	**Euc.**	**cos**	**Euc.**	**cos**
arb	0.03	-0.40	-0.33	-0.56	-0.59	-0.62	-0.61	-0.48	-0.71	-0.73
cmn	-0.27	0.29	0.08	0.07	0.01	0.15	0.16	0.37	0.06	0.11
ces	-0.18	0.28	0.03	0.36	0.05	0.13	0.40	0.20	0.11	0.14
eng	0.44	0.29	0.37	0.38	0.33	0.36	-0.13	-0.17	0.24	0.24
fin	-0.04	0.47	-0.32	0.28	-0.13	0.07	0.32	0.04	-0.13	-0.17
fra	-0.03	0.02	-0.15	-0.06	0.02	0.00	-0.12	-0.20	-0.12	-0.09
deu	0.59	0.47	0.04	0.10	0.38	0.34	0.02	0.00	-0.21	-0.20
hin	0.29	0.04	0.14	0.47	-0.18	-0.20	0.00	0.20	-0.25	-0.25
isl	0.03	0.25	-0.05	0.06	-0.18	-0.06	0.22	0.04	0.11	0.11
ind	0.23	0.02	-0.09	-0.04	-0.06	-0.25	-0.06	-0.18	-0.07	-0.07
ita	0.50	0.49	0.30	0.29	0.36	0.32	-0.10	-0.35	0.19	0.23
jpn	-0.16	0.17	-0.40	-0.21	-0.02	0.21	0.05	0.01	0.25	0.22
kor	0.15	0.07	0.01	0.17	0.15	0.20	-0.43	-0.25	0.09	0.06
pol	-0.08	0.09	-0.16	-0.02	-0.02	0.08	0.29	0.00	0.12	0.11
por	0.14	0.07	-0.02	0.03	0.10	0.09	-0.15	0.11	0.05	0.04
rus	-0.24	0.16	0.05	0.03	0.02	0.02	0.10	0.08	0.16	0.20
spa	0.34	0.09	0.13	0.07	0.08	0.08	-0.34	-0.34	-0.06	-0.09
swe	0.14	0.13	0.09	0.19	-0.09	-0.14	0.22	0.09	-0.02	-0.11
tha	0.02	-0.56	-0.15	-0.29	-0.20	-0.43	-0.47	-0.30	-0.58	-0.57
tur	-0.52	0.16	-0.34	-0.25	0.00	-0.06	-0.11	-0.22	-0.28	-0.26

Table A.69. UDPipe Spearman’s correlation coefficient for each PUD language and each typological strategy calculated between the language distances and the MLAS deltas. Values in green indicate a strong correlation, in yellow, a moderate one. Correlation coefficients between -0.50 and -0.40 are presented in red.

Annex 70.

			Right Choice (1)	Equal to right (2)	(1) + (2)	Negative delta (p<0,01) (3)	(1) + (2) - (3)	Inferior to right but positive (p<0,01) (4)	(1) + (2) - (3) + (4)
Msg All + Msg Lin + HD + VO	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.4	6	1	7	3	4	3	7
Msg All + Msg Lin + HD + VO	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.4	4	0	4	2	2	4	6
Msg All + HD	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.4	6	3	9	0	9	3	12
Msg All + HD	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.4	5	1	6	2	4	5	9
Msg All + VO	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.4	5	3	8	0	8	4	12
Msg All + VO	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.4	5	2	7	2	5	5	10
Msg Linear + HD	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.4	7	2	9	3	6	3	9
Msg Linear + HD	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.4	6	1	7	3	4	4	8
Msg Linear + VO	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.4	6	2	8	2	6	3	9
Msg Linear + VO	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.4	5	1	6	3	3	3	6
Msg All + Msg Lin + HD + VO	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.1	6	2	8	0	8	3	10
Msg All + Msg Lin + HD + VO	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.1	7	1	8	2	6	3	9
Msg All + HD	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.1	5	3	8	1	7	4	11
Msg All + HD	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.1	5	1	6	2	4	5	9
Msg All + VO	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.1	5	3	8	1	7	4	11

			Right Choice (1)	Equal to right (2)	(1) + (2)	Negative delta (p<0,01) (3)	(1) + (2) - (3)	Inferior to right but positive (p<0,01) (4)	(1) + (2) - (3) + (4)
Msg All + VO	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.1	6	2	8	3	5	3	8
Msg Linear + HD	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.1	3	0	3	2	1	7	8
Msg Linear + HD	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.1	5	3	8	0	8	4	11
Msg Linear + VO	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.1	5	0	5	2	3	6	9
Msg Linear + VO	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.1	5	3	8	0	8	4	11
Msg All + Msg Lin + HD + VO	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.7	4	2	6	2	4	5	9
Msg All + Msg Lin + HD + VO	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.7	7	2	9	3	6	3	9
Msg All + HD	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.7	2	0	2	2	0	9	9
Msg All + HD	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.7	6	2	8	2	6	3	9
Msg All + VO	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.7	6	1	7	3	4	3	7
Msg All + VO	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.7	7	1	8	1	7	3	10
Msg Linear + HD	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.7	6	1	7	2	5	4	9
Msg Linear + HD	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.7	5	3	8	1	7	4	11
Msg Linear + VO	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.7	6	1	7	2	5	5	10
Msg Linear + VO	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.7	5	3	8	1	7	4	11

			Right Choice (1)	Equal to right (2)	(1) + (2)	Negative delta (p<0,01) (3)	(1) + (2) - (3)	Inferior to right but positive (p<0,01) (4)	(1) + (2) - (3) + (4)
Msg All + Msg Lin + HD + VO	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.4	6	2	8	3	5	3	8
Msg All + Msg Lin + HD + VO	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.4	3	0	3	2	1	7	8
Msg All + HD	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.4	5	3	8	0	8	4	11
Msg All + HD	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.4	5	0	5	2	3	6	9
Msg All + VO	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.4	5	3	8	0	8	4	11
Msg All + VO	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.4	4	2	6	2	4	5	9
Msg Linear + HD	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.4	7	2	9	3	6	3	9
Msg Linear + HD	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.4	2	0	2	2	0	9	9
Msg Linear + VO	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.4	6	2	8	2	6	3	9
Msg Linear + VO	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.4	6	1	7	3	4	3	7
Msg All + Msg Lin + HD + VO	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.1	7	1	8	1	7	3	10
Msg All + Msg Lin + HD + VO	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.1	6	1	7	2	5	4	9
Msg All + HD	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.1	5	3	8	1	7	4	11
Msg All + HD	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.1	6	1	7	2	5	5	10
Msg All + VO	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.1	5	3	8	1	7	4	11
Msg All + VO	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.1	4	2	6	2	4	5	9

			Right Choice (1)	Equal to right (2)	(1) + (2)	Negative delta (p<0,01) (3)	(1) + (2) - (3)	Inferior to right but positive (p<0,01) (4)	(1) + (2) - (3) + (4)
Msg Linear + HD	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.1	6	2	8	3	5	3	8
Msg Linear + HD	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.1	6	1	7	2	5	4	9
Msg Linear + VO	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.1	6	1	7	2	5	3	8
Msg Linear + VO	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.1	6	1	7	3	4	3	7
Msg All + Msg Lin + HD + VO	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.7	4	1	5	3	2	2	4
Msg All + Msg Lin + HD + VO	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.7	2	0	2	2	0	8	8
Msg All + HD	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.7	5	2	7	1	6	4	11
Msg All + HD	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.7	4	0	4	2	2	7	9
Msg All + VO	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.7	5	2	7	1	6	4	11
Msg All + VO	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.7	4	2	6	2	4	5	9
Msg Linear + HD	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.7	7	2	9	2	7	3	9
Msg Linear + HD	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.7	2	0	2	2	0	9	9
Msg Linear + VO	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.7	6	2	8	2	6	3	9
Msg Linear + VO	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.7	6	1	7	3	4	3	7

Table A.70. Results of the linear regression experiments concerning the combination of the different typological methods (LAS). In green are presented the best scores and yellow the second-best ones.

Annex 71.

			Right Choice (1)	Equal to right (2)	(1) + (2)	Negative delta (p<0,01) (3)	(1) + (2) - (3)	Inferior to right but positive (p<0,01) (4)	(1) + (2) - (3) + (4)
Msg All + Msg Lin + HD + VO	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.4	4	4	8	2	6	2	8
Msg All + Msg Lin + HD + VO	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.4	5	3	8	2	6	7	13
Msg All + HD	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.4	5	4	9	2	7	5	11
Msg All + HD	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.4	5	4	9	2	7	5	12
Msg All + VO	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.4	5	4	9	2	7	5	11
Msg All + VO	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.4	4	3	7	2	5	7	12
Msg Linear + HD	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.4	4	4	8	2	6	4	10
Msg Linear + HD	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.4	4	4	8	2	6	3	9
Msg Linear + VO	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.4	4	3	7	2	5	2	7
Msg Linear + VO	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.4	4	4	8	2	6	3	9
Msg All + Msg Lin + HD + VO	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.1	5	3	8	3	5	5	10
Msg All + Msg Lin + HD + VO	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.1	4	2	6	2	4	7	11
Msg All + HD	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.1	5	4	9	3	6	5	11
Msg All + HD	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.1	5	4	9	2	7	5	12
Msg All + VO	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.1	5	4	9	3	6	5	11
Msg All + VO	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.1	4	3	7	2	5	7	12
Msg Linear + HD	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.1	4	4	8	2	6	2	8
Msg Linear + HD	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.1	4	4	8	2	6	3	9

			Right Choice (1)	Equal to right (2)	(1) + (2)	Negative delta (p<0,01) (3)	(1) + (2) - (3)	Inferior to right but positive (p<0,01) (4)	(1) + (2) - (3) + (4)
Msg Linear + VO	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.1	4	3	7	2	5	2	7
Msg Linear + VO	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.1	4	4	8	2	6	3	9
Msg All + Msg Lin + HD + VO	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.7	3	4	7	2	5	3	9
Msg All + Msg Lin + HD + VO	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.7	4	3	7	2	5	7	12
Msg All + HD	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.7	6	3	9	2	7	5	12
Msg All + HD	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.7	5	4	9	2	7	5	12
Msg All + VO	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.7	6	3	9	2	7	5	12
Msg All + VO	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.7	4	3	7	2	5	7	12
Msg Linear + HD	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.7	4	3	7	2	5	6	11
Msg Linear + HD	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.7	4	4	8	2	6	3	9
Msg Linear + VO	Euc	learning_rate = 0.5,TOL=1e-7,theta0=0.7	4	4	8	2	6	2	8
Msg Linear + VO	cos	learning_rate = 0.5,TOL=1e-7,theta0=0.7	4	4	8	2	6	3	9
Msg All + Msg Lin + HD + VO	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.4	4	4	8	2	6	2	8
Msg All + Msg Lin + HD + VO	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.4	4	4	8	2	6	5	11
Msg All + HD	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.4	5	4	9	2	7	5	11
Msg All + HD	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.4	4	3	7	2	5	6	11
Msg All + VO	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.4	5	4	9	2	7	5	11
Msg All + VO	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.4	4	4	8	2	6	5	11
Msg Linear + HD	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.4	4	4	8	2	6	4	10

			Right Choice (1)	Equal to right (2)	(1) + (2)	Negative delta (p<0,01) (3)	(1) + (2) - (3)	Inferior to right but positive (p<0,01) (4)	(1) + (2) - (3) + (4)
Msg Linear + HD	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.4	4	3	7	2	5	6	11
Msg Linear + VO	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.4	4	3	7	2	5	2	7
Msg Linear + VO	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.4	4	4	8	2	6	3	9
Msg All + Msg Lin + HD + VO	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.1	5	3	8	3	5	5	10
Msg All + Msg Lin + HD + VO	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.1	4	2	6	2	4	6	10
Msg All + HD	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.1	5	4	9	3	6	5	11
Msg All + HD	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.1	4	3	7	2	5	6	11
Msg All + VO	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.1	5	4	9	3	6	5	11
Msg All + VO	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.1	4	4	8	2	6	5	11
Msg Linear + HD	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.1	4	4	8	2	6	2	8
Msg Linear + HD	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.1	4	3	7	2	5	5	10
Msg Linear + VO	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.1	4	3	7	2	5	2	7
Msg Linear + VO	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.1	4	4	8	2	6	3	9
Msg All + Msg Lin + HD + VO	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.7	4	4	8	2	6	2	8
Msg All + Msg Lin + HD + VO	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.7	4	3	7	2	5	6	11
Msg All + HD	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.7	6	3	9	2	7	5	12
Msg All + HD	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.7	4	3	7	2	5	6	11
Msg All + VO	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.7	6	3	9	2	7	5	12
Msg All + VO	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.7	4	4	8	2	6	5	11

			Right Choice (1)	Equal to right (2)	(1) + (2)	Negative delta (p<0,01) (3)	(1) + (2) - (3)	Inferior to right but positive (p<0,01) (4)	(1) + (2) - (3) + (4)
Msg Linear + HD	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.7	4	3	7	2	5	6	11
Msg Linear + HD	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.7	4	3	7	2	5	6	11
Msg Linear + VO	Euc	learning_rate = 0.1,TOL=1e-7,theta0=0.7	4	4	8	2	6	2	8
Msg Linear + VO	cos	learning_rate = 0.1,TOL=1e-7,theta0=0.7	4	4	8	2	6	3	9

Table A.71. Results of the linear regression experiments concerning the combination of the different typological methods (MLAS). In green are presented the best scores and yellow the second-best ones.

Annex 72.

🟧 arb	DepRel labels: det:predet --> det (247) nmod:gmod --> nmod (437)
🟧 cmn	DepRel labels: mark:relcl --> mark:rel (626)
🟨 ces	Feats: VerbForm, Voice added to 2.10 (+397 and +246) Some changes in 10th column 5 nese být VERB --> 5 nese být AUX
🟩 eng	XPOS: NNS --> NNPS Some punctuations (--> -LRB- 10th column: OrigForm --> _ Lemmas: lower case in old (all lemmas)
🟨 fin	UPOS: VERB --> ADJ (2) PROPN --> ADJ (1) FEATS: Degree - 497 less in new one
🟨 fra	FEATS: Definite | PronType | Gender in 2.10 -t: PART in 2.7 and -t-il as one token in 2.10
🟨 deu	Feats: Number and Person --> NumType z --> z. (zu + dem) DepRel labels: det:predet --> det (32) nmod:gmod --> nmod (1)

🟧 hin	DepRel: compound:conjv --> compound:lvc (447) det:predet --> det (10) 10th column: OrigForm=-- and Bug=case-child in 2.7
🟩 isl	No difference
🟨 ind	FEATS: Number=Sing|PronType=Ind --> Definite=Ind|PronType=Art Polite=Form in 2.10 some changes in UPOS and heads
🟩 ita	10th column: OrigForm - 21 cases only in 2.7 SpaceAfter - 2886 in 2.10 / 2881 in 2.7
🟩 jpn	10th column: UniDicLemma --> UniDicInfo and BunsetuBIlabel Few changes in deprel
🟧 kor	FEATS: Case=Advb (1461 only in 2.7) Case=Comp (61 only in 2.7) DepRel: dep:prt --> dep (404)
🟨 pol	Comments: Wrong id and no English text in 2.7 FEATS: ConjType=Cmpr --> ConjType=Comp (111) PunctType=Peri --> PunctType=Dash

🟨 por	FEATS: Definite (4 --> 258) PronType (258 --> 2984) Token: naqueles --> em aqueles 10th column: OrigForm (28 only in 2.7)
🟨 rus	FEATS: Degree=Comp (95 --> 82) Degree=Pos (2522 --> 2506) 10th column: OrigForm (106 in 2.7) Some differences regarding SpaceAfter=No
🟩 spa	Lemmas: No lemmas in 2.7 but present in 2.10 FEATS: Few corrections in Tense and VerbForm
🟩 swe	10th column: Lang=en in 2.10 (2 tokens)
🟥 tha	UPOS: Many cases of VERB --> AUX PART --> AUX advmod --> obl
🟥 tur	UPOS: Many cases of NOUN --> VERB Few cases of ADJ --> VERB FEATS: VerbForm and Number differences Polite only in 2.10

Table A.72. Differences in the PUD annotations between UD v.2.7 and UD v.2.10. Languages in green present few or no changes, in yellow, some changes regarding FEATS, in orange, changes in FEATS and some DEPREL, and in red, important changes in terms of UPOS.

