UNIVERSITY OF GOTHENBURG

DEPARTMENT OF PHILOSOPHY, LINGUISTICS AND THEORY OF SCIENCE

THE LINGUISTIC STRUCTURE OF WIKIPEDIA

A multilingual analysis and comparison of the language used in Wikipedia articles

Patricia Grau Francitorra

Master's Thesis:	30 credits
Programme:	Master's Programme in Language Technology
Level:	Advanced level
Semester and year:	Spring, 2022
Supervisor	Aarne Ranta
Examiner	Peter Ljunglöf
Keywords	Abstract Wikipedia, Syntactic Analysis, Universal Dependencies, Grammatical Framework, UDPipe 2.0, Syntactic Patterns

Abstract

Wikipedia is a great source of knowledge, but due to its open-collaboration nature, it presents some limitations. Namely, the uneven distribution of content, the low overlap in topic coverage, the differences in the comprehensiveness of articles, and the low number of editors. For this reason, the Abstract Wikipedia project has been created; their objective is to construct language-independent (abstract) articles that can be rendered in any language. In this thesis, we have computationally analysed the language used in Wikipedia in order to find similarities between the language used in different articles. To do so, we have syntactically parsed articles of Wikipedia in different languages using UDPipe 2.0 and gathered the languages' recurrent syntactic patterns using Grammatical Framework's GF-UD. Then, we have compared the analyses with cosine similarity in two ways: based on dependency relations and based on linguistic patterns. We have seen that there is a basis for the Abstract Wikipedia project: there are syntactic similarities not only within one language, but also within multiple languages. In addition, we have found that semantically-related topics have a higher similarity than those which are not. Finally, we have gathered syntactic patterns of every language and compared them, which can constitute the basis of the creation of the Renderers for Abstract Wikipedia.

Preface

I would like to thank my supervisor Aarne Ranta for his support and patience while writing this thesis. I would also like to thank Denny Vrandečić and the team in Abstract Wikipedia for allowing me to make my contribution to the project.

Thank you to my classmates for two years of fun, and a very special thank you to my parents, who have supported me during my studies, and to Marc, for always being there.

Contents

1 Introduction

Wikipedia is the seventh most-visited website in the world (Wikimedia, 2022b) and the world's largest reference website (Wikimedia, 2022c). Their ambition is to create current and exhaustive encyclopedias through open collaboration, available in the highest possible number of languages. As of today, that number escalates to more than 300 languages. However, there are differences among the Wikipedias and their contributions. The lead developer of the Wikifunctions project states four main issues regarding the current state of Wikipedia (Vrandečić, 2018b):

1. **Uneven distribution of content**: the number of articles available in different languages varies greatly. The language with the highest coverage, English, has more than 6 million articles, while more than 50 languages, like Cree and Samoan, have only a few hundred articles or less.
2. **Low overlap in topic coverage**: the two most active Wikipedias are the English Wikipedia, with 5.6 million articles by 2018, and the German Wikipedia, with 2.1 million articles. However, only 1.1 million of the topics covered in the German Wikipedia are available in English. In fact, only 100,000 topics are common between the top ten most active Wikipedias.
3. **Differences in the comprehensiveness of articles**: some articles covering "local" topics often have information missing from others. For example, the Catalan Wikipedia for the writer Narcís Oller contains a detailed description of his life and work, while the English Wikipedia only provides two sentences.
4. **Low number of active editors**: more than half of Wikipedias have less than ten active volunteer editors, which poses a challenge to their development and maintenance.

In order to reduce these differences and make much more knowledge available in many more languages, the **Abstract Wikipedia** (AW) project was born. The objective of Abstract Wikipedia is to create "a Wikipedia written in an abstract language to be rendered into any natural language on request" (Vrandečić, 2018a, p. 1). They want to bridge the gap between formal knowledge representation languages and natural languages.

Writing an article about every topic in every desired language would be an arduous process; it would scale the problem to the number of topics multiplied with the number of languages. Instead, their solution is to construct an abstract representation of the topic which can be extended to any language. This would reduce the problem to the number of topics added to the number of languages (Vrandečić, 2018b).

Even if Machine Translation (MT) sounds like an enticing and more straightforward prospect, the developers of the Abstract Wikipedia project believe their idea to preferable. The reason is how differently information is conveyed in every language, especially in regards to morphological markedness or lexical or semantic distinction. Using MT would imply making the source language - which would most likely be English because of its status as interlingua in the scientific community - unnaturally and unnecessarily explicit, to be able to translate it to grammatically correct sentences in other languages.

The proposed solution would consists of three components: Content, Constructors, and Renderers. The Content stores the information of each topic, independently of the language, the Constructors specify the language of the Content, and the Renderers translate the language-independent information into natural language. Abstract Wikipedia should equally be sustained through open collaboration: all parts should be created, refined and maintained by the community, regardless of the language they speak.

In his paper, Vrandečić (2018b) describes the problems, desiderata and constraints of Abstract Wikipedia, and states that there is not a clearly defined solution for the task at hand at the time. However, the approach has changed since then. In the last GF Summer School, Vrandečić says that the current version of Abstract Wikipedia is inspired by Grammatical Framework; they want to apply their ideas in the project, and “use as much as Grammatical Framework as possible” (Vrandečić, 2021, 00:28:00).

1.1 Goals

The presented work aims to serve as groundwork for the Abstract Wikipedia project, by computationally analysing the language used in Wikipedia from a multilingual perspective. Our goal is to find the similarities among languages and topics in Wikipedia, as well as common syntactic patterns among them. This study could serve as a base for developing both the language interpretation of AW.

To do so, we have gathered Wikipedia articles which are available in multiple languages, and we have syntactically analysed them using a dependency-based parser, UDPipe 2.0. Then, we have compared the analyses with cosine similarity based on dependency relations and based on linguistic patterns. In addition, we have gathered the languages' recurrent syntactic patterns using Grammatical Framework's GF-UD and found common syntactic patterns among the languages.

1.2 Outline

Section 2 presents the related work, as well as the two main frameworks used in this thesis: Universal Dependencies (UD) and Grammatical Framework (GF). Section 3 explains how the data was gathered and syntactically analysed, and what algorithms were used to construct the dataset. Section 4 shows the cosine similarity measures within the dataset, as well as the recurring patterns gathered from the data. Section 5 interprets the results, and finally, section 6 contains the conclusions gathered from the results.

2 Background and related work

In this section, we present previous studies of the Wikipedia (section 2.1), as well as some algorithms which use dependency-based analyses (section 2.2). Then, we introduce Universal Dependencies (section 2.3), the basis of the dependency analysis in this thesis, and Grammatical Framework (section 2.4) and its underlying theory.