Annex 73.

	arb	bul	cmn	hrv	Ces	dan	nld	eng	est	fin	fra	deu	ell	hin	hun	isl	ind	gle	ita	jpn	kor	lav	lit	mlt	pol	por	ron	rus	slk	slv	spa	swe	tha	tur
arb	0.00	3.46	3.16	3.61	3.46	3.46	3.87	3.87	2.83	3.00	4.24	3.46	3.32	3.61	2.65	3.46	3.74	3.61	3.87	3.00	2.65	3.00	3.00	1.00	3.32	4.12	3.74	3.16	3.32	3.32	4.24	3.46	4.00	3.46
bul	3.46	0.00	3.16	2.24	2.45	3.16	3.61	3.61	2.83	3.00	4.00	3.16	3.00	3.32	2.65	3.16	3.74	3.32	3.61	3.00	2.65	2.24	2.24	3.61	2.24	3.87	3.46	2.00	2.24	1.73	4.00	3.16	4.00	3.46
cmn	3.16	3.16	0.00	3.32	3.16	3.16	3.61	3.61	2.45	2.65	4.00	3.16	3.00	3.32	2.24	3.16	3.46	3.32	3.61	2.65	2.24	2.65	2.65	3.32	3.00	3.87	3.46	2.83	3.00	3.00	4.00	3.16	3.74	3.16
hrv	3.61	2.24	3.32	0.00	2.65	3.32	3.74	3.74	3.00	3.16	4.12	3.32	3.16	3.46	2.83	3.32	3.87	3.46	3.74	3.16	2.83	2.45	2.45	3.74	2.45	4.00	3.61	2.24	2.45	1.41	4.12	3.32	4.12	3.61
ces	3.46	2.45	3.16	2.65	0.00	3.16	3.61	3.61	2.83	3.00	4.00	3.16	3.00	3.32	2.65	3.16	3.74	3.32	3.61	3.00	2.65	2.24	2.24	3.61	1.73	3.87	3.46	2.00	1.00	2.24	4.00	3.16	4.00	3.46
dan	3.46	3.16	3.16	3.32	3.16	0.00	3.00	3.00	2.83	3.00	4.00	2.45	3.00	3.32	2.65	2.00	3.74	3.32	3.61	3.00	2.65	2.65	2.65	3.61	3.00	3.87	3.46	2.83	3.00	3.00	4.00	1.41	4.00	3.46
nld	3.87	3.61	3.61	3.74	3.61	3.00	0.00	3.16	3.32	3.46	4.36	2.24	3.46	3.74	3.16	3.00	4.12	3.74	4.00	3.46	3.16	3.16	3.16	4.00	3.46	4.24	3.87	3.32	3.46	3.46	4.36	3.00	4.36	3.87
eng	3.87	3.61	3.61	3.74	3.61	3.00	3.16	0.00	3.32	3.46	4.36	2.65	3.46	3.74	3.16	3.00	4.12	3.74	4.00	3.46	3.16	3.16	3.16	4.00	3.46	4.24	3.87	3.32	3.46	3.46	4.36	3.00	4.36	3.87
est	2.83	2.83	2.45	3.00	2.83	2.83	3.32	3.32	0.00	1.00	3.74	2.83	2.65	3.00	1.00	2.83	3.16	3.00	3.32	2.24	1.73	2.24	2.24	3.00	2.65	3.61	3.16	2.45	2.65	2.65	3.74	2.83	3.46	2.83
fin	3.00	3.00	2.65	3.16	3.00	3.00	3.46	3.46	1.00	0.00	3.87	3.00	2.83	3.16	1.41	3.00	3.32	3.16	3.46	2.45	2.00	2.45	2.45	3.16	2.83	3.74	3.32	2.65	2.83	2.83	3.87	3.00	3.61	3.00
fra	4.24	4.00	4.00	4.12	4.00	4.00	4.36	4.36	3.74	3.87	0.00	4.00	3.87	4.12	3.61	4.00	4.47	4.12	2.65	3.87	3.61	3.61	3.61	4.36	3.87	2.24	2.83	3.74	3.87	3.87	2.45	4.00	4.69	4.24
deu	3.46	3.16	3.16	3.32	3.16	2.45	2.24	2.65	2.83	3.00	4.00	0.00	3.00	3.32	2.65	2.45	3.74	3.32	3.61	3.00	2.65	2.65	2.65	3.61	3.00	3.87	3.46	2.83	3.00	3.00	4.00	2.45	4.00	3.46
ell	3.32	3.00	3.00	3.16	3.00	3.00	3.46	3.46	2.65	2.83	3.87	3.00	0.00	3.16	2.45	3.00	3.61	3.16	3.46	2.83	2.45	2.45	2.45	3.46	2.83	3.74	3.32	2.65	2.83	2.83	3.87	3.00	3.87	3.32
hin	3.61	3.32	3.32	3.46	3.32	3.32	3.74	3.74	3.00	3.16	4.12	3.32	3.16	0.00	2.83	3.32	3.87	3.46	3.74	3.16	2.83	2.83	2.83	3.74	3.16	4.00	3.61	3.00	3.16	3.16	4.12	3.32	4.12	3.61
hun	2.65	2.65	2.24	2.83	2.65	2.65	3.16	3.16	1.00	1.41	3.61	2.65	2.45	2.83	0.00	2.65	3.00	2.83	3.16	2.00	1.41	2.00	2.00	2.83	2.45	3.46	3.00	2.24	2.45	2.45	3.61	2.65	3.32	2.65
isl	3.46	3.16	3.16	3.32	3.16	2.00	3.00	3.00	2.83	3.00	4.00	2.45	3.00	3.32	2.65	0.00	3.74	3.32	3.61	3.00	2.65	2.65	2.65	3.61	3.00	3.87	3.46	2.83	3.00	3.00	4.00	2.00	4.00	3.46
ind	3.74	3.74	3.46	3.87	3.74	3.74	4.12	4.12	3.16	3.32	4.47	3.74	3.61	3.87	3.00	3.74	0.00	3.87	4.12	3.32	3.00	3.32	3.32	3.87	3.61	4.36	4.00	3.46	3.61	3.61	4.47	3.74	4.24	3.74
gle	3.61	3.32	3.32	3.46	3.32	3.32	3.74	3.74	3.00	3.16	4.12	3.32	3.16	3.46	2.83	3.32	3.87	0.00	3.74	3.16	2.83	2.83	2.83	3.74	3.16	4.00	3.61	3.00	3.16	3.16	4.12	3.32	4.12	3.61
ita	3.87	3.61	3.61	3.74	3.61	3.61	4.00	4.00	3.32	3.46	2.65	3.61	3.46	3.74	3.16	3.61	4.12	3.74	0.00	3.46	3.16	3.16	3.16	4.00	3.46	2.45	2.24	3.32	3.46	3.46	2.65	3.61	4.36	3.87
jpn	3.00	3.00	2.65	3.16	3.00	3.00	3.46	3.46	2.24	2.45	3.87	3.00	2.83	3.16	2.00	3.00	3.32	3.16	3.46	0.00	2.00	2.45	2.45	3.16	2.83	3.74	3.32	2.65	2.83	2.83	3.87	3.00	3.61	3.00
kor	2.65	2.65	2.24	2.83	2.65	2.65	3.16	3.16	1.73	2.00	3.61	2.65	2.45	2.83	1.41	2.65	3.00	2.83	3.16	2.00	0.00	2.00	2.00	2.83	2.45	3.46	3.00	2.24	2.45	2.45	3.61	2.65	3.32	2.65
lav	3.00	2.24	2.65	2.45	2.24	2.65	3.16	3.16	2.24	2.45	3.61	2.65	2.45	2.83	2.00	2.65	3.32	2.83	3.16	2.45	2.00	0.00	0.00	3.16	2.00	3.46	3.00	1.73	2.00	2.00	3.61	2.65	3.61	3.00
lit	3.00	2.24	2.65	2.45	2.24	2.65	3.16	3.16	2.24	2.45	3.61	2.65	2.45	2.83	2.00	2.65	3.32	2.83	3.16	2.45	2.00	0.00	0.00	3.16	2.00	3.46	3.00	1.73	2.00	2.00	3.61	2.65	3.61	3.00
mlt	1.00	3.61	3.32	3.74	3.61	3.61	4.00	4.00	3.00	3.16	4.36	3.61	3.46	3.74	2.83	3.61	3.87	3.74	4.00	3.16	2.83	3.16	3.16	0.00	3.46	4.24	3.87	3.32	3.46	3.46	4.36	3.61	4.12	3.61
pol	3.32	2.24	3.00	2.45	1.73	3.00	3.46	3.46	2.65	2.83	3.87	3.00	2.83	3.16	2.45	3.00	3.61	3.16	3.46	2.83	2.45	2.00	2.00	3.46	0.00	3.74	3.32	1.73	1.41	2.00	3.87	3.00	3.87	3.32
por	4.12	3.87	3.87	4.00	3.87	3.87	4.24	4.24	3.61	3.74	2.24	3.87	3.74	4.00	3.46	3.87	4.36	4.00	2.45	3.74	3.46	3.46	3.46	4.24	3.74	0.00	2.65	3.61	3.74	3.74	1.00	3.87	4.58	4.12
ron	3.74	3.46	3.46	3.61	3.46	3.46	3.87	3.87	3.16	3.32	2.83	3.46	3.32	3.61	3.00	3.46	4.00	3.61	2.24	3.32	3.00	3.00	3.00	3.87	3.32	2.65	0.00	3.16	3.32	3.32	2.83	3.46	4.24	3.74
rus	3.16	2.00	2.83	2.24	2.00	2.83	3.32	3.32	2.45	2.65	3.74	2.83	2.65	3.00	2.24	2.83	3.46	3.00	3.32	2.65	2.24	1.73	1.73	3.32	1.73	3.61	3.16	0.00	1.73	1.73	3.74	2.83	3.74	3.16
slk	3.32	2.24	3.00	2.45	1.00	3.00	3.46	3.46	2.65	2.83	3.87	3.00	2.83	3.16	2.45	3.00	3.61	3.16	3.46	2.83	2.45	2.00	2.00	3.46	1.41	3.74	3.32	1.73	0.00	2.00	3.87	3.00	3.87	3.32
slv	3.32	1.73	3.00	1.41	2.24	3.00	3.46	3.46	2.65	2.83	3.87	3.00	2.83	3.16	2.45	3.00	3.61	3.16	3.46	2.83	2.45	2.00	2.00	3.46	2.00	3.74	3.32	1.73	2.00	0.00	3.87	3.00	3.87	3.32
spa	4.24	4.00	4.00	4.12	4.00	4.00	4.36	4.36	3.74	3.87	2.45	4.00	3.87	4.12	3.61	4.00	4.47	4.12	2.65	3.87	3.61	3.61	3.61	4.36	3.87	1.00	2.83	3.74	3.87	3.87	0.00	4.00	4.69	4.24
swe	3.46	3.16	3.16	3.32	3.16	1.41	3.00	3.00	2.83	3.00	4.00	2.45	3.00	3.32	2.65	2.00	3.74	3.32	3.61	3.00	2.65	2.65	2.65	3.61	3.00	3.87	3.46	2.83	3.00	3.00	4.00	0.00	4.00	3.46
tha	4.00	4.00	3.74	4.12	4.00	4.00	4.36	4.36	3.46	3.61	4.69	4.00	3.87	4.12	3.32	4.00	4.24	4.12	4.36	3.61	3.32	3.61	3.61	4.12	3.87	4.58	4.24	3.74	3.87	3.87	4.69	4.00	0.00	4.00
tur	3.46	3.46	3.16	3.61	3.46	3.46	3.87	3.87	2.83	3.00	4.24	3.46	3.32	3.61	2.65	3.46	3.74	3.61	3.87	3.00	2.65	3.00	3.00	3.61	3.32	4.12	3.74	3.16	3.32	3.32	4.24	3.46	4.00	0.00

Table A.73. Euclidean dissimilarity matrix regarding the analysis of the phylogenetic features of EU and PUD languages.

Annex 74.