2.1 Analysing the Wikipedia

The articles of Wikipedia have been used as training data for many language models, the most well-known and commonly used being BERT (Devlin et al., 2019). However, there has not been much research analysing its use of language, nor comparing the different languages in Wikipedia in a big scale. Some studies have focused on comparing two Wikipedias, such as Yasseri et al. (2012), whose goal is to analyse the difference in complexity between the simple English Wikipedia and the main English Wikipedia. Others have focused on studying specific topics, such as Joo (2020), who examined 132 Wikipedia articles related to information users, or Samoilenko & Yasseri (2014), who explored 400 Wikipedia articles on academics from different fields to see if there is any correlation between being in the Wikipedia and academic notability.

Some researchers have studied Wikipedia as a whole, like Massa & Scrinzi (2012) with Manypedia, Bao et al. (2012) with Omnipedia, and Ortega et al. (2008, 2009). The first two are tools that allow the user to explore and compare similarities and differences between the same Wikipedia topic in different languages. Manypedia does so by comparing the content of one article in one language with the content of the same in another one, which can be translated through Machine Translation. In their paper, Massa & Scrinzi talk about the Linguistic Points of View expressed in different languages, which relates to the differences in the comprehensiveness of articles that Abstract Wikipedia is trying to overcome (Vrandečić, 2018b). Omnipedia, on the other hand, shows the differences among various languages by "highlight[ing] the similarities and differences that exist among Wikipedia language editions, and mak[ing] salient information that is unique to each language as well as that which is shared more widely." (Bao et al., 2012, p. 1075).

The work of Ortega et al. (2008) analyzes the contributions of the Wikipedia of the top-ten language editions, based on the total number of articles. They point out that 10% of the total number of authors are responsible for more than 90% of the total number of contributions. The authors also mention that this inequality has been consistent in the last few editions of every language. A similar more in-depth analysis was done later by the same author (Ortega Soto, 2009), using WikiXRay. The difference of contribution is, once more, one of the issues that the Abstract Wikipedia wants to solve Vrandečić (2018b).

In addition, some studies have analysed the gender bias of Wikipedia. They have found that, overall, the majority of editors of Wikipedia are male (Antin et al., 2011; Hill & Shaw, 2013; Wikimedia, 2022a). This might not necessarily be reflected in the language used in Wikipedia - which is the topic of this thesis - but rather in the choice of articles and their length.

2.2 Dependency-based analysis

Several Natural Language Processing (NLP) approaches currently used revolve around the distributional hypothesis (Harris, 1954) or word embeddings. Not many have exploited the use of dependency analysis as its basis, especially not to compare different languages in a big scale.

Erkan et al. (2007) is one of the earliest authors who compare two sentences using dependency parsing, in the field of Biomedicine. In their study, they introduce a way of extracting relations among two (or more) protein names in a sentence. They analyse the sentences using the Stanford Parser, creating a linguistic path from one protein name to the other. Then, given two dependency trees, they calculate the similarity in two ways: using cosine similarity or using edit distance among the paths between the protein names. They argue that "Unlike syntactic parsing, dependency parsing captures the semantic predicate argument relationships between the entities in addition to the syntactic relationships." (Erkan et al., 2007, p. 235). This is similar to the work of Liu & El-Gohary (2017), in the field of Civil Engineering. They present a similarity-based dependency parsing methodology that extracts entities and relations to automate the relation extraction of bridge inspection reports. They represent the dependencies of the sentences based on sentence configurations, and then compare them using cosine similarity.

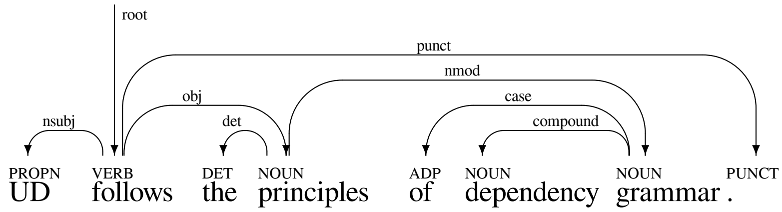
On a broader perspective, there is the model of Levy & Goldberg (2014), who exploited dependency-based word embeddings. Instead of using linear contexts to calculate the embeddings, they use syntactic contexts that are derived from automatically produced dependency parse trees to train a Skip-Gram model. Their results show that dependency-based contexts produce different kinds of similarities than the the Skip-Gram neural embedding model.

There are some language applications that have benefited from dependency-based analysis, such as Multi-Document Summarization (MDS) (Radev et al., 2008), Text Similarity (Inan, 2020), and Question - Answering (QA) (Tran et al., 2015). Radev et al. propose computing sentence similarity for MDS based on dependency parsing of sentences, instead of using a bag-of-words model. They create "bigram units", which represent a branch in a dependency tree, and calculate cosine similarity passing them through their kernels. Inan is also concerned with text similarity. They use SimiT, an unsupervised hybrid method based on: 1) an embedding model that produces sentence representations created with spaCy dependency parser and 2) ConceptNet, a lexicon-based embedding model. They combine both vector representations and calculate soft cosine similarity, obtaining good results. In QA there is JAIST, an answering scoring approach created to solve Task 3 in SemEval2015. One of the features they use is dependency cosine similarity, in which they represent the questions and answers as a bag-of-word-dependency, where words are associated with their dependency relation. A sentence (question) is made of the dependencies of its words, and together with other features, it is then vectorised and used to calculate cosine similarity with another sentence (potential answer). Together with their other features, they achieve good results in the main task.

2.3 Universal Dependencies

In 2016, Nivre et al. developed Universal Dependencies (UD), "an open community effort to create cross-linguistically consistent treebank annotation for many languages within a dependency-based lexicalist framework" (Nivre et al., 2016, p. 1659). Their objective was to support multilingual research by unifying annotation schemes in different languages, creating cross-linguistically consistent morphosyntactic annotation guidelines, as well as corpora following these guidelines. They wanted to explore the parallelism between constructions across different languages, in spite of their typological differences. By 2020, UD includes 183 treebanks for 104 languages, with contributions from more than 400 researchers around the world (de Marneffe et al., 2021d).

UD combines previous initiatives, like the universal Stanford dependencies, an extended version of the Google universal tag set, and a revised subset of the Interset feature inventory. It follows the principles of dependency grammar: a linguistic utterance can be divided into clauses and phrases, which contain a head and elements that ultimately depend on it. It is a binary asymmetrical relation, which is represented with arrow diagrams. The following diagram illustrates these relations with one sentence from this paragraph. The sentence has been parsed with UDPipe 2.0 (explained in section 3.1.1) and printed out using GF-UD (section 3.1.2):



The head of the sentence is the root, “follows” in this case, and all other tokens ultimately depend on it. Multiword units have their own heads whose elements depend on them. For example, the noun phrase (NP) “the principles of dependency grammar” has “principles” as its head, “the” as the determiner and “of dependency grammar” as a prepositional phrase (PP) depending on it. Each arrow in the diagram represents a dependency relation: nsubj, obj, det, nmod, case, compound, punct. There are 37 syntactic relations in UD which can be found in de Marneffe et al. (2021d), page 266.