	arb	bul	cmn	hrv	ces	dan	nld	eng	est	fin	fra	deu	ell	hin	hun	isl	ind	gle	ita	jpn	kor	lav	lit	mlt	pol	por	ron	rus	slk	slv	spa	swe	tha	tur
arb	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.07	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
bul	1.00	0.00	1.00	0.38	0.50	0.83	0.86	0.86	1.00	1.00	0.88	0.83	0.82	0.85	1.00	0.83	1.00	0.85	0.86	1.00	1.00	0.53	0.53	1.00	0.45	0.88	0.86	0.39	0.45	0.27	0.88	0.83	1.00	1.00
cmn	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
hrv	1.00	0.38	1.00	0.00	0.54	0.85	0.87	0.87	1.00	1.00	0.89	0.85	0.83	0.86	1.00	0.85	1.00	0.86	0.87	1.00	1.00	0.56	0.56	1.00	0.49	0.89	0.87	0.43	0.49	0.15	0.89	0.85	1.00	1.00
ces	1.00	0.50	1.00	0.54	0.00	0.83	0.86	0.86	1.00	1.00	0.88	0.83	0.82	0.85	1.00	0.83	1.00	0.85	0.86	1.00	1.00	0.53	0.53	1.00	0.27	0.88	0.86	0.39	0.09	0.45	0.88	0.83	1.00	1.00
dan	1.00	0.83	1.00	0.85	0.83	0.00	0.59	0.59	1.00	1.00	0.88	0.50	0.82	0.85	1.00	0.33	1.00	0.85	0.86	1.00	1.00	0.76	0.76	1.00	0.82	0.88	0.86	0.80	0.82	0.82	0.88	0.17	1.00	1.00
nld	1.00	0.86	1.00	0.87	0.86	0.59	0.00	0.56	1.00	1.00	0.90	0.32	0.85	0.87	1.00	0.59	1.00	0.87	0.89	1.00	1.00	0.81	0.81	1.00	0.85	0.90	0.88	0.83	0.85	0.85	0.90	0.59	1.00	1.00
eng	1.00	0.86	1.00	0.87	0.86	0.59	0.56	0.00	1.00	1.00	0.90	0.46	0.85	0.87	1.00	0.59	1.00	0.87	0.89	1.00	1.00	0.81	0.81	1.00	0.85	0.90	0.88	0.83	0.85	0.85	0.90	0.59	1.00	1.00
est	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.18	1.00	1.00	1.00	1.00	0.29	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
fin	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.18	0.00	1.00	1.00	1.00	1.00	0.42	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
fra	1.00	0.88	1.00	0.89	0.88	0.88	0.90	0.90	1.00	1.00	0.00	0.88	0.87	0.89	1.00	0.88	1.00	0.89	0.33	1.00	1.00	0.83	0.83	1.00	0.87	0.22	0.39	0.86	0.87	0.87	0.25	0.88	1.00	1.00
deu	1.00	0.83	1.00	0.85	0.83	0.50	0.32	0.46	1.00	1.00	0.88	0.00	0.82	0.85	1.00	0.50	1.00	0.85	0.86	1.00	1.00	0.76	0.76	1.00	0.82	0.88	0.86	0.80	0.82	0.82	0.88	0.50	1.00	1.00
ell	1.00	0.82	1.00	0.83	0.82	0.82	0.85	0.85	1.00	1.00	0.87	0.82	0.00	0.83	1.00	0.82	1.00	0.83	0.85	1.00	1.00	0.74	0.74	1.00	0.80	0.87	0.84	0.78	0.80	0.80	0.87	0.82	1.00	1.00
hin	1.00	0.85	1.00	0.86	0.85	0.85	0.87	0.87	1.00	1.00	0.89	0.85	0.83	0.00	1.00	0.85	1.00	0.86	0.87	1.00	1.00	0.78	0.78	1.00	0.83	0.89	0.87	0.81	0.83	0.83	0.89	0.85	1.00	1.00
hun	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.29	0.42	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
isl	1.00	0.83	1.00	0.85	0.83	0.33	0.59	0.59	1.00	1.00	0.88	0.50	0.82	0.85	1.00	0.00	1.00	0.85	0.86	1.00	1.00	0.76	0.76	1.00	0.82	0.88	0.86	0.80	0.82	0.82	0.88	0.33	1.00	1.00
ind	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
gle	1.00	0.85	1.00	0.86	0.85	0.85	0.87	0.87	1.00	1.00	0.89	0.85	0.83	0.86	1.00	0.85	1.00	0.00	0.87	1.00	1.00	0.78	0.78	1.00	0.83	0.89	0.87	0.81	0.83	0.83	0.89	0.85	1.00	1.00
ita	1.00	0.86	1.00	0.87	0.86	0.86	0.89	0.89	1.00	1.00	0.33	0.86	0.85	0.87	1.00	0.86	1.00	0.87	0.00	1.00	1.00	0.81	0.81	1.00	0.85	0.30	0.29	0.83	0.85	0.85	0.33	0.86	1.00	1.00
jpn	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
kor	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
lav	1.00	0.53	1.00	0.56	0.53	0.76	0.81	0.81	1.00	1.00	0.83	0.76	0.74	0.78	1.00	0.76	1.00	0.78	0.81	1.00	1.00	0.00	0.00	1.00	0.48	0.83	0.80	0.42	0.48	0.48	0.83	0.76	1.00	1.00
lit	1.00	0.53	1.00	0.56	0.53	0.76	0.81	0.81	1.00	1.00	0.83	0.76	0.74	0.78	1.00	0.76	1.00	0.78	0.81	1.00	1.00	0.00	0.00	1.00	0.48	0.83	0.80	0.42	0.48	0.48	0.83	0.76	1.00	1.00
mlt	0.07	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
pol	1.00	0.45	1.00	0.49	0.27	0.82	0.85	0.85	1.00	1.00	0.87	0.82	0.80	0.83	1.00	0.82	1.00	0.83	0.85	1.00	1.00	0.48	0.48	1.00	0.00	0.87	0.84	0.33	0.20	0.40	0.87	0.82	1.00	1.00
por	1.00	0.88	1.00	0.89	0.88	0.88	0.90	0.90	1.00	1.00	0.22	0.88	0.87	0.89	1.00	0.88	1.00	0.89	0.30	1.00	1.00	0.83	0.83	1.00	0.87	0.00	0.36	0.85	0.87	0.87	0.04	0.88	1.00	1.00
ron	1.00	0.86	1.00	0.87	0.86	0.86	0.88	0.88	1.00	1.00	0.39	0.86	0.84	0.87	1.00	0.86	1.00	0.87	0.29	1.00	1.00	0.80	0.80	1.00	0.84	0.36	0.00	0.82	0.84	0.84	0.39	0.86	1.00	1.00
rus	1.00	0.39	1.00	0.43	0.39	0.80	0.83	0.83	1.00	1.00	0.86	0.80	0.78	0.81	1.00	0.80	1.00	0.81	0.83	1.00	1.00	0.42	0.42	1.00	0.33	0.85	0.82	0.00	0.33	0.33	0.86	0.80	1.00	1.00
slk	1.00	0.45	1.00	0.49	0.09	0.82	0.85	0.85	1.00	1.00	0.87	0.82	0.80	0.83	1.00	0.82	1.00	0.83	0.85	1.00	1.00	0.48	0.48	1.00	0.20	0.87	0.84	0.33	0.00	0.40	0.87	0.82	1.00	1.00
slv	1.00	0.27	1.00	0.15	0.45	0.82	0.85	0.85	1.00	1.00	0.87	0.82	0.80	0.83	1.00	0.82	1.00	0.83	0.85	1.00	1.00	0.48	0.48	1.00	0.40	0.87	0.84	0.33	0.40	0.00	0.87	0.82	1.00	1.00
spa	1.00	0.88	1.00	0.89	0.88	0.88	0.90	0.90	1.00	1.00	0.25	0.88	0.87	0.89	1.00	0.88	1.00	0.89	0.33	1.00	1.00	0.83	0.83	1.00	0.87	0.04	0.39	0.86	0.87	0.87	0.00	0.88	1.00	1.00
swe	1.00	0.83	1.00	0.85	0.83	0.17	0.59	0.59	1.00	1.00	0.88	0.50	0.82	0.85	1.00	0.33	1.00	0.85	0.86	1.00	1.00	0.76	0.76	1.00	0.82	0.88	0.86	0.80	0.82	0.82	0.88	0.00	1.00	1.00
tha	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00
tur	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00

Table A.74. Euclidean dissimilarity matrix regarding the analysis of the phylogenetic features of EU and PUD languages.

Annex 75.

Language	Test set	LAS	MLAS
bul	BTB	92.40	83.43
hrv	SET	89.79	72.72
ces	PDT	92.88	87.13
dan	DDT	84.50	73.76
nld	Alpino	91.21	82.81
eng	EWT	88.50	79.80
est	EDT	86.67	79.20
fin	PUD	86.58	77.83
fra	GSD	91.45	81.61
deu	GSD	83.59	61.27
ell	GDT	92.15	77.89
hun	Szeged	84.88	64.27
gle	IDT	69.28	34.39
ita	ParTUT	93.68	86.83
lav	LVTB	85.09	69.51
lit	HSE	69.34	36.21
mlt	MUDT	75.56	58.14
pol	LFG	94.58	76.50
por	GSD	92.54	85.96
ron	RRT	88.56	79.20
slk	SNK	93.81	77.33
slv	SSJ	93.07	81.55
spa	AnCora	90.50	83.43
swe	Talbanken	89.03	80.72

Table A.75. UDify LAS and MLAS scores obtained by Kondratyuk and Straka (2019) for the EU languages using a multilingual parsing model trained with 124 corpora from 75 languages.

Annex 76.

Language	Corpus (UD v.2.7)	Number of Tokens
arb	PUD	20,751
bul	BTB	14,186
cmn	PUD	21,415
hrv	SET	21,887
ces	PUD	18,565
dan	DDT	18,542
nld	Alpino	14,982
eng	PUD	21,176
est	EDT	14,085
fin	PUD	15,807
fra	PUD	24,137
deu	PUD	21,001
ell	GDT	26,464
hin	PUD	23,829
hun	Szeged	22,396
isl	PUD	18,831
ind	PUD	19,030
gle	IDT	24,043
ita	PUD	22,182
jpn	PUD	28,784
kor	PUD	16,584
lav	LVTB	16,166
lit	ALKSNIS	20,353
mlt	MUDT	20,068
pol	PUD	18,338
por	PUD	21,917
ron	RRT	23,031
rus	PUD	19,355
slk	SNK	9,582
slv	SSJ	19,085
spa	PUD	22,822
swe	PUD	19,076
tha	PUD	22,322
tur	PUD	16,536

Table A.76. Information regarding source and number of tokens of the corpora built for the typological experiments regarding EU and PUD languages. For PUD languages, the typological corpus concerns all 1,000 sentences provided in this collection. For the other languages, 1,000 sentences were randomly selected from the training-set of the mentioned corpus (Hungarian being the exception, with 90 sentences from its development-set).

Annex 77.

Dependency relation labels			
acl	cc	expl:pv	obj:lvc
acl:cleft	cc:preconj	expl:subj	obl
acl:relcl	ccomp	fixed	obl:agent
advcl	ccomp:cleft	flat	obl:arg
advcl:relcl	ccomp:obj	flat:foreign	obl:cmpr
advcl:tcl	ccomp:obl	flat:name	obl:loc
advmod	ccomp:pmod	goeswith	obl:mod
advmod:arg	ccomp:pred	iobj	obl:npmod
advmod:emph	clf	list	obl:patient
advmod:locy	compound	mark	obl:poss
advmod:mode	compound:a	mark:adv	obl:prep
advmod:neg	compound:conjv	mark:prt	obl:tmod
advmod:obl	compound:lvc	mark:relcl	orphan
advmod:que	compound:nn	nmod	parataxis
advmod:tfrom	compound:preverb	nmod:agent	parataxis:insert
advmod:tlocy	compound:prt	nmod:arg	parataxis:obj
advmod:tmod	compound:redup	nmod:att	punct
advmod:to	conj	nmod:attlvc	reparandum
advmod:tto	cop	nmod:flat	root
amod	cop:expl	nmod:gmod	vocative
amod:att	cop:own	nmod:gobj	xcomp
amod:attlvc	csubj	nmod:gsubj	xcomp:ds
amod:flat	csubj:cleft	nmod:lmod	xcomp:pred
amod:mode	csubj:cop	nmod:npmod	xcomp:subj
amod:obl	csubj:pass	nmod:obl	
appos	dep	nmod:obllvc	
aux	dep:prt	nmod:pmod	
aux:caus	det	nmod:poss	
aux:clitic	det:numgov	nmod:pred	
aux:cnd	det:nummod	nmod:tmod	
aux:neg	det:poss	nsubj	
aux:part	det:predet	nsubj:caus	
aux:pass	discourse	nsubj:cop	
aux:q	discourse:sp	nsubj:lvc	
aux:tense	dislocated	nsubj:pass	
case	expl	nummod	
case:adv	expl:comp	nummod:entity	
case:det	expl:impers	nummod:gov	
case:loc	expl:pass	obj	
case:voc	expl:poss	obj:agent	

Table A.77. List of the dependency relation labels found in the collection composed of the typological corpora of PUD and EU languages.

Annex 78.

Language	Number of patterns
arb	2,208
bul	1,348
cmn	2,552
hrv	2,612
ces	2,053
dan	2,023
nld	2,422
eng	2,599
est	1,541
fin	1,764
fra	1,928
deu	1,826
ell	1,853
hin	2,842
hun	2,356
isl	2,710
ind	1,664
gle	2,850
ita	2,090
jpn	1,287
kor	1,418
lav	1,840
lit	1,868
mlt	2,794
pol	2,257
por	2,023
ron	2,506
rus	2,072
slk	1,049
slv	1,288
spa	1,996
swe	2,508
tha	2,665
tur	2,144

Table A.78. Number of linear patterns extracted using MarsaGram toll for each PUD and EU languages.

Annex 79.