The treebanks in UD use the CoNLL-U format, where one line represents each token from the sentence, whose information is tab separated. There are 10 columns per line, which represent:

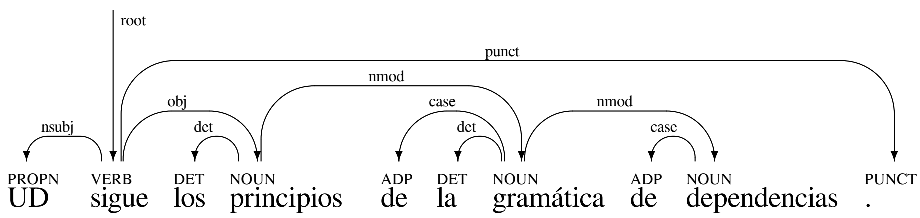
- **ID**: a unique id per each token in the sentence.
- **FORM**: the word form of the token.
- **LEMMA**: the base form of the token, an “abstract representation” (Crystal, 2008) of the word.
- **UPOS**: the universal part of speech tag of the token, of a series of 17 different tags.
- **XPOSTAG**: the optional language-specific part of speech tag of the token.
- **FEATS**: the morphological features of the token.
- **HEAD**: the ID of the token on which the token depends.
- **DEPS**: secondary additional dependencies.
- **DEPREL**: the dependency relation between the token and its head.
- **MISC**: other miscellaneous information of the word, like its range.

Table 1 shows the CoNLL-U structure for the previously analysed sentence.

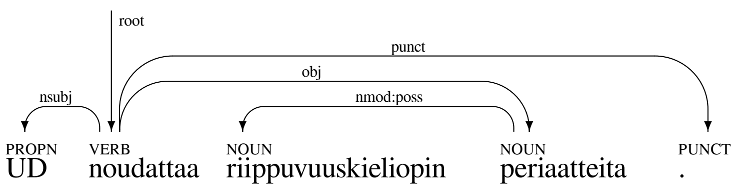
ID	FORM	LEMMA	UPOS	XPOSTAG	FEATS	HEAD	DEPREL	DEPS	MISC
1	UD	UD	PROPN	NNP	Number=Sing	2	nsubj	_	TokenRange=0:2
2	follows	follow	VERB	VBZ	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin	0	root	_	TokenRange=3:10
3	the	the	DET	DT	Definite=Def PronType=Art	4	det	_	TokenRange=11:14
4	principles	principle	NOUN	NNS	Number=Plur	2	obj	_	TokenRange=15:25
5	of	of	ADP	IN	_	7	case	_	TokenRange=26:28
6	dependency	dependency	NOUN	NN	Number=Sing	7	compound	_	TokenRange=29:39
7	grammar	grammar	NOUN	NN	Number=Sing	4	nmod	_	SpaceAfter=No TokenRange=40:47
8	.	.	PUNCT	.	_	2	punct	_	SpaceAfter=No TokenRange=47:48

Table 1: CoNLL-U analysis of *UD follows the principles of dependency grammar.* parsed using UDPipe 2.0.

The translation of the same sentence to Spanish results in this analysis, which uses similar POS tags and dependency relations, even if the order of some dependents changes:



And in Finnish, despite its typological differences with English or Spanish:



2.4 Grammatical Framework

Grammatical Framework (GF) is a programming language for multilingual grammar applications that can both parse and generate language. It defines interlinguas (abstract syntaxes) and reversible mappings from them to individual languages (concrete syntaxes) (Ranta et al., 2020). The specific linguistic structures in the concrete syntaxes are defined in the Resource Grammar Library (RGL), which implements the morphology and syntax of the languages. Its core theory of abstract + concrete syntax relates directly to the aspiration of Abstract Wikipedia.

GF uses abstract syntax trees, which contain the information of dependency trees and phrase structure trees, overlooking word order and lexical items. An abstract syntax tree can generate text in different languages from the lineralisation functions written for those languages. Figure 1 represents the abstract syntax tree for the previously analysed English, Spanish and Finnish sentences. It shows that the same concepts and structures apply to different languages, although not necessarily in the same way.

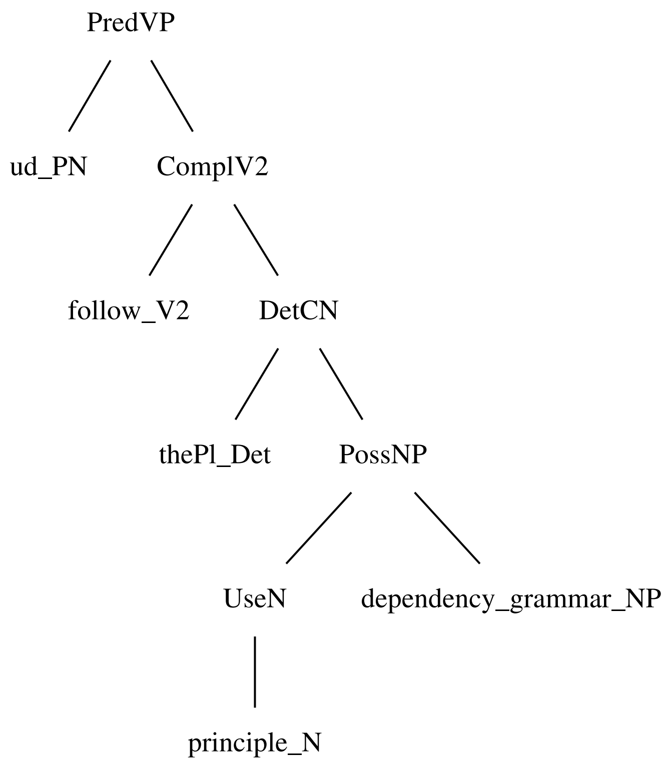


Figure 1: Abstract Syntax Tree of “UD follows the principles of dependency grammar”

Abstract syntax trees assume syntactic parallelisms among languages, similarly to UD. There has been work done to translate from GF to UD (Kolachina & Ranta, 2016) and vice versa (Ranta & Kolachina, 2017), which has been assembled to create GF-UD (Ranta et al., 2022). The algorithms of GF-UD used in the present thesis are explained in section 3.1.2.

3 Materials and Methods

3.1 Dataset

A dataset was generated to analyse the language used in Wikipedia from a multilingual perspective. The data was gathered from Wikipedia's own web page: "Wikipedia articles written in the greatest number of languages" (Wikimedia, 2022d). It contains, at the time of retrieval, 62 articles covering a variety of topics, written in at least 100 of the languages available in Wikipedia. The complete list of topics can be found in Appendix A.

Wikipedia pages contain more information than plain text: there are lists, tables, footnotes, etc. which do not give much information about the language used. For this reason, only raw text (<p>) and titles were extracted, using the Python library BeautifulSoup (Richardson, 2007). The text was first extracted in English, parsed using UDPipe 2.0, and then extracted and parsed in other languages using the same parsing tool. The similarity among articles was calculated using GF-UD's cosine similarity measure. Then, the similarity among languages was calculated using linguistic patterns, and finally, recurrent patterns among the languages were found.