	arb	bul	cmn	hrv	ces	dan	nld	eng	est	fin	fra	deu	ell	hin	hun	isl	ind	gle	ita	jpn	kor	lav	lit	mlt	pol	por	ron	rus	slk	slv	spa	swe	tha	tur
arb	0.00	0.60	0.90	0.76	0.76	0.81	0.75	0.68	0.95	0.96	0.66	0.79	0.73	0.90	0.98	0.88	0.81	0.78	0.69	0.99	0.97	0.86	0.89	0.84	0.75	0.68	0.76	0.74	0.69	0.73	0.74	0.68	0.70	0.97
bul	0.60	0.00	0.73	0.64	0.64	0.73	0.64	0.55	0.86	0.89	0.59	0.69	0.65	0.90	0.95	0.84	0.64	0.77	0.61	0.98	0.96	0.77	0.82	0.76	0.68	0.60	0.66	0.62	0.55	0.61	0.65	0.37	0.69	0.92
cmn	0.90	0.73	0.00	0.90	0.93	0.91	0.93	0.82	0.91	0.94	0.86	0.90	0.89	0.96	0.93	0.92	0.87	0.91	0.88	0.87	0.81	0.91	0.94	0.88	0.93	0.88	0.86	0.89	0.91	0.92	0.90	0.75	0.86	0.82
hrv	0.76	0.64	0.90	0.00	0.60	0.81	0.73	0.54	0.66	0.75	0.67	0.55	0.71	0.83	0.95	0.81	0.70	0.82	0.71	0.98	0.98	0.62	0.62	0.56	0.67	0.58	0.73	0.53	0.59	0.50	0.59	0.67	0.66	0.85
ces	0.76	0.64	0.93	0.60	0.00	0.82	0.73	0.62	0.81	0.84	0.63	0.45	0.79	0.89	0.95	0.80	0.84	0.78	0.70	0.99	0.98	0.71	0.85	0.77	0.63	0.67	0.77	0.52	0.59	0.64	0.62	0.63	0.72	0.90
dan	0.81	0.73	0.91	0.81	0.82	0.00	0.71	0.69	0.88	0.92	0.70	0.76	0.74	0.96	0.88	0.89	0.81	0.83	0.71	0.98	0.94	0.86	0.89	0.86	0.82	0.70	0.70	0.81	0.78	0.80	0.74	0.72	0.87	0.93
nld	0.75	0.64	0.93	0.73	0.73	0.71	0.00	0.60	0.89	0.94	0.54	0.66	0.61	0.93	0.88	0.86	0.75	0.79	0.52	0.93	0.94	0.86	0.88	0.82	0.73	0.55	0.78	0.72	0.64	0.72	0.56	0.67	0.83	0.92
eng	0.68	0.55	0.82	0.54	0.62	0.69	0.60	0.00	0.77	0.81	0.40	0.54	0.49	0.86	0.87	0.75	0.76	0.73	0.45	0.97	0.91	0.65	0.82	0.60	0.69	0.48	0.64	0.56	0.58	0.60	0.51	0.49	0.71	0.83
est	0.95	0.86	0.91	0.66	0.81	0.88	0.89	0.77	0.00	0.60	0.73	0.64	0.88	0.81	0.91	0.87	0.88	0.87	0.83	0.96	0.92	0.81	0.88	0.89	0.88	0.89	0.86	0.60	0.90	0.86	0.77	0.86	0.80	0.82
fin	0.96	0.89	0.94	0.75	0.84	0.92	0.94	0.81	0.60	0.00	0.78	0.74	0.94	0.83	0.89	0.82	0.86	0.89	0.91	0.96	0.93	0.78	0.92	0.90	0.93	0.91	0.94	0.66	0.91	0.88	0.80	0.84	0.80	0.94
fra	0.66	0.59	0.86	0.67	0.63	0.70	0.54	0.40	0.73	0.78	0.00	0.52	0.52	0.87	0.84	0.78	0.72	0.63	0.41	0.97	0.90	0.78	0.84	0.68	0.65	0.36	0.70	0.44	0.62	0.66	0.46	0.51	0.68	0.89
deu	0.79	0.69	0.90	0.55	0.45	0.76	0.66	0.54	0.64	0.74	0.52	0.00	0.68	0.81	0.86	0.75	0.77	0.83	0.58	0.97	0.92	0.76	0.85	0.71	0.71	0.61	0.81	0.48	0.74	0.69	0.48	0.64	0.61	0.87
ell	0.73	0.65	0.89	0.71	0.79	0.74	0.61	0.49	0.88	0.94	0.52	0.68	0.00	0.86	0.77	0.88	0.84	0.77	0.47	0.99	0.96	0.85	0.87	0.72	0.82	0.58	0.72	0.78	0.74	0.74	0.61	0.70	0.78	0.92
hin	0.90	0.90	0.96	0.83	0.89	0.96	0.93	0.86	0.81	0.83	0.87	0.81	0.86	0.00	0.90	0.92	0.95	0.92	0.88	0.70	0.89	0.89	0.96	0.95	0.93	0.82	0.86	0.84	0.93	0.94	0.75	0.90	0.76	0.84
hun	0.98	0.95	0.93	0.95	0.95	0.88	0.88	0.87	0.91	0.89	0.84	0.86	0.77	0.90	0.00	0.95	0.93	0.95	0.88	0.95	0.92	0.86	0.95	0.86	0.98	0.88	0.93	0.92	0.95	0.92	0.90	0.90	0.96	0.93
isl	0.88	0.84	0.92	0.81	0.80	0.89	0.86	0.75	0.87	0.82	0.78	0.75	0.88	0.92	0.95	0.00	0.90	0.91	0.81	0.99	0.96	0.81	0.93	0.81	0.80	0.85	0.88	0.73	0.88	0.87	0.79	0.82	0.79	0.92
ind	0.81	0.64	0.87	0.70	0.84	0.81	0.75	0.76	0.88	0.86	0.72	0.77	0.84	0.95	0.93	0.90	0.00	0.76	0.71	0.88	0.93	0.84	0.74	0.74	0.84	0.67	0.72	0.77	0.82	0.80	0.74	0.75	0.70	0.88
gle	0.78	0.77	0.91	0.82	0.78	0.83	0.79	0.73	0.87	0.89	0.63	0.83	0.77	0.92	0.95	0.91	0.76	0.00	0.73	0.95	0.99	0.90	0.93	0.86	0.81	0.74	0.74	0.76	0.83	0.83	0.78	0.82	0.71	0.87
ita	0.69	0.61	0.88	0.71	0.70	0.71	0.52	0.45	0.83	0.91	0.41	0.58	0.47	0.88	0.88	0.81	0.71	0.73	0.00	0.93	0.94	0.81	0.87	0.74	0.71	0.51	0.74	0.66	0.71	0.75	0.36	0.66	0.72	0.93
jpn	0.99	0.98	0.87	0.98	0.99	0.98	0.93	0.97	0.96	0.96	0.97	0.97	0.99	0.70	0.95	0.99	0.88	0.95	0.93	0.00	0.85	0.99	0.98	0.94	1.00	0.97	0.96	0.99	0.98	0.95	0.95	0.98	0.91	0.82
kor	0.97	0.96	0.81	0.98	0.98	0.94	0.94	0.91	0.92	0.93	0.90	0.92	0.96	0.89	0.92	0.96	0.93	0.99	0.94	0.85	0.00	0.92	0.97	0.95	0.99	0.92	0.96	0.95	0.97	0.95	0.93	0.91	0.95	0.85
lav	0.86	0.77	0.91	0.62	0.71	0.86	0.86	0.65	0.81	0.78	0.78	0.76	0.85	0.89	0.86	0.81	0.84	0.90	0.81	0.99	0.92	0.00	0.86	0.75	0.81	0.82	0.85	0.74	0.71	0.60	0.79	0.71	0.82	0.89
lit	0.89	0.82	0.94	0.62	0.85	0.89	0.88	0.82	0.88	0.92	0.84	0.85	0.87	0.96	0.95	0.93	0.74	0.93	0.87	0.98	0.97	0.86	0.00	0.79	0.88	0.67	0.90	0.86	0.79	0.81	0.86	0.86	0.92	0.96
mlt	0.84	0.76	0.88	0.56	0.77	0.86	0.82	0.60	0.89	0.90	0.68	0.71	0.72	0.95	0.86	0.81	0.74	0.86	0.74	0.94	0.95	0.75	0.79	0.00	0.84	0.70	0.78	0.72	0.77	0.57	0.77	0.72	0.79	0.63
pol	0.75	0.68	0.93	0.67	0.63	0.82	0.73	0.69	0.88	0.93	0.65	0.71	0.82	0.93	0.98	0.80	0.84	0.81	0.71	1.00	0.99	0.81	0.88	0.84	0.00	0.70	0.75	0.61	0.62	0.70	0.68	0.73	0.75	0.96
por	0.68	0.60	0.88	0.58	0.67	0.70	0.55	0.48	0.89	0.91	0.36	0.61	0.58	0.82	0.88	0.85	0.67	0.74	0.51	0.97	0.92	0.82	0.67	0.70	0.70	0.00	0.67	0.63	0.60	0.70	0.35	0.57	0.73	0.93
ron	0.76	0.66	0.86	0.73	0.77	0.70	0.78	0.64	0.86	0.94	0.70	0.81	0.72	0.86	0.93	0.88	0.72	0.74	0.74	0.96	0.96	0.85	0.90	0.78	0.75	0.67	0.00	0.75	0.71	0.78	0.69	0.69	0.72	0.85
rus	0.74	0.62	0.89	0.53	0.52	0.81	0.72	0.56	0.60	0.66	0.44	0.48	0.78	0.84	0.92	0.73	0.77	0.76	0.66	0.99	0.95	0.74	0.86	0.72	0.61	0.63	0.75	0.00	0.59	0.67	0.53	0.64	0.53	0.89
slk	0.69	0.55	0.91	0.59	0.59	0.78	0.64	0.58	0.90	0.91	0.62	0.74	0.74	0.93	0.95	0.88	0.82	0.83	0.71	0.98	0.97	0.71	0.79	0.77	0.62	0.60	0.71	0.59	0.00	0.50	0.63	0.57	0.78	0.89
slv	0.73	0.61	0.92	0.50	0.64	0.80	0.72	0.60	0.86	0.88	0.66	0.69	0.74	0.94	0.92	0.87	0.80	0.83	0.75	0.95	0.95	0.60	0.81	0.57	0.70	0.70	0.78	0.67	0.50	0.00	0.71	0.63	0.80	0.90
spa	0.74	0.65	0.90	0.59	0.62	0.74	0.56	0.51	0.77	0.80	0.46	0.48	0.61	0.75	0.90	0.79	0.74	0.78	0.36	0.95	0.93	0.79	0.86	0.77	0.68	0.35	0.69	0.53	0.63	0.71	0.00	0.65	0.66	0.94
swe	0.68	0.37	0.75	0.67	0.63	0.72	0.67	0.49	0.86	0.84	0.51	0.64	0.70	0.90	0.90	0.82	0.75	0.82	0.66	0.98	0.91	0.71	0.86	0.72	0.73	0.57	0.69	0.64	0.57	0.63	0.65	0.00	0.74	0.89
tha	0.70	0.69	0.86	0.66	0.72	0.87	0.83	0.71	0.80	0.80	0.68	0.61	0.78	0.76	0.96	0.79	0.70	0.71	0.72	0.91	0.95	0.82	0.92	0.79	0.75	0.73	0.72	0.53	0.78	0.80	0.66	0.74	0.00	0.84
tur	0.97	0.92	0.82	0.85	0.90	0.93	0.92	0.83	0.82	0.94	0.89	0.87	0.92	0.84	0.93	0.92	0.88	0.87	0.93	0.82	0.85	0.89	0.96	0.63	0.96	0.93	0.85	0.89	0.89	0.90	0.94	0.89	0.84	0.00

Table A.79. Cosine dissimilarity matrix for EU and PUD languages regarding MarsaGram linear properties.

Annex 80.

Language	Number of patterns
arb	16,460
bul	8,439
cmn	18,070
hrv	20,560
ces	16,706
dan	12,839
nld	20,709
eng	20,517
est	10,563
fin	13,374
fra	13,656
deu	13,225
ell	11,972
hin	22,106
hun	21,900
isl	22,199
ind	10,889
gle	25,989
ita	15,380
jpn	5,226
kor	8,860
lav	14,456
lit	14,643
mlt	22,206
pol	17,592
por	13,994
ron	21,193
rus	16,827
slk	8,283
slv	7,220
spa	14,021
swe	19,795
tha	19,403
tur	17,508

Table A.80. Number of all possible patterns extracted using MarsaGram toll for each PUD and EU languages.

Annex 81.

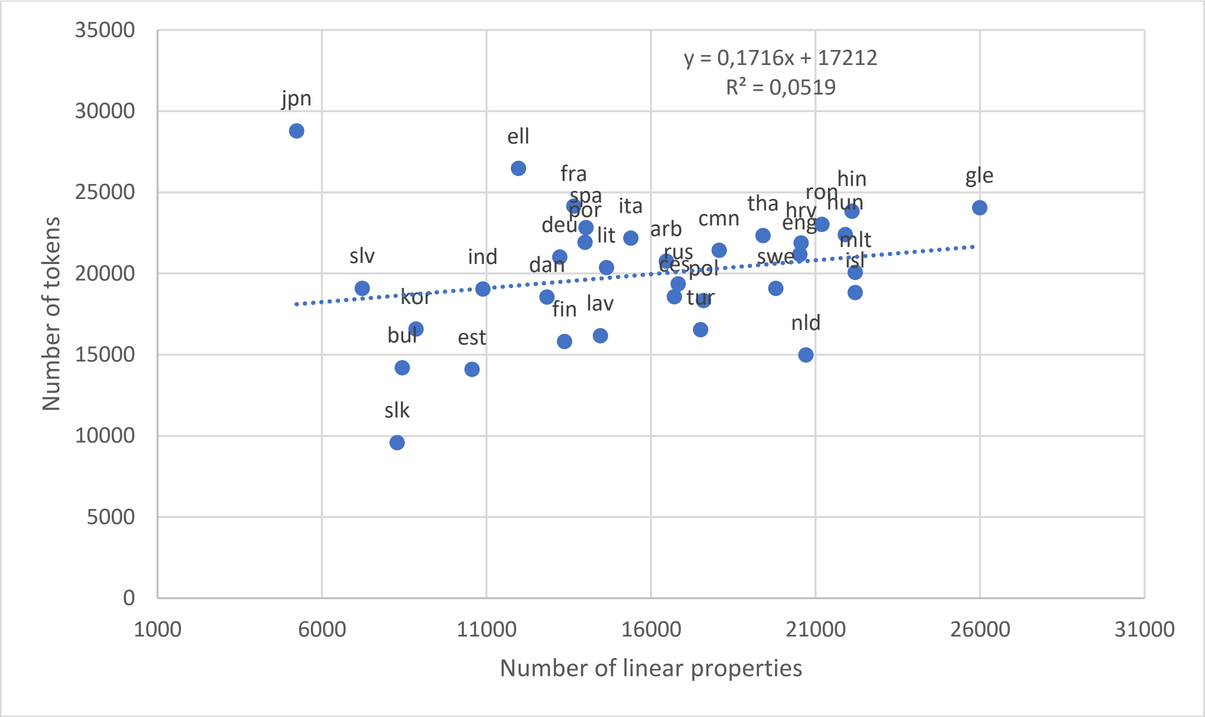


Figure A.1. Graph representing the number of extracted patterns (all properties) in relation with the corpora size.

Annex 82.

MarsaGram Pattern
NOUN-+_exclude_AUX-cop_NOUN-appos
NOUN-+_exclude_CCONJ-cc_NOUN-appos
NOUN-+_exclude_NUM-nummod_NOUN-appos
NOUN-+_precede_*_NOUN-appos
NOUN-+_unicity_AUX-cop_-
NOUN-+_unicity_CCONJ-cc_-
NOUN-+_unicity_NOUN-appos_-
PROPN-+_unicity_AUX-cop_-
PROPN-+_unicity_CCONJ-cc_-
VERB-+_exclude_ADV-advmod_NOUN-advcl
VERB-+_exclude_ADV-advmod_PROPN-obj
VERB-+_exclude_ADV-advmod_VERB-ccomp
VERB-+_exclude_NOUN-nsubj_PROPN-nsubj
VERB-+_exclude_NOUN-nsubj_PROPN-obj
VERB-+_exclude_NOUN-obj_PRON-obj
VERB-+_exclude_NOUN-obj_PROPN-obj
VERB-+_exclude_NOUN-obj_VERB-ccomp
VERB-+_exclude_PRON-nsubj_PROPN-obj
VERB-+_exclude_PROPN-nsubj_PROPN-obj
VERB-+_precede_CCONJ-cc_*
VERB-+_unicity_CCONJ-cc_-
VERB-+_unicity_NOUN-advcl_-
VERB-+_unicity_NOUN-nsubj_-
VERB-+_unicity_NOUN-obj_-
VERB-+_unicity_PRON-nsubj_-
VERB-+_unicity_PRON-obj_-
VERB-+_unicity_PROPN-nsubj_-
VERB-+_unicity_PROPN-obj_-
VERB-+_unicity_VERB-ccomp_-

Table A.81. List of MarsaGram patterns present in all PUD and EU corpora.

Annex 83.