3.1.1 UDPipe 2.0

UDPipe is a trainable pipeline which performs sentence segmentation, tokenization, POS tagging, lemmatization and dependency parsing (Straka & Straková, 2017). It is language-agnostic and can be trained with CoNLL-U data in any language. There are 66 trained models available in UDPipe based on UD treebanks. The complete list of models is found in Appendix B. UDPipe 2.0 is a Prototype presented at the CONLL 2018 UD Shared Task which has yielded great results, greatly surpassing the UDPipe 1.2 baseline. It uses artificial networks, mostly RNNs, and is trained with both CoNLL-U data and pretrained word embeddings (Straka, 2018).

Not all languages available in Wikipedia have a UDPipe model that can parse them. Of the 66 available languages with models, a maximum of 58 were used to syntactically analyse the Wikipedia's topics. If a language has multiple models, one of the most recent ones was chosen to parse the articles in that language. The list of languages used as well as their frequency in the corpus can be found in Appendix C[1]. The raw data from BeautifulSoup was parsed using the UDPipe API (Lindat & CLARIAH-CZ, 2022) and saved according to language and topic.

3.1.2 GF-UD

GF-UD is a software for dependency trees and interlingual syntax which has many features to analyse, visualise, parse, compare and convert trees in different formats. The diagrams of section 2.3 were made using gfud *conll2latex* from a CoNLL-U file.

One of the main GF-UD features used in this project is the cosine similarity measure. GF-UD's cosine similarity "compares two treebanks with respect to feature combinations, by computing the cosine similarity of the two frequency lists" (Ranta et al., 2022). It is necessary to specify what feature combinations GF-UD must look at, such as the surface forms of the words (FORM), their part of speech tag (POS), or the dependency labels (DEPREL). Given the multilingual perspective of the data, the cosine similarity was calculated based on the dependency labels, both among languages and topics.

[1]Even though Norwegian is available in Wikipedia and as a UDPipe model, there is no data in this language (Bokmål or Nynorsk) due to an early fault in the code that has since been solved.

GF-UD can also be used to evaluate a ConLL-U file against a gold standard, with the measure *eval*. If specified with the option *units*, GF-UD shows the scores sentence by sentence, starting from the lowest score, and marking differing lines with a vertical line. An example of such can be found in the following section (3.1.3), where the output of the UDPipe 2.0 parser is compared to a gold standard.

GF-UD is used to extract the linguistic patterns in the parsed data, using *pattern-replace*, *reduced2conll*, and *conll2tree*. *Pattern-replace* replaces or deletes subtrees that satisfy a certain pattern, or flattens trees below a given depth (Ranta et al., 2022). Therefore, it can be used to look for elements that directly depend on other elements, such as the root of a tree. *Reduced2conll* creates CoNLL-U files from data with missing columns, and *conll2tree* returns the data in a hyerarchical structure. These have been used to obtain linguistic patterns in section 3.2.

3.1.3 UDPipe 2.0 Evaluation and Gold Standard

The parser, UDPipe 2.0, was compared against a gold standard in three languages to evaluate its performance. First, a topic was chosen at random: Russia. Considering the length of the topic, only the introductory part was used for the gold standard. Second, the text was gathered using BeautifulSoup and pre-tokenised using UDPipe 2.0 in three languages: English, Spanish and French. Finally, the gold standards were created based on the UD Treebanks for that language, which matched the treebanks of the model of the pre-trained parser.

The treebanks that worked as a base for the gold standard were: EWT for English (Silveira et al., 2021), Ancora for Spanish (Taulé et al., 2008), and GSD for French (de Marneffe et al., 2021c). Table 2 shows the size of the treebanks and their similarities. The three treebanks have a similar number of sentences, although the number of tokens and words does vary substantially depending on the language. A relevant measure for the UDPipe 2.0 evaluation is the number of multi-word tokens: the English treebank has circa 3k multi-word tokens, while the Spanish and French ones have more than three times the amount. However, they seem to use them differently when analysing the tokens. For example, English separates the word "don't" directly into "do" and "n't", whereas Spanish and French would write it first as "don't" in one line, and then separate it into "do" and "n't" in the following lines, if they had the word. The three languages use all UPOS, except for GSD, which is missing *part*. Of the 37 dependency relations of UD, all missing relations from the treebanks are part of the rare relations (based on the distinction of Ranta, 2020).

	EWT	**AnCora**	**GSD**
Sentences	16 621	17 662	16 341
Tokens	251 489	547 203	389 196
Syntactic words	254 825	559 782	400 221
Multi-word tokens	3 333	12 557	11 025
UPOS	17/17	17/17	16/17
Missing relations	clf	clf dislocated goeswith reparandum	clf list

Table 2: Information about UD Treebanks EWT, AnCora and GSD

The creation of the gold standard was a long process and it was developed over two months. The treebanks were used as a reference for the gold standard and served as the last say when there was a doubt, in spite of their possible incongruities. Even though Universal Dependencies wants to follow cross-linguistically consistent morphosyntactic annotations, we have found some interlingual discrepancies among the annotation of structures which a priori seem the same. For instance, the dependency in the noun phrase "Soviet Union" (PROPN + PROPN) is analysed as a compound in English, an adjectival modifier (amod) in French ("Union soviétique", PROP + ADJ) and a flat in Spanish ("Unión Soviética", PROP + PROPN). It makes sense that they are analysed as a compound in English, because it is an endocentric (headed) multi-word expression, and as an amod in French, because they follow the expected structure. However, it is surprising that both Romance languages analyse it differently regarding the POS tags (possibly because of the capitalisation), and that Spanish analyses it differently to English regarding the dependency if we assume the same POS tags. This is only an observation, since this analysis is not the goal of this thesis, but we invite the reader to do a critical study of the UD treebanks and their consistency.

The gold standard file for English contained 25 sentences with an average length of 23.88 tokens, the Spanish gold standard had 29 sentences with an average length of 30.9 tokens, and the French gold standard had 24 sentences with an average sentence length of 31.42 tokens. The gold standard file structure was adapted to the reduced CoNLL-U file, containing only ID, FORM, UPOS, HEAD and DEPREL.

As mentioned in section 3.1.2, GF-UD has an own evaluation measure: *eval*. When called with the option *units*, GF-UD shows the scores sentence by sentence, starting from the lowest score, and marking differing lines with a vertical line. It can be called with the Labelled Attachment Score (LAS) option or the Unlabelled Attachment Score (UAS) option - the first calculates the score based on the head and its label, whereas UAS only looks at the head to calculate similarities. The following is an example of a GF-UD *eval* LAS *units* comparison of a sentence from the gold standard (left) vs. UDPipe (right). It shows the differences between the two analyses with a vertical lines in the 6th, 10th and 11th token:

```
    UDScore {udScore = 0.8333333333333334, udMatching = 1,
        udTotalLength = 12, udSamesLength = 10, udPerfectMatch = 0}
1  In  _  ADP  _  _  2  case                 1  In  _  ADP  _  _  2  case
2  988  _  NUM  _  _  5  obl                 2  988  _  NUM  _  _  5  obl
3  ,  _  PUNCT  _  _  5  punct               3  ,  _  PUNCT  _  _  5  punct
4  it  _  PRON  _  _  5  nsubj               4  it  _  PRON  _  _  5  nsubj
5  adopted  _  VERB  _  _  0  root           5  adopted  _  VERB  _  _  0  root
6  Orthodox  _  ADJ  _  _  7  amod       |   6  Orthodox  _  PROPN  _  _  7  compound
7  Christianity  _  PROPN  _  _  5  obj      7  Christianity  _  PROPN  _  _  5  obj
8  from  _  ADP  _  _  11  case              8  from  _  ADP  _  _  11  case
9  the  _  DET  _  _  11  det                9  the  _  DET  _  _  11  det
10  Byzantine  _  ADJ  _  _  11  amod    |   10  Byzantine  _  PROPN  _  _  11  compound
11  Empire  _  NOUN  _  _  5  obl        |   11  Empire  _  PROPN  _  _  5  obl
12  .  _  PUNCT  _  _  5  punct              12  .  _  PUNCT  _  _  5  punct
```

GF-UD evaluation measure provides the following results, both per sentence and for the total file:

- **udScore**: the score of the sentence or file, calculated diving udSamesLength by udTotalLength.
- **udMatching**: when comparing two sentences, it is 1 if the tokens are the same for both sentences and 0 otherwise. When comparing two files, it returns the sum of all sentence udMatching values.
- **udTotalLength**: the total number of words of the sentences.
- **udSamesLength**: the number of words with matching HEAD and DEPREL.
- **udPerfectMatch**: when comparing two sentences, it is 1 if the sentence is analysed the same in both files and 0 otherwise. When comparing two files, it returns the sum of all sentence udPerfectMatch values.

However, it does not work perfectly with missalignments. This might not be worrying for English data, because there are not that many multi-word tokens in the language (based on the EWT data, table 2). Per contra, languages with a considerable amount of multi-word tokens (like the French words *du*, *des*, *au*, *aux*, etc. or Spanish words *del*, *al*, etc.) might be more affected by it. This is illustrated in the following sentence comparison by GF-UD, taken from the gold standard and parsed text in Spanish:

```
    UDScore {udScore = 0.5789473684210527, udMatching = 0,
        udTotalLength = 19, udSamesLength = 11, udPerfectMatch = 0}
1  Posee  _  VERB  _  _  0  root           1  Posee  _  VERB  _  _  0  root
2  las  _  DET  _  _  4  det               2  las  _  DET  _  _  4  det
3  mayores  _  ADJ  _  _  4  amod          3  mayores  _  ADJ  _  _  4  amod
4  reservas  _  NOUN  _  _  1  obj         4  reservas  _  NOUN  _  _  1  obj
5  de  _  ADP  _  _  6  case               5  de  _  ADP  _  _  6  case
6  recursos  _  NOUN  _  _  4  nmod        6  recursos  _  NOUN  _  _  4  nmod
7  forestales  _  ADJ  _  _  6  amod       7  forestales  _  ADJ  _  _  6  amod
8  y  _  CCONJ  _  _  11  cc               8  y  _  CCONJ  _  _  11  cc
9  la  _  DET  _  _  11  det               9  la  _  DET  _  _  11  det
10  cuarta  _  ADJ  _  _  11  amod         10  cuarta  _  ADJ  _  _  11  amod
11  parte  _  NOUN  _  _  4  conj          11  parte  _  NOUN  _  _  4  conj
12-13  del  _  _  _  _  _  _           |   12  del  _  ADP  _  _  13  case
12  de  _  ADP  _  _  14  case         |   13  agua  _  NOUN  _  _  11  nmod
13  el  _  DET  _  _  14  det          |   14  dulce  _  ADJ  _  _  13  amod
14  agua  _  NOUN  _  _  11  nmod      |   15  sin  _  ADP  _  _  16  mark
15  dulce  _  ADJ  _  _  14  amod      |   16  congelar  _  VERB  _  _  13  acl
16  sin  _  ADP  _  _  17  mark        |   17  del  _  ADP  _  _  18  case
17  congelar  _  VERB  _  _  14  acl  |   18  mundo  _  NOUN  _  _  16  obl
18-19  del  _  _  _  _  _  _           |   19  .  _  PUNCT  _  _  1  punct
```

The sentence "Posee las mayores reservas de recursos forestales y la cuarta parte del agua dulce sin congelar del mundo." can be translated to '[It] has the largest reserves of forest resources and a quarter of the world's unfrozen fresh water.'. It contains "del", which is a contraction of the ADP "de" ('of') and the DET "el" ('the'). According to the AnCora treebank and UD's word segmentation rules (Nivre et al., 2016, p. 1660), such contractions should be separated into its parts. The chosen parser does not separate them, which causes missalignments between the words, and ultimately counts mistakes in correct sentences. In the case of this sentence, all words after the first contraction "del" have the same POS tag, head, and refer to the same element of the sentence in both the gold standard and the parsed text, yet they are considered wrong.

To amend the missalignments, we have created a **new evaluation measure**. It is almost identical to GF-UD's *eval*, but tries to overcome the missalignment issues. Similarly to GF-UD's *eval* LAS *units*, it returns the scores sentence by sentence, starting from the lowest score, and marking differing lines with a vertical line, given the gold standard and the text to be parsed. It also returns the same metrics as GF-UD's *eval* micro LAS measure, both per sentence and for the whole file. Instead of only looking at the ID, it compares the lines of two files based on the head that they refer to. In addition, it can work with bad sentence tokenisation, when the parsed sentences have been split into more parts than the sentences in the gold standard. The previous sentence is evaluated here with the new measure:

```
    # UDScore {udScore = 0.8571428571428571, udMatching = 1,
        udTotalLength = 21, udSamesLength = 18, udPerfectMatch = 0}
1  Posee  _  VERB  _  _  0  root               1  Posee  _  VERB  _  _  0  root
2  las  _  DET  _  _  4  det                   2  las  _  DET  _  _  4  det
3  mayores  _  ADJ  _  _  4  amod              3  mayores  _  ADJ  _  _  4  amod
4  reservas  _  NOUN  _  _  1  obj             4  reservas  _  NOUN  _  _  1  obj
5  de  _  ADP  _  _  6  case                   5  de  _  ADP  _  _  6  case
6  recursos  _  NOUN  _  _  4  nmod            6  recursos  _  NOUN  _  _  4  nmod
7  forestales  _  ADJ  _  _  6  amod           7  forestales  _  ADJ  _  _  6  amod
```

```
8  y  _  CCONJ  _  _  11  cc                         8  y  _  CCONJ  _  _  11  cc
9  la  _  DET  _  _  11  det                         9  la  _  DET  _  _  11  det
10  cuarta  _  ADJ  _  _  11  amod                   10  cuarta  _  ADJ  _  _  11  amod
11  parte  _  NOUN  _  _  4  conj                    11  parte  _  NOUN  _  _  4  conj
12-13  del  _  _  _  _  _  _                      |  12  del  _  ADP  _  _  13  case
12  de  _  ADP  _  _  14  case
13  el  _  DET  _  _  14  det
14  agua  _  NOUN  _  _  11  nmod                    13  agua  _  NOUN  _  _  11  nmod
15  dulce  _  ADJ  _  _  14  amod                    14  dulce  _  ADJ  _  _  13  amod
16  sin  _  ADP  _  _  17  mark                      15  sin  _  ADP  _  _  16  mark
17  congelar  _  VERB  _  _  14  acl                16  congelar  _  VERB  _  _  13  acl
18-19  del  _  _  _  _  _  _                      |  17  del  _  ADP  _  _  18  case
18  de  _  ADP  _  _  20  case
19  el  _  DET  _  _  20  det
20  mundo  _  NOUN  _  _  11  nmod                |  18  mundo  _  NOUN  _  _  16  obl
21  .  _  PUNCT  _  _  1  punct                      19  .  _  PUNCT  _  _  1  punct
```