	arb	bul	cmn	hrv	ces	dan	nld	eng	est	fin	fra	deu	ell	hin	hun	isl	ind	gle	ita	jpn	kor	lav	lit	mlt	pol	por	ron	rus	slk	slv	spa	swe	tha	tur
arb	0.0	17.1	19.2	19.2	18.6	18.6	19.8	19.0	17.5	18.9	18.9	18.9	18.3	20.4	21.1	19.6	17.4	18.9	19.5	18.7	18.0	18.6	18.4	19.4	19.1	18.9	18.3	18.9	17.2	17.7	19.3	17.8	17.4	17.9
bul	17.1	0.0	18.3	17.1	17.2	17.7	18.6	17.7	16.2	17.6	18.3	17.7	16.8	20.4	20.0	18.7	16.3	18.9	18.5	17.9	17.2	17.0	17.1	18.1	17.9	18.6	16.8	16.9	15.4	15.2	18.4	16.7	17.0	17.5
cmn	19.2	18.3	0.0	20.1	19.9	19.3	20.2	19.8	18.0	18.9	19.7	19.4	18.7	21.4	20.8	20.5	17.8	20.3	20.9	19.4	16.7	18.7	17.9	19.3	19.9	20.4	19.3	20.0	17.8	18.0	20.5	19.2	17.8	18.1
hrv	19.2	17.1	20.1	0.0	18.2	19.4	19.7	19.1	17.4	18.9	19.9	19.3	18.2	21.4	21.3	20.0	18.5	20.2	19.6	19.6	19.3	17.7	18.1	19.1	19.5	19.9	18.2	18.2	17.5	15.8	19.7	18.2	19.2	18.0
ces	18.6	17.2	19.9	18.2	0.0	19.8	19.8	18.9	17.1	18.6	19.4	18.6	18.3	21.6	21.2	19.3	17.9	20.1	19.5	19.6	18.8	18.3	17.5	19.1	17.7	19.3	18.5	18.1	16.0	16.9	19.3	18.0	18.7	18.1
dan	18.6	17.7	19.3	19.4	19.8	0.0	19.1	18.8	17.7	18.9	19.0	19.2	18.0	21.3	21.2	19.7	17.9	19.6	20.0	19.4	18.3	19.1	18.5	19.2	19.5	19.7	18.4	19.6	17.6	17.8	19.5	18.2	18.2	18.8
nld	19.8	18.6	20.2	19.7	19.8	19.1	0.0	18.5	18.4	19.7	18.6	18.9	17.5	21.6	21.6	20.4	18.3	19.5	18.9	20.7	19.7	19.3	19.2	19.6	20.7	19.4	18.6	19.9	18.8	18.5	19.8	17.7	19.6	19.0
eng	19.0	17.7	19.8	19.1	18.9	18.8	18.5	0.0	18.2	19.0	17.9	17.1	17.7	20.4	21.6	19.1	16.9	19.8	17.4	19.8	19.2	19.1	19.3	19.1	20.0	17.3	17.9	18.5	18.4	17.9	18.5	16.2	18.1	19.0
est	17.5	16.2	18.0	17.4	17.1	17.7	18.4	18.2	0.0	14.9	17.8	17.5	16.8	21.0	19.2	18.5	16.2	18.4	18.9	18.1	16.5	16.4	16.3	17.6	18.3	18.2	17.6	17.7	16.1	15.7	18.6	17.2	17.1	16.2
fin	18.9	17.6	18.9	18.9	18.6	18.9	19.7	19.0	14.9	0.0	18.9	18.4	18.2	21.3	20.1	18.6	17.1	19.7	20.1	19.4	17.4	17.6	17.8	18.5	19.5	19.7	18.9	18.4	17.4	17.1	19.7	17.6	17.7	17.4
fra	18.9	18.3	19.7	19.9	19.4	19.0	18.6	17.9	17.8	18.9	0.0	18.2	17.3	21.3	21.3	20.1	17.4	19.5	17.8	20.1	19.1	18.9	19.0	18.8	19.8	17.2	18.9	18.6	18.3	18.2	17.5	17.8	18.4	19.2
deu	18.9	17.7	19.4	19.3	18.6	19.2	18.9	17.1	17.5	18.4	18.2	0.0	17.3	20.7	21.1	19.8	17.2	19.9	18.2	20.2	19.1	18.6	18.3	18.9	19.7	18.2	18.9	18.4	17.6	17.4	18.4	17.6	18.9	18.4
ell	18.3	16.8	18.7	18.2	18.3	18.0	17.5	17.7	16.8	18.2	17.3	17.3	0.0	21.0	19.9	19.4	16.7	19.1	17.4	19.0	17.9	17.8	17.4	17.7	19.5	17.9	17.4	18.3	16.8	16.7	18.0	17.3	18.0	17.9
hin	20.4	20.4	21.4	21.4	21.6	21.3	21.6	20.4	21.0	21.3	21.3	20.7	21.0	0.0	23.6	21.8	19.9	21.8	21.2	21.0	20.8	21.4	21.2	21.5	21.9	20.6	20.7	21.4	20.5	19.8	21.2	19.9	19.9	20.9
hun	21.1	20.0	20.8	21.3	21.2	21.2	21.6	21.6	19.2	20.1	21.3	21.1	19.9	23.6	0.0	22.1	19.7	21.6	22.2	21.2	19.1	19.9	19.8	20.5	21.3	21.7	21.1	21.4	19.5	19.5	21.7	21.1	20.2	19.5
isl	19.6	18.7	20.5	20.0	19.3	19.7	20.4	19.1	18.5	18.6	20.1	19.8	19.4	21.8	22.1	0.0	18.4	20.7	20.3	20.3	19.5	19.4	19.6	19.6	20.5	20.1	19.6	19.6	18.8	18.5	20.7	17.5	18.8	19.3
ind	17.4	16.3	17.8	18.5	17.9	17.9	18.3	16.9	16.2	17.1	17.4	17.2	16.7	19.9	19.7	18.4	0.0	18.3	18.1	18.0	16.2	17.1	16.7	17.3	18.2	17.5	17.7	17.5	16.2	16.1	18.2	16.7	16.3	17.2
gle	18.9	18.9	20.3	20.2	20.1	19.6	19.5	19.8	18.4	19.7	19.5	19.9	19.1	21.8	21.6	20.7	18.3	0.0	19.8	20.0	19.0	19.7	19.4	20.1	20.5	19.8	19.3	20.1	18.8	19.0	20.1	19.3	18.9	19.1
ita	19.5	18.5	20.9	19.6	19.5	20.0	18.9	17.4	18.9	20.1	17.8	18.2	17.4	21.2	22.2	20.3	18.1	19.8	0.0	20.6	20.3	19.8	19.8	19.8	20.6	18.0	19.1	19.5	19.0	18.5	18.1	18.4	19.1	19.8
jpn	18.7	17.9	19.4	19.6	19.6	19.4	20.7	19.8	18.1	19.4	20.1	20.2	19.0	21.0	21.2	20.3	18.0	20.0	20.6	0.0	17.6	19.6	19.0	19.3	20.2	20.4	19.3	20.3	18.1	17.7	20.5	18.9	17.9	18.8
kor	18.0	17.2	16.7	19.3	18.8	18.3	19.7	19.2	16.5	17.4	19.1	19.1	17.9	20.8	19.1	19.5	16.2	19.0	20.3	17.6	0.0	17.8	16.8	17.9	18.7	19.4	18.5	19.3	16.7	16.8	19.5	18.4	15.9	16.8
lav	18.6	17.0	18.7	17.7	18.3	19.1	19.3	19.1	16.4	17.6	18.9	18.6	17.8	21.4	19.9	19.4	17.1	19.7	19.8	19.6	17.8	0.0	17.0	18.2	19.5	19.4	18.0	17.8	16.8	16.3	19.6	18.0	18.3	17.6
lit	18.4	17.1	17.9	18.1	17.5	18.5	19.2	19.3	16.3	17.8	19.0	18.3	17.4	21.2	19.8	19.6	16.7	19.4	19.8	19.0	16.8	17.0	0.0	17.8	17.9	19.6	18.4	18.4	15.0	16.2	19.5	18.6	17.8	17.4
mlt	19.4	18.1	19.3	19.1	19.1	19.2	19.6	19.1	17.6	18.5	18.8	18.9	17.7	21.5	20.5	19.6	17.3	20.1	19.8	19.3	17.9	18.2	17.8	0.0	19.7	19.7	19.0	19.2	17.6	17.5	19.6	18.5	18.1	18.2
pol	19.1	17.9	19.9	19.5	17.7	19.5	20.7	20.0	18.3	19.5	19.8	19.7	19.5	21.9	21.3	20.5	18.2	20.5	20.6	20.2	18.7	19.5	17.9	19.7	0.0	20.3	19.5	19.3	16.8	17.6	20.1	19.6	19.0	18.9
por	18.9	18.6	20.4	19.9	19.3	19.7	19.4	17.3	18.2	19.7	17.2	18.2	17.9	20.6	21.7	20.1	17.5	19.8	18.0	20.4	19.4	19.4	19.6	19.7	20.3	0.0	18.7	19.3	18.6	18.4	16.0	18.6	18.7	19.5
ron	18.3	16.8	19.3	18.2	18.5	18.4	18.6	17.9	17.6	18.9	18.9	18.9	17.4	20.7	21.1	19.6	17.7	19.3	19.1	19.3	18.5	18.0	18.4	19.0	19.5	18.7	0.0	18.3	17.6	17.2	18.9	17.2	18.1	18.5
rus	18.9	16.9	20.0	18.2	18.1	19.6	19.9	18.5	17.7	18.4	18.6	18.4	18.3	21.4	21.4	19.6	17.5	20.1	19.5	20.3	19.3	17.8	18.4	19.2	19.3	19.3	18.3	0.0	18.0	17.6	19.5	17.4	18.7	18.7
slk	17.2	15.4	17.8	17.5	16.0	17.6	18.8	18.4	16.1	17.4	18.3	17.6	16.8	20.5	19.5	18.8	16.2	18.8	19.0	18.1	16.7	16.8	15.0	17.6	16.8	18.6	17.6	18.0	0.0	14.9	18.4	17.6	16.9	17.0
slv	17.7	15.2	18.0	15.8	16.9	17.8	18.5	17.9	15.7	17.1	18.2	17.4	16.7	19.8	19.5	18.5	16.1	19.0	18.5	17.7	16.8	16.3	16.2	17.5	17.6	18.4	17.2	17.6	14.9	0.0	18.1	16.9	17.2	17.2
spa	19.3	18.4	20.5	19.7	19.3	19.5	19.8	18.5	18.6	19.7	17.5	18.4	18.0	21.2	21.7	20.7	18.2	20.1	18.1	20.5	19.5	19.6	19.5	19.6	20.1	16.0	18.9	19.5	18.4	18.1	0.0	18.9	18.8	19.7
swe	17.8	16.7	19.2	18.2	18.0	18.2	17.7	16.2	17.2	17.6	17.8	17.6	17.3	19.9	21.1	17.5	16.7	19.3	18.4	18.9	18.4	18.0	18.6	18.5	19.6	18.6	17.2	17.4	17.6	16.9	18.9	0.0	17.5	18.2
tha	17.4	17.0	17.8	19.2	18.7	18.2	19.6	18.1	17.1	17.7	18.4	18.9	18.0	19.9	20.2	18.8	16.3	18.9	19.1	17.9	15.9	18.3	17.8	18.1	19.0	18.7	18.1	18.7	16.9	17.2	18.8	17.5	0.0	17.8
tur	17.9	17.5	18.1	18.0	18.1	18.8	19.0	19.0	16.2	17.4	19.2	18.4	17.9	20.9	19.5	19.3	17.2	19.1	19.8	18.8	16.8	17.6	17.4	18.2	18.9	19.5	18.5	18.7	17.0	17.2	19.7	18.2	17.8	0.0

Table A.82. Euclidean dissimilarity matrix for EU and PUD languages regarding MarsaGram all properties.

Annex 84.

Head and dependent feature
ADV_advmod_precedes_ADJ
ADV_advmod_precedes_NOUN
ADV_advmod_precedes_VERB
CCONJ_cc_precedes_NOUN
CCONJ_cc_precedes_PROPN
CCONJ_cc_precedes_VERB
NOUN_appos_follows_NOUN
NUM_nummod_precedes_NOUN
PRON_nsubj_precedes_VERB
PROPN_nsubj_precedes_VERB
PUNCT_punct_precedes_ADJ
PUNCT_punct_precedes_NOUN
PUNCT_punct_precedes_PROPN
PUNCT_punct_precedes_VERB
PUNCT_punct_follows_ADJ
PUNCT_punct_follows_NOUN
PUNCT_punct_follows_NUM
PUNCT_punct_follows_PRON
PUNCT_punct_follows_PROPN
PUNCT_punct_follows_VERB
VERB_advcl_precedes_VERB

Table A.83. List of head directionality features present in all PUD and EU corpora.

Annex 85.

	Number of features
arb	530
bul	396
cmn	491
hrv	700
ces	593
dan	555
nld	651
eng	607
est	464
fin	490
fra	462
deu	478
ell	465
hin	571
hun	655
isl	637
ind	431
gle	665
ita	504
jpn	243
kor	309
lav	564
lit	569
mlt	661
pol	588
por	496
ron	618
rus	572
slk	390
slv	379
spa	479
swe	590
tha	543
tur	481

Table A.84. Number of extracted head directionality patterns per corpus from the language-set composed of PUD and EU languages.

Annex 86.

	Left-branching features	**Right-branching features**
arb	36.32	56.98
bul	57.87	36.64
cmn	60.06	35.70
hrv	55.51	35.37
ces	57.13	35.80
dan	54.73	38.84
nld	63.57	27.66
eng	63.71	30.73
est	57.19	35.09
fin	55.76	35.62
fra	58.28	34.32
deu	66.76	26.83
ell	60.68	34.07
hin	54.15	36.46
hun	66.16	26.06
isl	51.08	39.58
ind	41.88	49.78
gle	42.23	48.41
ita	57.09	36.80
jpn	45.85	52.12
kor	79.85	14.80
lav	62.12	30.41
lit	61.99	29.08
mlt	49.91	42.70
pol	48.22	41.74
por	57.90	35.05
ron	48.96	42.93
rus	54.50	37.78
slk	55.12	40.20
slv	61.70	32.83
spa	57.74	35.20
swe	58.75	35.01
tha	38.96	52.20
tur	69.91	22.05

Table A.85. Frequency of right and left-branching features for each PUD and EU language.

Annex 87.