The new evaluation measure addresses missalignments in three cases: **morphologically disparity**: when a word has not been morphologically separated in the parsed text (such as the previous sentence); **extra split**: when a word has been split into more pieces in the parsed text, compared to the gold standard; and **no split**: when a word has not been split in the parsed text, but it has in the gold standard. Examples of *extra split* and *no split* can be found in Appendix D.

3.2 Recurring patterns

Recurring patterns were found using shell commands and GF-UD. First, all parsed files of the same language were concatenated into a single file. Then, the sentences were pruned on a top-level, keeping only the root and the head of the elements that directly depend on the root. The following is an example of the pruning on a top level. From the sentence "The human body contains from 55% to 78% water, depending on body size.", extracted from the topic "Water" in English, we get[2]:

```
## PRUNE TRUE 1
# sent_id = 306
# text = The human body contains from 55% to 78% water, depending on body size.
# newtext = body contains % water , size .
1       body     body     NOUN     NN       _      2       nsubj   _       ADJUSTED=True
2       contains     contain VERB    VBZ _      0       root    _       ADJUSTED=True
                                                                         |ORIG_LABEL=root
3       %        %        SYM      NN       _      2       obl     _       ADJUSTED=True
4       water    water    NOUN     NN       _      2       obl     _       ADJUSTED=True
5       ,        ,        PUNCT    ,        _      2       punct   _       ADJUSTED=True
6       size     size     NOUN     NN       _      2       obl     _       ADJUSTED=True
7       .        .        PUNCT    .        _      2       punct   _       ADJUSTED=True
```

After pruning, only the root and the head of the elements that directly depend on the root are kept: "body" as the nsubj, the root "contains", "%" for the first oblique, "water" for the second oblique, "size" for the adjunct, and the punctuation mark. Because we are interested in structures and not in word forms, the output is reduced to its ID, UPOS, HEAD and DEPREL columns. Then, the result is hierarchically ordered, leaving the root before all other elements that depend on it.

[2]Because of space limitations, morphological and miscellaneous information have been omitted

```
2       _       _       VERB    _       _       0       root    _       _
    1   _       _       NOUN    _       _       2       nsubj   _       _
    3   _       _       SYM     _       _       2       obl     _       _
    4   _       _       NOUN    _       _       2       obl     _       _
    5   _       _       PUNCT   _       _       2       punct   _       _
    6   _       _       NOUN    _       _       2       obl     _       _
    7   _       _       PUNCT   _       _       2       punct   _       _
```

Finally, the output is further reduced into UPOS and DEPREL, thus leaving linguistic information that can be analysed cross-linguistically. For each parsed sentence, we obtain the root with its part of speech tag followed by all other dependencies and their part of speech tags. Because we want to disregard word order, the non-root elements are sorted alphabetically in a future step, before obtaining the results.

```
VERB    root
NOUN    nsubj
SYM     obl
NOUN    obl
PUNCT   punct
NOUN    obl
PUNCT   punct
```

The recurring pattern distributions per language can be found in the GitHub repository. They show, per language, the structures that make up the text, and their frequency in the text, in descending order. These distributions can be used in the future to create the languages used for Abstract Wikipedia, and could be built with Grammatical Framework.

4 Results

This sections contains the results of the UDPipe 2.0 evaluation (section 4.1), and the comparative analysis, first based in dependency relations (section 4.2) and second, in linguistic patterns (section 4.3). Additionally, it presents the recurring syntactic patterns found in the analysed languages, and their distribution (section 4.4).

4.1 UDPipe 2.0 Evaluation

Using the new evaluation measure and the gold standards explained in section 3.1.3, we evaluate UDPipe 2.0. The parsing of the sentences get the results shown in table 3.

	udScore from GF-UD	**udScore from new evaluation measure**
English	0.8777	0.8848
Spanish	0.8470	0.8317
French	0.8645	0.9086

Table 3: Evaluation of UDPipe 2.0 parsing - Labelled Attachment Scores (LAS)

Overall, the results of the Evaluation of UDPipe 2.0 are quite high, achieving a minimum of 0.83 and a maximum of 0.9. From the table we see that the results from GF-UD evaluation measure and the new evaluation measure do not vary considerably. Nonetheless, we believe that the new evaluation can be helpful when comparing two files in detail. In addition, it yields better results than GF-UD's *eval* when analysing data with a high number of missalignments. It also repairs bad sentence tokenisation when the parsed sentences have been split into more parts than the sentences in the gold standard, and shows a message with its occurrence.

4.2 DEPREL-based cosine similarity

As detailed in section 3.1.2, the cosine similarity measure was calculated using GF-UD's cosine similarity measure based on the dependency labels (DEPREL). It was computed interlinguistically, comparing the topic among different languages, and intralinguistically, comparing the different topics available for each language. Section 4.2.1 refers to the interlinguistic comparison, and 4.2.2 refers to the intralinguistic one. The full data of DEPREL-based cosine similarity per topic and per language can be found in the project's GitHub repository.

4.2.1 DEPREL-based cosine similarity per Wikipedia topic

Table 4 presents the maximum, minimum, and average cosine similarity values calculated on dependency labels per Wikipedia topic. It also contains the languages of the files that were compared when calculating the cosine similarity. Every line represents a topic. For instance, the first topic is "Adolf Hitler", which has received a maximum cosine similarity value of 0.9928 when comparing the Catalan and Spanish file on the topic, a minimum cosine similarity value of 0.2138 when comparing Japanese and Sanskrit, and an average similarity of 0.7383.