	arb	bul	cmn	hrv	ces	dan	nld	eng	est	fin	fra	deu	ell	hin	hun	isl	ind	gle	ita	jpn	kor	lav	lit	mlt	pol	por	ron	rus	slk	slv	spa	swe	tha	tur
arb	0.00	0.13	0.21	0.14	0.14	0.14	0.17	0.16	0.19	0.19	0.17	0.19	0.19	0.25	0.25	0.13	0.11	0.10	0.16	0.30	0.25	0.18	0.21	0.17	0.13	0.15	0.09	0.14	0.17	0.15	0.15	0.14	0.15	0.23
bul	0.13	0.00	0.18	0.07	0.07	0.10	0.14	0.11	0.16	0.16	0.17	0.14	0.16	0.23	0.24	0.09	0.12	0.13	0.15	0.30	0.24	0.15	0.18	0.17	0.10	0.15	0.11	0.07	0.11	0.07	0.15	0.09	0.17	0.20
cmn	0.21	0.18	0.00	0.17	0.16	0.14	0.17	0.15	0.13	0.14	0.21	0.18	0.20	0.20	0.19	0.14	0.16	0.20	0.20	0.25	0.16	0.13	0.16	0.16	0.15	0.20	0.18	0.17	0.15	0.16	0.20	0.15	0.17	0.17
hrv	0.14	0.07	0.17	0.00	0.05	0.10	0.14	0.11	0.14	0.14	0.17	0.14	0.14	0.20	0.22	0.09	0.12	0.14	0.16	0.28	0.22	0.14	0.16	0.16	0.10	0.16	0.12	0.06	0.11	0.06	0.16	0.09	0.16	0.18
ces	0.14	0.07	0.16	0.05	0.00	0.09	0.13	0.10	0.13	0.13	0.16	0.13	0.14	0.20	0.21	0.09	0.12	0.14	0.15	0.28	0.22	0.12	0.15	0.15	0.08	0.15	0.12	0.05	0.09	0.06	0.15	0.08	0.16	0.17
dan	0.14	0.10	0.14	0.10	0.09	0.00	0.08	0.07	0.12	0.12	0.13	0.10	0.12	0.20	0.18	0.08	0.10	0.13	0.12	0.27	0.20	0.11	0.15	0.12	0.10	0.12	0.10	0.10	0.10	0.09	0.12	0.05	0.14	0.16
nld	0.17	0.14	0.17	0.14	0.13	0.08	0.00	0.07	0.15	0.16	0.09	0.05	0.09	0.20	0.16	0.13	0.14	0.13	0.08	0.28	0.21	0.14	0.18	0.12	0.13	0.09	0.13	0.13	0.14	0.13	0.09	0.09	0.17	0.17
eng	0.16	0.11	0.15	0.11	0.10	0.07	0.07	0.00	0.14	0.14	0.10	0.07	0.09	0.20	0.18	0.11	0.12	0.13	0.09	0.28	0.20	0.13	0.16	0.12	0.11	0.09	0.11	0.10	0.12	0.10	0.10	0.06	0.15	0.17
est	0.19	0.16	0.13	0.14	0.13	0.12	0.15	0.14	0.00	0.08	0.20	0.16	0.18	0.17	0.18	0.12	0.15	0.19	0.19	0.24	0.20	0.07	0.10	0.15	0.12	0.19	0.17	0.14	0.12	0.12	0.20	0.13	0.17	0.14
fin	0.19	0.16	0.14	0.14	0.13	0.12	0.16	0.14	0.08	0.00	0.21	0.17	0.19	0.17	0.19	0.11	0.14	0.19	0.20	0.26	0.19	0.11	0.14	0.15	0.13	0.20	0.17	0.13	0.12	0.12	0.20	0.12	0.16	0.14
fra	0.17	0.17	0.21	0.17	0.16	0.13	0.09	0.10	0.20	0.21	0.00	0.09	0.09	0.24	0.19	0.17	0.15	0.12	0.05	0.31	0.25	0.19	0.22	0.14	0.16	0.04	0.12	0.16	0.18	0.17	0.05	0.13	0.18	0.22
deu	0.19	0.14	0.18	0.14	0.13	0.10	0.05	0.07	0.16	0.17	0.09	0.00	0.08	0.20	0.17	0.15	0.15	0.14	0.09	0.29	0.22	0.15	0.18	0.13	0.14	0.09	0.14	0.13	0.14	0.13	0.09	0.10	0.18	0.18
ell	0.19	0.16	0.20	0.14	0.14	0.12	0.09	0.09	0.18	0.19	0.09	0.08	0.00	0.23	0.19	0.16	0.16	0.15	0.09	0.31	0.24	0.17	0.20	0.14	0.17	0.09	0.14	0.14	0.16	0.14	0.09	0.12	0.20	0.20
hin	0.25	0.23	0.20	0.20	0.20	0.20	0.20	0.20	0.17	0.17	0.24	0.20	0.23	0.00	0.21	0.20	0.22	0.24	0.24	0.15	0.21	0.19	0.20	0.20	0.20	0.24	0.23	0.21	0.20	0.20	0.24	0.20	0.22	0.15
hun	0.25	0.24	0.19	0.22	0.21	0.18	0.16	0.18	0.18	0.19	0.19	0.17	0.19	0.21	0.00	0.20	0.20	0.22	0.19	0.28	0.23	0.19	0.21	0.17	0.20	0.19	0.21	0.22	0.20	0.21	0.20	0.19	0.22	0.19
isl	0.13	0.09	0.14	0.09	0.09	0.08	0.13	0.11	0.12	0.11	0.17	0.15	0.16	0.20	0.20	0.00	0.09	0.13	0.16	0.27	0.20	0.12	0.15	0.13	0.08	0.16	0.11	0.10	0.11	0.09	0.16	0.07	0.12	0.17
ind	0.11	0.12	0.16	0.12	0.12	0.10	0.14	0.12	0.15	0.14	0.15	0.15	0.16	0.22	0.20	0.09	0.00	0.11	0.15	0.28	0.21	0.14	0.17	0.13	0.11	0.14	0.09	0.11	0.13	0.12	0.14	0.11	0.13	0.19
gle	0.10	0.13	0.20	0.14	0.14	0.13	0.13	0.13	0.19	0.19	0.12	0.14	0.15	0.24	0.22	0.13	0.11	0.00	0.11	0.30	0.23	0.18	0.20	0.15	0.13	0.11	0.08	0.14	0.17	0.14	0.11	0.12	0.15	0.21
ita	0.16	0.15	0.20	0.16	0.15	0.12	0.08	0.09	0.19	0.20	0.05	0.09	0.09	0.24	0.19	0.16	0.15	0.11	0.00	0.31	0.24	0.18	0.21	0.13	0.15	0.04	0.11	0.15	0.17	0.15	0.05	0.12	0.17	0.21
jpn	0.30	0.30	0.25	0.28	0.28	0.27	0.28	0.28	0.24	0.26	0.31	0.29	0.31	0.15	0.28	0.27	0.28	0.30	0.31	0.00	0.26	0.26	0.26	0.28	0.27	0.31	0.29	0.29	0.28	0.29	0.31	0.28	0.28	0.24
kor	0.25	0.24	0.16	0.22	0.22	0.20	0.21	0.20	0.20	0.19	0.25	0.22	0.24	0.21	0.23	0.20	0.21	0.23	0.24	0.26	0.00	0.21	0.21	0.21	0.20	0.24	0.23	0.23	0.23	0.22	0.25	0.21	0.21	0.17
lav	0.18	0.15	0.13	0.14	0.12	0.11	0.14	0.13	0.07	0.11	0.19	0.15	0.17	0.19	0.19	0.12	0.14	0.18	0.18	0.26	0.21	0.00	0.08	0.14	0.12	0.18	0.15	0.13	0.11	0.12	0.18	0.12	0.16	0.17
lit	0.21	0.18	0.16	0.16	0.15	0.15	0.18	0.16	0.10	0.14	0.22	0.18	0.20	0.20	0.21	0.15	0.17	0.20	0.21	0.26	0.21	0.08	0.00	0.17	0.14	0.21	0.19	0.16	0.15	0.16	0.21	0.15	0.18	0.18
mlt	0.17	0.17	0.16	0.16	0.15	0.12	0.12	0.12	0.15	0.15	0.14	0.13	0.14	0.20	0.17	0.13	0.13	0.15	0.13	0.28	0.21	0.14	0.17	0.00	0.14	0.13	0.13	0.16	0.15	0.15	0.14	0.13	0.15	0.18
pol	0.13	0.10	0.15	0.10	0.08	0.10	0.13	0.11	0.12	0.13	0.16	0.14	0.17	0.20	0.20	0.08	0.11	0.13	0.15	0.27	0.20	0.12	0.14	0.14	0.00	0.15	0.11	0.10	0.11	0.10	0.15	0.09	0.13	0.17
por	0.15	0.15	0.20	0.16	0.15	0.12	0.09	0.09	0.19	0.20	0.04	0.09	0.09	0.24	0.19	0.16	0.14	0.11	0.04	0.31	0.24	0.18	0.21	0.13	0.15	0.00	0.11	0.15	0.17	0.15	0.03	0.12	0.17	0.21
ron	0.09	0.11	0.18	0.12	0.12	0.10	0.13	0.11	0.17	0.17	0.12	0.14	0.14	0.23	0.21	0.11	0.09	0.08	0.11	0.29	0.23	0.15	0.19	0.13	0.11	0.11	0.00	0.12	0.14	0.12	0.11	0.11	0.14	0.20
rus	0.14	0.07	0.17	0.06	0.05	0.10	0.13	0.10	0.14	0.13	0.16	0.13	0.14	0.21	0.22	0.10	0.11	0.14	0.15	0.29	0.23	0.13	0.16	0.16	0.10	0.15	0.12	0.00	0.10	0.07	0.15	0.08	0.16	0.19
slk	0.17	0.11	0.15	0.11	0.09	0.10	0.14	0.12	0.12	0.12	0.18	0.14	0.16	0.20	0.20	0.11	0.13	0.17	0.17	0.28	0.23	0.11	0.15	0.15	0.11	0.17	0.14	0.10	0.00	0.09	0.17	0.11	0.18	0.18
slv	0.15	0.07	0.16	0.06	0.06	0.09	0.13	0.10	0.12	0.12	0.17	0.13	0.14	0.20	0.21	0.09	0.12	0.14	0.15	0.29	0.22	0.12	0.16	0.15	0.10	0.15	0.12	0.07	0.09	0.00	0.15	0.08	0.16	0.18
spa	0.15	0.15	0.20	0.16	0.15	0.12	0.09	0.10	0.20	0.20	0.05	0.09	0.09	0.24	0.20	0.16	0.14	0.11	0.05	0.31	0.25	0.18	0.21	0.14	0.15	0.03	0.11	0.15	0.17	0.15	0.00	0.12	0.17	0.22
swe	0.14	0.09	0.15	0.09	0.08	0.05	0.09	0.06	0.13	0.12	0.13	0.10	0.12	0.20	0.19	0.07	0.11	0.12	0.12	0.28	0.21	0.12	0.15	0.13	0.09	0.12	0.11	0.08	0.11	0.08	0.12	0.00	0.14	0.17
tha	0.15	0.17	0.17	0.16	0.16	0.14	0.17	0.15	0.17	0.16	0.18	0.18	0.20	0.22	0.22	0.12	0.13	0.15	0.17	0.28	0.21	0.16	0.18	0.15	0.13	0.17	0.14	0.16	0.18	0.16	0.17	0.14	0.00	0.20
tur	0.23	0.20	0.17	0.18	0.17	0.16	0.17	0.17	0.14	0.14	0.22	0.18	0.20	0.15	0.19	0.17	0.19	0.21	0.21	0.24	0.17	0.17	0.18	0.18	0.17	0.21	0.20	0.19	0.18	0.18	0.22	0.17	0.20	0.00

Table A.86. Euclidean dissimilarity matrix for EU and PUD languages regarding head and dependent relative position.

Annex 88.

	arb	bul	cmn	hrv	ces	dan	nld	eng	est	fin	fra	deu	ell	hin	hun	isl	ind	gle	ita	jpn	kor	lav	lit	mlt	pol	por	ron	rus	slk	slv	spa	swe	tha	tur
arb	0.00	0.15	0.27	0.18	0.18	0.18	0.23	0.20	0.24	0.25	0.22	0.24	0.25	0.33	0.33	0.17	0.12	0.11	0.20	0.40	0.32	0.23	0.27	0.22	0.16	0.19	0.09	0.17	0.21	0.18	0.19	0.17	0.19	0.29
bul	0.15	0.00	0.22	0.07	0.07	0.11	0.17	0.13	0.19	0.19	0.21	0.17	0.19	0.30	0.31	0.10	0.13	0.16	0.19	0.40	0.31	0.18	0.22	0.21	0.11	0.19	0.12	0.07	0.12	0.06	0.19	0.09	0.20	0.25
cmn	0.27	0.22	0.00	0.22	0.21	0.17	0.21	0.19	0.16	0.17	0.27	0.23	0.26	0.26	0.26	0.18	0.19	0.26	0.26	0.33	0.19	0.16	0.19	0.20	0.19	0.26	0.23	0.22	0.18	0.20	0.27	0.19	0.20	0.21
hrv	0.18	0.07	0.22	0.00	0.04	0.12	0.17	0.13	0.16	0.17	0.22	0.17	0.17	0.27	0.30	0.11	0.14	0.17	0.20	0.38	0.29	0.16	0.20	0.20	0.12	0.20	0.14	0.06	0.13	0.05	0.20	0.10	0.20	0.22
ces	0.18	0.07	0.21	0.04	0.00	0.11	0.16	0.12	0.15	0.15	0.21	0.15	0.17	0.27	0.28	0.10	0.14	0.17	0.19	0.38	0.28	0.15	0.18	0.19	0.08	0.19	0.14	0.04	0.09	0.06	0.19	0.08	0.20	0.21
dan	0.18	0.11	0.17	0.12	0.11	0.00	0.10	0.07	0.13	0.14	0.16	0.12	0.14	0.26	0.24	0.09	0.12	0.15	0.14	0.36	0.26	0.13	0.19	0.14	0.11	0.15	0.12	0.12	0.11	0.10	0.15	0.05	0.17	0.20
nld	0.23	0.17	0.21	0.17	0.16	0.10	0.00	0.06	0.18	0.21	0.10	0.05	0.09	0.27	0.22	0.17	0.17	0.16	0.09	0.38	0.28	0.18	0.23	0.15	0.17	0.10	0.15	0.17	0.17	0.15	0.11	0.10	0.22	0.21
eng	0.20	0.13	0.19	0.13	0.12	0.07	0.06	0.00	0.17	0.18	0.11	0.07	0.10	0.26	0.23	0.13	0.14	0.16	0.10	0.38	0.26	0.16	0.21	0.14	0.13	0.10	0.13	0.12	0.14	0.11	0.11	0.04	0.19	0.21
est	0.24	0.19	0.16	0.16	0.15	0.13	0.18	0.17	0.00	0.08	0.26	0.19	0.22	0.22	0.24	0.14	0.17	0.24	0.25	0.31	0.24	0.06	0.10	0.18	0.15	0.25	0.21	0.16	0.13	0.13	0.25	0.15	0.20	0.17
fin	0.25	0.19	0.17	0.17	0.15	0.14	0.21	0.18	0.08	0.00	0.27	0.21	0.24	0.22	0.25	0.13	0.16	0.25	0.27	0.34	0.24	0.12	0.17	0.19	0.16	0.26	0.21	0.16	0.14	0.14	0.26	0.14	0.20	0.16
fra	0.22	0.21	0.27	0.22	0.21	0.16	0.10	0.11	0.26	0.27	0.00	0.10	0.10	0.33	0.25	0.22	0.19	0.15	0.04	0.43	0.33	0.24	0.28	0.17	0.21	0.02	0.15	0.20	0.23	0.21	0.03	0.15	0.23	0.28
deu	0.24	0.17	0.23	0.17	0.15	0.12	0.05	0.07	0.19	0.21	0.10	0.00	0.08	0.26	0.22	0.18	0.18	0.18	0.10	0.39	0.29	0.18	0.23	0.16	0.18	0.10	0.17	0.15	0.17	0.15	0.10	0.11	0.23	0.22
ell	0.25	0.19	0.26	0.17	0.17	0.14	0.09	0.10	0.22	0.24	0.10	0.08	0.00	0.31	0.24	0.21	0.20	0.18	0.10	0.42	0.32	0.22	0.25	0.17	0.21	0.10	0.17	0.17	0.19	0.16	0.10	0.13	0.25	0.26
hin	0.33	0.30	0.26	0.27	0.27	0.26	0.27	0.26	0.22	0.22	0.33	0.26	0.31	0.00	0.30	0.27	0.28	0.32	0.32	0.20	0.28	0.25	0.27	0.27	0.26	0.32	0.31	0.28	0.27	0.26	0.33	0.26	0.29	0.19
hun	0.33	0.31	0.26	0.30	0.28	0.24	0.22	0.23	0.24	0.25	0.25	0.22	0.24	0.30	0.00	0.28	0.27	0.29	0.26	0.38	0.29	0.24	0.27	0.22	0.27	0.26	0.28	0.30	0.26	0.27	0.26	0.25	0.29	0.25
isl	0.17	0.10	0.18	0.11	0.10	0.09	0.17	0.13	0.14	0.13	0.22	0.18	0.21	0.27	0.28	0.00	0.10	0.17	0.20	0.36	0.26	0.15	0.19	0.16	0.10	0.20	0.14	0.12	0.13	0.10	0.20	0.07	0.14	0.22
ind	0.12	0.13	0.19	0.14	0.14	0.12	0.17	0.14	0.17	0.16	0.19	0.18	0.20	0.28	0.27	0.10	0.00	0.13	0.18	0.37	0.26	0.16	0.21	0.15	0.12	0.17	0.09	0.13	0.15	0.13	0.17	0.12	0.15	0.24
gle	0.11	0.16	0.26	0.17	0.17	0.15	0.16	0.16	0.24	0.25	0.15	0.18	0.18	0.32	0.29	0.17	0.13	0.00	0.14	0.41	0.31	0.23	0.26	0.19	0.16	0.13	0.10	0.17	0.21	0.18	0.13	0.14	0.19	0.28
ita	0.20	0.19	0.26	0.20	0.19	0.14	0.09	0.10	0.25	0.27	0.04	0.10	0.10	0.32	0.26	0.20	0.18	0.14	0.00	0.42	0.32	0.23	0.28	0.16	0.20	0.03	0.13	0.20	0.22	0.19	0.04	0.14	0.22	0.27
jpn	0.40	0.40	0.33	0.38	0.38	0.36	0.38	0.38	0.31	0.34	0.43	0.39	0.42	0.20	0.38	0.36	0.37	0.41	0.42	0.00	0.33	0.34	0.34	0.37	0.37	0.42	0.39	0.40	0.37	0.37	0.42	0.37	0.37	0.31
kor	0.32	0.31	0.19	0.29	0.28	0.26	0.28	0.26	0.24	0.24	0.33	0.29	0.32	0.28	0.29	0.26	0.26	0.31	0.32	0.33	0.00	0.26	0.27	0.26	0.26	0.32	0.29	0.30	0.29	0.28	0.33	0.27	0.27	0.21
lav	0.23	0.18	0.16	0.16	0.15	0.13	0.18	0.16	0.06	0.12	0.24	0.18	0.22	0.25	0.24	0.15	0.16	0.23	0.23	0.34	0.26	0.00	0.08	0.17	0.14	0.23	0.19	0.15	0.12	0.13	0.23	0.14	0.20	0.20
lit	0.27	0.22	0.19	0.20	0.18	0.19	0.23	0.21	0.10	0.17	0.28	0.23	0.25	0.27	0.27	0.19	0.21	0.26	0.28	0.34	0.27	0.08	0.00	0.21	0.17	0.28	0.24	0.20	0.18	0.18	0.28	0.19	0.23	0.22
mlt	0.22	0.21	0.20	0.20	0.19	0.14	0.15	0.14	0.18	0.19	0.17	0.16	0.17	0.27	0.22	0.16	0.15	0.19	0.16	0.37	0.26	0.17	0.21	0.00	0.17	0.16	0.17	0.20	0.18	0.17	0.17	0.15	0.18	0.22
pol	0.16	0.11	0.19	0.12	0.08	0.11	0.17	0.13	0.15	0.16	0.21	0.18	0.21	0.26	0.27	0.10	0.12	0.16	0.20	0.37	0.26	0.14	0.17	0.17	0.00	0.19	0.14	0.11	0.12	0.11	0.19	0.11	0.16	0.21
por	0.19	0.19	0.26	0.20	0.19	0.15	0.10	0.10	0.25	0.26	0.02	0.10	0.10	0.32	0.26	0.20	0.17	0.13	0.03	0.42	0.32	0.23	0.28	0.16	0.19	0.00	0.12	0.19	0.22	0.19	0.00	0.14	0.22	0.28
ron	0.09	0.12	0.23	0.14	0.14	0.12	0.15	0.13	0.21	0.21	0.15	0.17	0.17	0.31	0.28	0.14	0.09	0.10	0.13	0.39	0.29	0.19	0.24	0.17	0.14	0.12	0.00	0.14	0.17	0.13	0.12	0.12	0.17	0.26
rus	0.17	0.07	0.22	0.06	0.04	0.12	0.17	0.12	0.16	0.16	0.20	0.15	0.17	0.28	0.30	0.12	0.13	0.17	0.20	0.40	0.30	0.15	0.20	0.20	0.11	0.19	0.14	0.00	0.11	0.07	0.19	0.08	0.20	0.24
slk	0.21	0.12	0.18	0.13	0.09	0.11	0.17	0.14	0.13	0.14	0.23	0.17	0.19	0.27	0.26	0.13	0.15	0.21	0.22	0.37	0.29	0.12	0.18	0.18	0.12	0.22	0.17	0.11	0.00	0.09	0.21	0.12	0.22	0.22
slv	0.18	0.06	0.20	0.05	0.06	0.10	0.15	0.11	0.13	0.14	0.21	0.15	0.16	0.26	0.27	0.10	0.13	0.18	0.19	0.37	0.28	0.13	0.18	0.17	0.11	0.19	0.13	0.07	0.09	0.00	0.18	0.08	0.20	0.22
spa	0.19	0.19	0.27	0.20	0.19	0.15	0.11	0.11	0.25	0.26	0.03	0.10	0.10	0.33	0.26	0.20	0.17	0.13	0.04	0.42	0.33	0.23	0.28	0.17	0.19	0.00	0.12	0.19	0.21	0.18	0.00	0.14	0.22	0.28
swe	0.17	0.09	0.19	0.10	0.08	0.05	0.10	0.04	0.15	0.14	0.15	0.11	0.13	0.26	0.25	0.07	0.12	0.14	0.14	0.37	0.27	0.14	0.19	0.15	0.11	0.14	0.12	0.08	0.12	0.08	0.14	0.00	0.16	0.21
tha	0.19	0.20	0.20	0.20	0.20	0.17	0.22	0.19	0.20	0.20	0.23	0.23	0.25	0.29	0.29	0.14	0.15	0.19	0.22	0.37	0.27	0.20	0.23	0.18	0.16	0.22	0.17	0.20	0.22	0.20	0.22	0.16	0.00	0.25
tur	0.29	0.25	0.21	0.22	0.21	0.20	0.21	0.21	0.17	0.16	0.28	0.22	0.26	0.19	0.25	0.22	0.24	0.28	0.27	0.31	0.21	0.20	0.22	0.22	0.21	0.28	0.26	0.24	0.22	0.22	0.28	0.21	0.25	0.00