	Max sim	Max lang	Min sim	Min lang	Avg sim
Adolf Hitler	0.9928	Catalan, Spanish	0.2138	Japanese, Sanskrit	0.7383
Africa	0.9960	Catalan, Spanish	0.1238	Gothic, Hungarian	0.7105
Asia	0.9896	Belarusian, Ukrainian	0.1076	Gothic, Hungarian	0.7125
Association Football	0.9968	Catalan, Spanish	0.1195	Gothic, Hungarian	0.7208
Barack Obama	0.9887	Catalan, Spanish	0.1396	Japanese, Sanskrit	0.6949
Bible	0.9895	Catalan, Spanish	0.2033	Classical Chinese, Japanese	0.7521
Buddha	0.9892	Czech, Slovak	0.0926	Gothic, Hungarian	0.7435
Buddhism	0.9890	Belarusian, Ukrainian	0.1734	Japanese, Sanskrit	0.7623
China	0.9948	Belarusian, Ukrainian	0.1296	Gothic, North Sami	0.7045
Christianity	0.9952	Catalan, Spanish	0.1638	Japanese, Sanskrit	0.7344
Christmas	0.9870	Catalan, Spanish	0.1624	Japanese, Sanskrit	0.7299
Dog	0.9962	Catalan, Spanish	0.1416	Gothic, Hungarian	0.7166
Earth	0.9981	Catalan, Spanish	0.1180	Gothic, Hungarian	0.7265
English Language	0.9936	Belarusian, Ukrainian	0.0394	Gothic, Hungarian	0.7139
Europe	0.9937	Catalan, Spanish	0.1657	Japanese, Sanskrit	0.7139
Eye	0.9965	Catalan, Spanish	0.1296	Gothic, Hungarian	0.7023
George W	0.9923	Czech, Slovak	0.0816	Latin, Sanskrit	0.6590
Ghana	0.9929	Catalan, Spanish	0.1766	Kazakh, Sanskrit	0.7045
Gold	0.9915	Czech, Slovak	0.1619	Japanese, Sanskrit	0.7240
Hinduism	0.9970	Croatian, Serbian	0.0525	Gothic, Hungarian	0.7586
Human	0.9912	Catalan, Spanish	0.1770	Japanese, Sanskrit	0.7486
India	0.9983	Catalan, Spanish	0.1274	Gothic, Hungarian	0.7158
Internet	0.9919	Catalan, Spanish	0.2012	Japanese, Sanskrit	0.7387
Iran	0.9945	Catalan, Spanish	0.0891	Galician, Sanskrit	0.6988

Iraq	0.9939	Catalan, Spanish	0.1193	Galician, Sanskrit	0.6975
Iron	0.9983	Catalan, Spanish	0.1469	Japanese, Sanskrit	0.7392
Islam	0.9965	Czech, Slovak	0.1238	Gothic, Hungarian	0.7418
Italy	0.9891	Belarusian, Ukrainian	0.0857	Gothic, Hungarian	0.7108
Japan	0.9942	Belarusian, Ukrainian	0.0762	Gothic, Kazakh	0.7063
Jesus	0.9904	Catalan, Spanish	0.0418	Gothic, Hungarian	0.7343
Judaism	0.9933	Catalan, Spanish	0.2179	Classical Chinese, Japanese	0.7563
Julius Caesar	0.9915	Belarusian, Ukrainian	0.0690	Gothic, Sanskrit	0.7180
Koran	0.9903	Catalan, Spanish	0.1566	Japanese, Sanskrit	0.7488
Maize	0.9881	Catalan, Spanish	0.2728	Hungarian, Japanese	0.7642
Milk	0.9962	Catalan, Spanish	0.1207	Gothic, Hungarian	0.7169
Mohandas Karamchand Gandhi	0.9956	Czech, Slovak	0.1779	Japanese, Sanskrit	0.7398
Money	0.9837	Catalan, Spanish	0.0308	Gothic, Hungarian	0.6843
Moon	0.9956	Catalan, Spanish	0.1814	Japanese, Sanskrit	0.7337
Moses	0.9880	Catalan, Spanish	0.3501	Hungarian, Japanese	0.7918
Muhammad	0.9909	Czech, Slovak	0.0879	Gothic, Hungarian	0.7489
New York City	0.9964	Catalan, Spanish	0.0602	Gothic, Hungarian	0.6897
Niger	0.9930	Catalan, Spanish	0.1468	Sanskrit, Wolof	0.6969
Osama Bin Laden	0.9849	Catalan, Spanish	0.1510	Japanese, Sanskrit	0.6884
Paris	0.9983	Catalan, Spanish	0.0671	Gothic, Hungarian	0.6954
Periodic Table	0.9916	Belarusian, Ukrainian	0.1039	Latin, Uyghur	0.7343
Pope Benedict Xvi	0.9879	Catalan, Spanish	0.2579	Japanese, Marathi	0.7251
Pope John Paul Ii	0.9851	Catalan, Spanish	0.1676	Latin, Urdu	0.7129
Religion	0.9976	Croatian, Serbian	0.1613	Japanese, Sanskrit	0.7532
Rice	0.9952	Catalan, Spanish	0.1519	Japanese, Sanskrit	0.7395

Roman Catholic Church	0.9907	Catalan, Spanish	0.2371	Japanese, Telugu	0.7265
Rome	0.9938	Catalan, Spanish	0.1437	Galician, Sanskrit	0.7008
Russia	0.9953	Catalan, Spanish	0.1178	Gothic, Hungarian	0.7084
Silver	0.9859	Catalan, Spanish	0.1280	Old Church Slavonic, Uyghur	0.7153
South Africa	0.9972	Catalan, Spanish	0.0901	Gothic, Wolof	0.7013
South America	0.9901	Catalan, Spanish	0.1656	Galician, Sanskrit	0.7103
Soviet Union	0.9934	Croatian, Serbian	0.1065	Gothic, Hungarian	0.7009
Sun	0.9974	Catalan, Spanish	0.1246	Gothic, Hungarian	0.7056
United Kingdom	0.9944	Belarusian, Ukrainian	0.2332	Hungarian, Urdu	0.7221
United States	0.9956	Catalan, Spanish	0.0646	Gothic, Hungarian	0.7082
Water	0.9948	Catalan, Spanish	0.0658	Gothic, Hungarian	0.7163
Wikipedia	0.9954	Catalan, Spanish	0.1039	Japanese, Sanskrit	0.7056
World War Ii	0.9924	Czech, Slovak	0.0873	Japanese, Sanskrit	0.7054

Table 4: Cosine similarity based on DEPREL per Wikipedia topic

The maximum cosine similarity per topic is quite high, reaching a total maximum of 0.9983 in the topic of "Paris" between Catalan vs. Spanish (marked in blue in the table). 43 out of 62 times, the language comparison which achieves the highest cosine similarity measure is Catalan and Spanish. 9 times, it is Belarusian vs. Ukrainian; 7 times, Czech vs. Slovak; and 3 times, Croatian vs. Serbian. The average maximum value is 0.9928, and the lowest maximum is 0.9837.