Table A.87. Dissimilarity matrix for EU and PUD languages regarding combination of the MarsaGram all properties and the head and dependent relative position distances (Euclidean).

Annex 89.

	arb	bul	cmn	hrv	ces	dan	nld	eng	est	fin	fra	deu	ell	hin	hun	isl	ind	gle	ita	jpn	kor	lav	lit	mlt	pol	por	ron	rus	slk	slv	spa	swe	tha	tur
arb	0.00	0.02	0.01	0.02	0.02	0.01	0.48	0.00	0.11	0.04	0.03	0.40	0.02	0.99	0.26	0.00	0.00	0.03	0.01	0.99	1.00	0.04	0.03	0.01	0.01	0.01	0.02	0.00	0.04	0.09	0.01	0.00	0.03	0.99
bul	0.02	0.00	0.02	0.01	0.00	0.04	0.42	0.02	0.08	0.02	0.01	0.35	0.00	0.93	0.21	0.03	0.02	0.04	0.01	0.93	0.93	0.02	0.01	0.01	0.01	0.01	0.00	0.01	0.02	0.04	0.00	0.02	0.03	0.93
cmn	0.01	0.02	0.00	0.03	0.02	0.03	0.50	0.00	0.12	0.04	0.04	0.43	0.03	1.00	0.28	0.01	0.00	0.03	0.01	1.00	1.00	0.05	0.03	0.01	0.01	0.01	0.02	0.01	0.07	0.11	0.01	0.01	0.01	1.00
hrv	0.02	0.01	0.03	0.00	0.00	0.04	0.39	0.02	0.06	0.02	0.02	0.32	0.01	0.88	0.19	0.03	0.02	0.04	0.01	0.89	0.89	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.04	0.01	0.02	0.03	0.88
ces	0.02	0.00	0.02	0.00	0.00	0.04	0.39	0.02	0.06	0.01	0.02	0.33	0.01	0.88	0.19	0.03	0.02	0.04	0.01	0.89	0.89	0.02	0.00	0.01	0.01	0.01	0.01	0.01	0.03	0.05	0.01	0.02	0.03	0.88
dan	0.01	0.04	0.03	0.04	0.04	0.00	0.48	0.02	0.12	0.05	0.06	0.40	0.05	0.98	0.27	0.01	0.02	0.04	0.03	0.98	0.98	0.06	0.05	0.02	0.03	0.03	0.04	0.02	0.05	0.11	0.04	0.01	0.03	0.98
nld	0.48	0.42	0.50	0.39	0.39	0.48	0.00	0.49	0.16	0.28	0.45	0.01	0.44	0.16	0.04	0.49	0.50	0.50	0.49	0.17	0.17	0.28	0.36	0.47	0.45	0.48	0.46	0.45	0.36	0.28	0.48	0.49	0.48	0.16
eng	0.00	0.02	0.00	0.02	0.02	0.02	0.49	0.00	0.11	0.04	0.03	0.41	0.02	0.99	0.27	0.01	0.00	0.02	0.00	0.99	0.99	0.05	0.02	0.00	0.00	0.00	0.02	0.00	0.06	0.10	0.01	0.00	0.02	0.99
est	0.11	0.08	0.12	0.06	0.06	0.12	0.16	0.11	0.00	0.02	0.11	0.11	0.09	0.55	0.04	0.12	0.12	0.13	0.11	0.56	0.56	0.02	0.05	0.10	0.09	0.11	0.10	0.09	0.07	0.05	0.10	0.11	0.12	0.55
fin	0.04	0.02	0.04	0.02	0.01	0.05	0.28	0.04	0.02	0.00	0.04	0.22	0.03	0.74	0.11	0.04	0.04	0.06	0.04	0.74	0.75	0.00	0.01	0.03	0.03	0.04	0.03	0.02	0.03	0.04	0.03	0.04	0.05	0.74
fra	0.03	0.01	0.04	0.02	0.02	0.06	0.45	0.03	0.11	0.04	0.00	0.39	0.00	0.98	0.25	0.04	0.04	0.05	0.02	0.99	0.99	0.03	0.03	0.03	0.03	0.02	0.00	0.02	0.03	0.04	0.01	0.04	0.05	0.98
deu	0.40	0.35	0.43	0.32	0.33	0.40	0.01	0.41	0.11	0.22	0.39	0.00	0.38	0.21	0.03	0.41	0.42	0.43	0.41	0.22	0.22	0.23	0.30	0.39	0.38	0.41	0.39	0.38	0.29	0.23	0.40	0.41	0.41	0.21
ell	0.02	0.00	0.03	0.01	0.01	0.05	0.44	0.02	0.09	0.03	0.00	0.38	0.00	0.96	0.23	0.03	0.03	0.04	0.01	0.97	0.97	0.02	0.02	0.02	0.02	0.01	0.00	0.01	0.03	0.04	0.01	0.03	0.04	0.96
hin	0.99	0.93	1.00	0.88	0.88	0.98	0.16	0.99	0.55	0.74	0.98	0.21	0.96	0.00	0.33	0.99	1.00	0.99	0.99	0.00	0.00	0.76	0.83	0.97	0.95	0.99	0.98	0.95	0.87	0.76	0.99	0.99	0.97	0.00
hun	0.26	0.21	0.28	0.19	0.19	0.27	0.04	0.27	0.04	0.11	0.25	0.03	0.23	0.33	0.00	0.27	0.28	0.28	0.26	0.34	0.34	0.12	0.17	0.25	0.24	0.26	0.24	0.24	0.18	0.13	0.25	0.27	0.27	0.33
isl	0.00	0.03	0.01	0.03	0.03	0.01	0.49	0.01	0.12	0.04	0.04	0.41	0.03	0.99	0.27	0.00	0.00	0.03	0.01	0.99	0.99	0.05	0.03	0.01	0.01	0.02	0.03	0.01	0.05	0.11	0.02	0.00	0.03	0.99
ind	0.00	0.02	0.00	0.02	0.02	0.02	0.50	0.00	0.12	0.04	0.04	0.42	0.03	1.00	0.28	0.00	0.00	0.02	0.01	1.00	1.00	0.05	0.03	0.00	0.00	0.01	0.02	0.00	0.06	0.11	0.01	0.00	0.02	1.00
gle	0.03	0.04	0.03	0.04	0.04	0.04	0.50	0.02	0.13	0.06	0.05	0.43	0.04	0.99	0.28	0.03	0.02	0.00	0.03	0.99	0.99	0.07	0.04	0.02	0.02	0.03	0.04	0.02	0.08	0.12	0.03	0.02	0.04	0.99
ita	0.01	0.01	0.01	0.01	0.01	0.03	0.49	0.00	0.11	0.04	0.02	0.41	0.01	0.99	0.26	0.01	0.01	0.03	0.00	1.00	1.00	0.04	0.02	0.00	0.00	0.00	0.01	0.00	0.05	0.08	0.00	0.01	0.02	0.99
jpn	0.99	0.93	1.00	0.89	0.89	0.98	0.17	0.99	0.56	0.74	0.99	0.22	0.97	0.00	0.34	0.99	1.00	0.99	1.00	0.00	0.00	0.76	0.84	0.97	0.95	0.99	0.98	0.95	0.88	0.77	0.99	0.99	0.97	0.00
kor	1.00	0.93	1.00	0.89	0.89	0.98	0.17	0.99	0.56	0.75	0.99	0.22	0.97	0.00	0.34	0.99	1.00	0.99	1.00	0.00	0.00	0.77	0.84	0.97	0.95	1.00	0.98	0.96	0.88	0.78	0.99	0.99	0.97	0.00
lav	0.04	0.02	0.05	0.01	0.02	0.06	0.28	0.05	0.02	0.00	0.03	0.23	0.02	0.76	0.12	0.05	0.05	0.07	0.04	0.76	0.77	0.00	0.01	0.04	0.03	0.04	0.03	0.03	0.02	0.02	0.03	0.05	0.06	0.76
lit	0.03	0.01	0.03	0.01	0.00	0.05	0.36	0.02	0.05	0.01	0.03	0.30	0.02	0.83	0.17	0.03	0.03	0.04	0.02	0.84	0.84	0.01	0.00	0.01	0.01	0.02	0.02	0.01	0.04	0.06	0.01	0.03	0.03	0.84
mlt	0.01	0.01	0.01	0.01	0.01	0.02	0.47	0.00	0.10	0.03	0.03	0.39	0.02	0.97	0.25	0.01	0.00	0.02	0.00	0.97	0.97	0.04	0.01	0.00	0.00	0.00	0.01	0.00	0.05	0.09	0.00	0.01	0.02	0.97
pol	0.01	0.01	0.01	0.01	0.01	0.03	0.45	0.00	0.09	0.03	0.03	0.38	0.02	0.95	0.24	0.01	0.00	0.02	0.00	0.95	0.95	0.03	0.01	0.00	0.00	0.00	0.01	0.00	0.05	0.09	0.01	0.01	0.02	0.95
por	0.01	0.01	0.01	0.01	0.01	0.03	0.48	0.00	0.11	0.04	0.02	0.41	0.01	0.99	0.26	0.02	0.01	0.03	0.00	0.99	1.00	0.04	0.02	0.00	0.00	0.00	0.01	0.00	0.05	0.08	0.00	0.01	0.02	0.99
ron	0.02	0.00	0.02	0.01	0.01	0.04	0.46	0.02	0.10	0.03	0.00	0.39	0.00	0.98	0.24	0.03	0.02	0.04	0.01	0.98	0.98	0.03	0.02	0.01	0.01	0.01	0.00	0.01	0.03	0.05	0.00	0.02	0.03	0.98
rus	0.00	0.01	0.01	0.01	0.01	0.02	0.45	0.00	0.09	0.02	0.02	0.38	0.01	0.95	0.24	0.01	0.00	0.02	0.00	0.95	0.96	0.03	0.01	0.00	0.00	0.00	0.01	0.00	0.04	0.08	0.00	0.00	0.02	0.95
slk	0.04	0.02	0.07	0.02	0.03	0.05	0.36	0.06	0.07	0.03	0.03	0.29	0.03	0.87	0.18	0.05	0.06	0.08	0.05	0.88	0.88	0.02	0.04	0.05	0.05	0.05	0.03	0.04	0.00	0.02	0.04	0.05	0.07	0.87
slv	0.09	0.04	0.11	0.04	0.05	0.11	0.28	0.10	0.05	0.04	0.04	0.23	0.04	0.76	0.13	0.11	0.11	0.12	0.08	0.77	0.78	0.02	0.06	0.09	0.09	0.08	0.05	0.08	0.02	0.00	0.07	0.10	0.11	0.76
spa	0.01	0.00	0.01	0.01	0.01	0.04	0.48	0.01	0.10	0.03	0.01	0.40	0.01	0.99	0.25	0.02	0.01	0.03	0.00	0.99	0.99	0.03	0.01	0.00	0.01	0.00	0.00	0.00	0.04	0.07	0.00	0.01	0.03	0.99
swe	0.00	0.02	0.01	0.02	0.02	0.01	0.49	0.00	0.11	0.04	0.04	0.41	0.03	0.99	0.27	0.00	0.00	0.02	0.01	0.99	0.99	0.05	0.03	0.01	0.01	0.01	0.02	0.00	0.05	0.10	0.01	0.00	0.02	0.99
tha	0.03	0.03	0.01	0.03	0.03	0.03	0.48	0.02	0.12	0.05	0.05	0.41	0.04	0.97	0.27	0.03	0.02	0.04	0.02	0.97	0.97	0.06	0.03	0.02	0.02	0.02	0.03	0.02	0.07	0.11	0.03	0.02	0.00	0.97
tur	0.99	0.93	1.00	0.88	0.88	0.98	0.16	0.99	0.55	0.74	0.98	0.21	0.96	0.00	0.33	0.99	1.00	0.99	0.99	0.00	0.00	0.76	0.84	0.97	0.95	0.99	0.98	0.95	0.87	0.76	0.99	0.99	0.97	0.00

Table A.88. Cosine dissimilarity matrix for EU and PUD languages regarding the verb and object relative position.

Annex 90.

	arb	bul	cmn	hrv	ces	dan	nld	eng	est	fin	fra	deu	ell	hin	hun	isl	ind	gle	ita	jpn	kor	lav	lit	mlt	pol	por	ron	rus	slk	slv	spa	swe	tha	tur
arb	0.00	0.64	0.79	0.69	0.67	0.71	0.70	0.67	0.71	0.77	0.67	0.66	0.70	0.67	0.85	0.70	0.70	0.68	0.66	0.74	0.90	0.72	0.77	0.79	0.73	0.65	0.66	0.66	0.70	0.71	0.68	0.62	0.73	0.71
bul	0.64	0.00	0.78	0.58	0.62	0.70	0.67	0.62	0.67	0.73	0.67	0.62	0.64	0.71	0.82	0.68	0.68	0.74	0.63	0.74	0.91	0.65	0.73	0.74	0.69	0.67	0.60	0.56	0.61	0.57	0.66	0.58	0.76	0.74
cmn	0.79	0.78	0.00	0.79	0.80	0.80	0.76	0.76	0.80	0.81	0.77	0.72	0.77	0.77	0.87	0.80	0.78	0.83	0.79	0.84	0.82	0.76	0.77	0.81	0.82	0.79	0.77	0.78	0.78	0.78	0.80	0.75	0.80	0.76
hrv	0.69	0.58	0.79	0.00	0.59	0.70	0.65	0.62	0.63	0.70	0.69	0.64	0.63	0.70	0.80	0.67	0.71	0.72	0.62	0.75	0.92	0.59	0.67	0.69	0.70	0.67	0.60	0.57	0.64	0.50	0.65	0.59	0.79	0.64
ces	0.67	0.62	0.80	0.59	0.00	0.76	0.67	0.63	0.64	0.71	0.67	0.61	0.66	0.73	0.82	0.65	0.70	0.74	0.63	0.77	0.91	0.66	0.65	0.72	0.60	0.65	0.64	0.58	0.56	0.60	0.65	0.60	0.78	0.67
dan	0.71	0.70	0.80	0.70	0.76	0.00	0.66	0.66	0.73	0.77	0.68	0.68	0.68	0.74	0.86	0.71	0.75	0.74	0.69	0.80	0.93	0.76	0.78	0.77	0.76	0.71	0.67	0.72	0.73	0.72	0.70	0.65	0.80	0.78
nld	0.70	0.67	0.76	0.65	0.67	0.66	0.00	0.57	0.67	0.72	0.58	0.59	0.56	0.69	0.79	0.68	0.66	0.65	0.56	0.80	0.90	0.68	0.72	0.70	0.75	0.61	0.60	0.65	0.70	0.66	0.64	0.54	0.78	0.69
eng	0.67	0.62	0.76	0.62	0.63	0.66	0.57	0.00	0.68	0.70	0.55	0.50	0.59	0.62	0.81	0.61	0.58	0.69	0.48	0.75	0.89	0.68	0.75	0.68	0.72	0.50	0.58	0.58	0.71	0.64	0.57	0.46	0.69	0.71
est	0.71	0.67	0.80	0.63	0.64	0.73	0.67	0.68	0.00	0.55	0.67	0.63	0.67	0.79	0.79	0.69	0.71	0.73	0.68	0.80	0.90	0.63	0.70	0.73	0.75	0.67	0.69	0.64	0.70	0.64	0.70	0.65	0.82	0.66
fin	0.77	0.73	0.81	0.70	0.71	0.77	0.72	0.70	0.55	0.00	0.70	0.66	0.73	0.77	0.81	0.66	0.72	0.78	0.73	0.84	0.90	0.68	0.76	0.75	0.79	0.74	0.74	0.66	0.76	0.70	0.74	0.64	0.80	0.71
fra	0.67	0.67	0.77	0.69	0.67	0.68	0.58	0.55	0.67	0.70	0.00	0.57	0.57	0.69	0.80	0.69	0.63	0.68	0.51	0.78	0.90	0.68	0.74	0.67	0.72	0.50	0.65	0.60	0.71	0.68	0.52	0.57	0.73	0.74
deu	0.66	0.62	0.72	0.64	0.61	0.68	0.59	0.50	0.63	0.66	0.57	0.00	0.56	0.65	0.77	0.66	0.61	0.70	0.53	0.78	0.87	0.65	0.67	0.67	0.70	0.55	0.64	0.58	0.64	0.61	0.57	0.55	0.75	0.66
ell	0.70	0.64	0.77	0.63	0.66	0.68	0.56	0.59	0.67	0.73	0.57	0.56	0.00	0.73	0.77	0.70	0.67	0.71	0.53	0.78	0.91	0.67	0.70	0.66	0.77	0.59	0.61	0.63	0.68	0.64	0.60	0.59	0.79	0.72
hin	0.67	0.71	0.77	0.70	0.73	0.74	0.69	0.62	0.79	0.77	0.69	0.65	0.73	0.00	0.86	0.71	0.70	0.73	0.64	0.74	0.88	0.75	0.79	0.76	0.76	0.63	0.68	0.69	0.75	0.68	0.67	0.62	0.72	0.75
hun	0.85	0.82	0.87	0.80	0.82	0.86	0.79	0.81	0.79	0.81	0.80	0.77	0.77	0.86	0.00	0.84	0.83	0.84	0.81	0.89	0.92	0.77	0.82	0.82	0.85	0.81	0.82	0.80	0.82	0.80	0.81	0.81	0.91	0.78
isl	0.70	0.68	0.80	0.67	0.65	0.71	0.68	0.61	0.69	0.66	0.69	0.66	0.70	0.71	0.84	0.00	0.68	0.74	0.65	0.78	0.89	0.70	0.77	0.71	0.75	0.67	0.68	0.64	0.72	0.68	0.71	0.54	0.73	0.72
ind	0.70	0.68	0.78	0.71	0.70	0.75	0.66	0.58	0.71	0.72	0.63	0.61	0.67	0.70	0.83	0.68	0.00	0.72	0.62	0.79	0.87	0.69	0.73	0.71	0.74	0.61	0.70	0.63	0.72	0.68	0.68	0.61	0.75	0.75
gle	0.68	0.74	0.83	0.72	0.74	0.74	0.65	0.69	0.73	0.78	0.68	0.70	0.71	0.73	0.84	0.74	0.72	0.00	0.65	0.79	0.92	0.75	0.80	0.79	0.79	0.68	0.69	0.71	0.76	0.76	0.70	0.68	0.79	0.74
ita	0.66	0.63	0.79	0.62	0.63	0.69	0.56	0.48	0.68	0.73	0.51	0.53	0.53	0.64	0.81	0.65	0.62	0.65	0.00	0.76	0.92	0.69	0.74	0.68	0.72	0.51	0.61	0.61	0.69	0.64	0.52	0.57	0.71	0.72
jpn	0.74	0.74	0.84	0.75	0.77	0.80	0.80	0.75	0.80	0.84	0.78	0.78	0.78	0.74	0.89	0.78	0.79	0.79	0.76	0.00	0.91	0.83	0.85	0.80	0.84	0.78	0.76	0.79	0.80	0.75	0.79	0.72	0.80	0.81
kor	0.90	0.91	0.82	0.92	0.91	0.93	0.90	0.89	0.90	0.90	0.90	0.87	0.91	0.88	0.92	0.89	0.87	0.92	0.92	0.91	0.00	0.90	0.91	0.91	0.93	0.89	0.90	0.90	0.93	0.91	0.90	0.87	0.88	0.86
lav	0.72	0.65	0.76	0.59	0.66	0.76	0.68	0.68	0.63	0.68	0.68	0.65	0.67	0.75	0.77	0.70	0.69	0.75	0.69	0.83	0.90	0.00	0.67	0.70	0.77	0.69	0.65	0.59	0.67	0.61	0.71	0.64	0.82	0.69
lit	0.77	0.73	0.77	0.67	0.65	0.78	0.72	0.75	0.70	0.76	0.74	0.67	0.70	0.79	0.82	0.77	0.73	0.80	0.74	0.85	0.91	0.67	0.00	0.74	0.70	0.76	0.74	0.69	0.60	0.67	0.76	0.75	0.87	0.75
mlt	0.79	0.74	0.81	0.69	0.72	0.77	0.70	0.68	0.73	0.75	0.67	0.67	0.66	0.76	0.82	0.71	0.71	0.79	0.68	0.80	0.91	0.70	0.74	0.00	0.78	0.71	0.72	0.69	0.74	0.70	0.71	0.68	0.80	0.74
pol	0.73	0.69	0.82	0.70	0.60	0.76	0.75	0.72	0.75	0.79	0.72	0.70	0.77	0.76	0.85	0.75	0.74	0.79	0.72	0.84	0.93	0.77	0.70	0.78	0.00	0.74	0.74	0.68	0.63	0.68	0.72	0.73	0.83	0.76
por	0.65	0.67	0.79	0.67	0.65	0.71	0.61	0.50	0.67	0.74	0.50	0.55	0.59	0.63	0.81	0.67	0.61	0.68	0.51	0.78	0.89	0.69	0.76	0.71	0.74	0.00	0.62	0.62	0.71	0.67	0.42	0.60	0.72	0.74
ron	0.66	0.60	0.77	0.60	0.64	0.67	0.60	0.58	0.69	0.74	0.65	0.64	0.61	0.68	0.82	0.68	0.70	0.69	0.61	0.76	0.90	0.65	0.74	0.72	0.74	0.62	0.00	0.60	0.69	0.64	0.64	0.56	0.75	0.72
rus	0.66	0.56	0.78	0.57	0.58	0.72	0.65	0.58	0.64	0.66	0.60	0.58	0.63	0.69	0.80	0.64	0.63	0.71	0.61	0.79	0.90	0.59	0.69	0.69	0.68	0.62	0.60	0.00	0.67	0.62	0.64	0.54	0.74	0.69
slk	0.70	0.61	0.78	0.64	0.56	0.73	0.70	0.71	0.70	0.76	0.71	0.64	0.68	0.75	0.82	0.72	0.72	0.76	0.69	0.80	0.93	0.67	0.60	0.74	0.63	0.71	0.69	0.67	0.00	0.59	0.69	0.69	0.81	0.73
slv	0.71	0.57	0.78	0.50	0.60	0.72	0.66	0.64	0.64	0.70	0.68	0.61	0.64	0.68	0.80	0.68	0.68	0.76	0.64	0.75	0.91	0.61	0.67	0.70	0.68	0.67	0.64	0.62	0.59	0.00	0.65	0.61	0.80	0.72
spa	0.68	0.66	0.80	0.65	0.65	0.70	0.64	0.57	0.70	0.74	0.52	0.57	0.60	0.67	0.81	0.71	0.68	0.70	0.52	0.79	0.90	0.71	0.76	0.71	0.72	0.42	0.64	0.64	0.69	0.65	0.00	0.63	0.73	0.75
swe	0.62	0.58	0.75	0.59	0.60	0.65	0.54	0.46	0.65	0.64	0.57	0.55	0.59	0.62	0.81	0.54	0.61	0.68	0.57	0.72	0.87	0.64	0.75	0.68	0.73	0.60	0.56	0.54	0.69	0.61	0.63	0.00	0.68	0.69
tha	0.73	0.76	0.80	0.79	0.78	0.80	0.78	0.69	0.82	0.80	0.73	0.75	0.79	0.72	0.91	0.73	0.75	0.79	0.71	0.80	0.88	0.82	0.87	0.80	0.83	0.72	0.75	0.74	0.81	0.80	0.73	0.68	0.00	0.82
tur	0.71	0.74	0.76	0.64	0.67	0.78	0.69	0.71	0.66	0.71	0.74	0.66	0.72	0.75	0.78	0.72	0.75	0.74	0.72	0.81	0.86	0.69	0.75	0.74	0.76	0.74	0.72	0.69	0.73	0.72	0.75	0.69	0.82	0.00

Table A.89. Cosine dissimilarity matrix for EU and PUD languages regarding MarsaGram all properties

Biography of the author

Diego Fernando Válio Antunes Alves was born on September 1st, 1983 in São Paulo, Brazil. He graduated in Chemical Engineering in 2006 at the Polytechnic School of the University of São Paulo and at the Paris Technical Institute of Chemistry as part of a double degree program. After 10 years of experience in the cosmetic industry, he obtained his bachelor's degree in Modern Letters (Comparative Literature) at Sorbonne Nouvelle in 2017. In 2019, he received his MA in Computational Linguistics at Sorbonne University. He is currently employed as an Early-researcher for the CLEOPATRA project at the Linguistic Institute of the Faculty of Humanities and Social Sciences of the University of Zagreb. His broad research interests include Corpus-based Typology and Natural Language Processing for low-resource languages, particularly dependency parsing. Throughout his academic career, he has published 11 research papers in international conference proceedings.

List of Publications

Alves, D., Tadić, M., & Bekavac, B. (2022, June). Multilingual Comparative Analysis of Deep-Learning Dependency Parsing Results Using Parallel Corpora. In *Proceedings of the BUCC Workshop within LREC 2022* (pp. 33-42).

Alves, D., Thakkar, G., Amaral, G., Kuculo, T., & Tadić, M. (2022). Building Multilingual Corpora for a Complex Named Entity Recognition and Classification Hierarchy using Wikipedia and DBpedia. *arXiv preprint arXiv:2212.07429*.

Alves, D., Thakkar, G., & Tadić, M. (2022). Building and Evaluating Universal Named-Entity Recognition English corpus. *arXiv preprint arXiv:2212.07162*.

Alves, D., Bekavac, B., & Tadić, M. (2021, December). Typological Approach to Improve Dependency Parsing for Croatian Language. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)* (pp. 1-11).

Alves, D., Bekavac, B., & Tadić, M. (2021). The Optimization of Portuguese Named-Entity Recognition and Classification by Combining Local Grammars and Conditional Random Fields Trained with a Parsed Corpus. In *Formalising Natural Languages: Applications to Natural Language Processing and Digital Humanities: 14th International Conference, NooJ 2020, Zagreb, Croatia, June 5–7, 2020, Revised Selected Papers 14* (pp. 196-205). Springer International Publishing.

Alves, D., Kuculo, T., Amaral, G., Thakkar, G., & Tadic, M. (2020). UNER: Universal Named-Entity RecognitionFramework. *arXiv preprint arXiv:2010.12406*.

Alves, D., Salimbajevs, A., & Pinnis, M. (2020). Data augmentation for pipeline-based speech translation. In *Human Language Technologies–The Baltic Perspective* (pp. 73-79). IOS Press.

Alves, D., Thakkar, G., & Tadić, M. (2020). Evaluating language tools for fifteen EU-official under-resourced languages. *arXiv preprint arXiv:2010.12428*.

Alves, D., Thakkar, G., & Tadić, M. (2020). Natural Language Processing Chains Inside a Cross-lingual Event-Centric Knowledge Pipeline for European Union Under-resourced Languages. *arXiv preprint arXiv:2010.12433*.

Gottschalk, S., Kacupaj, E., Abdollahi, S., Alves, D., Amaral, G., Koutsiana, E., ... & Thakkar, G. (2021). Oekg: The open event knowledge graph. In *CLEOPATRA 2021 Cross-lingual Event-centric Open Analytics 2021, April 12 2021, Ljubiljana, Slovenia* (Vol. 2829). Aachen, Germany: RWTH Aachen.

Sarajlić, J., Thakkar, G., Alves, D., & Preradović, N. M. (2022). Quotations, Coreference Resolution, and Sentiment Annotations in Croatian News Articles: An Exploratory Study. *arXiv preprint arXiv:2212.07172*.