The minimum cosine similarity per topic has a mean of 0.1369, and a maximum value of 0.3501. The total minimum is 0.0308 in the topic of "Money" when comparing Gothic vs. Hungarian (marked in yellow in the table). The comparison of these two languages get the minimum cosine similarity value a total of 23 times, followed by Japanese and Sanskrit, which get the lowest value 18 times, and Galician and Sanskrit, which happen 4 times. The other language comparison that cause a minimum cosine similarity value per topic occur 2 or less times. The average cosine similarity per topic is a range between 0.6590 - 0.7918, and the average cosine similarity among all of them is 0.7213.

4.2.2 DEPREL-based cosine similarity per Language

Table 5 presents the maximum, minimum, and average cosine similarity values calculated on dependency labels per language analysed. In addition, it contains the topic comparison that caused the maximum and minimum cosine similarity. Every line of the table represents a language.

	Max sim	Max lang	Min sim	Min lang	Avg sim
Afrikaans	0.9992	Iran, South Africa	0.5646	Association Football, Religion	0.9540
Arabic	0.9990	Gold, Silver	0.9023	Muhammad, Soviet Union	0.9794
Armenian	0.9976	Italy, United Kingdom	0.7021	George W. Bush, South Africa	0.9544
Basque	0.9974	Europe, United States	0.7796	English Language, Osama Bin Laden	0.9605
Belarusian	0.9982	China, South Africa	0.8031	Money, Osama Bin Laden	0.9596
Bulgarian	0.9983	South Africa, United States	0.8746	Association Football, World War II	0.9759
Catalan	0.9986	Eye, Milk	0.9084	English Language, Money	0.9790
Chinese	0.9967	Japan, United States	0.8518	China, Moses	0.9653
Classical Chinese	0.9967	China, Japan	0.6790	Human, South Africa	0.9437
Croatian	0.9972	Italy, United States	0.8362	Islam, Osama Bin Laden	0.9594
Czech	0.9985	India, Japan	0.8859	Osama Bin Laden, Roman Catholic Church	0.9721
Danish	0.9975	China, Soviet Union	0.8338	Asia, Pope Benedict XVI	0.9709
Dutch	0.9976	Italy, United States	0.8813	Barack Obama, Sun	0.9709
English	0.9987	Moon, Sun	0.9016	Julius Caesar, New York City	0.9789
Estonian	0.9952	India, United States	0.7429	Pope John Paul II, South America	0.9376
Finnish	0.9965	China, Iran	0.8676	Asia, Osama Bin Laden	0.9648
French	0.9991	India, Iran	0.9458	George W. Bush, Hinduism	0.9896
Galician	0.9994	China, India	0.9196	Iran, Moses	0.9843
German	0.9986	Iran, Russia	0.9116	Rome, Sun	0.9807
Gothic	0.9956	New York City, United States	0.3261	Jesus, South America	0.7577
Greek	0.9985	China, Russia	0.8917	Julius Caesar, Rice	0.9774
Hebrew	0.9987	Iran, Italy	0.8576	Roman Catholic Church, Silver	0.9774
Hindi	0.9990	Africa, South America	0.6068	Niger, Osama Bin Laden	0.9408
Hungarian	0.9988	Moon, Sun	0.8758	English Language, Eye	0.9787

Indonesian	0.9979	Christianity, Judaism	0.7741	Osama Bin Laden, Rice	0.9474
Irish	0.9940	Koran, Sun	0.5729	Eye, Human	0.8935
Italian	0.9993	China, Russia	0.9055	English Language, World War II	0.9850
Japanese	0.9995	Barack Obama, George W. Bush	0.9123	Human, Osama Bin Laden	0.9849
Kazakh	0.9992	Human, South America	0.6408	Ghana, Osama Bin Laden	0.9661
Korean	0.9976	Russia, United States	0.7852	Pope John Paul II, Religion	0.9545
Latin	0.9957	India, Paris	0.3312	Maize, Periodic Table	0.8497
Latvian	0.9971	India, Russia	0.7414	Osama Bin Laden, Water	0.9562
Lithuanian	0.9982	Asia, Europe	0.8072	Koran, Osama Bin Laden	0.9624
Maltese	0.9937	Bible, Europe	0.4955	Osama Bin Laden, Wikipedia	0.8956
Marathi	0.9966	India, United States	0.7681	Europe, Moses	0.9517
North Sami	0.9922	Asia, Europe	0.3700	Iraq, Wikipedia	0.8106
Old Church Slavonic	0.9870	Italy, Russia	0.3665	Christianity, Silver	0.8564
Persian	0.9978	India, Japan	0.9091	Silver, South America	0.9749
Polish	0.9966	Iran, Soviet Union	0.7832	Maize, Rice	0.9681
Portuguese	0.9996	India, Iran	0.9393	Barack Obama, Religion	0.9876
Romanian	0.9976	Italy, Russia	0.7435	George W. Bush, Money	0.9588
Russian	0.9985	Iraq, South Africa	0.8949	Osama Bin Laden, Water	0.9777
Sanskrit	0.9991	Asia, South America	0.4339	English Language, Iran	0.8992
Scottish Gaelic	0.9874	English Language, Japan	0.5830	Africa, Buddha	0.8769
Serbian	0.9979	China, United States	0.8654	Osama Bin Laden, Religion	0.9695
Slovak	0.9984	China, Japan	0.7998	Asia, Barack Obama	0.9604
Slovenian	0.9968	China, Italy	0.7203	English Language, Pope John Paul II	0.9515
Spanish	0.9988	Russia, South Africa	0.9235	Asia, Barack Obama	0.9824
Swedish	0.9977	Islam, Judaism	0.8787	Money, Pope John Paul Ii	0.9696

Tamil	0.9979	Iraq, New York City	0.7920	Bible, George W. Bush	0.9636
Telugu	0.9982	Iran, Iraq	0.4820	George W. Bush, Roman Catholic Church	0.9314
Turkish	0.9976	Moon, Sun	0.7113	Eye, George W. Bush	0.9585
Ukrainian	0.9987	Silver, Water	0.8806	Barack Obama, Europe	0.9765
Urdu	0.9965	Jesus, Moses	0.5210	English Language, Pope John Paul Ii	0.9263
Uyghur	0.9904	Islam, Gandhi	0.4292	Eye, Periodic Table	0.8332
Vietnamese	0.9991	India, Italy	0.8835	Koran, New York City	0.9768
Welsh	0.9969	South Africa, United States	0.6765	Association Football, Money	0.9368
Wolof	0.9997	Ghana, Niger	0.2582	English Language, Water	0.6791

Table 5: Cosine similarity based on DEPREL per language

The maximum DEPREL-based cosine similarity per language is also quite high, with a value of 0.9997. It is found in Wolof, when comparing the topics of “Ghana” and “Niger” (marked in blue in the table). When comparing topics within a language, there is not a combination of topics that clearly point to a high cosine similarity measure. The topics “Moon” and “Sun” get the maximum value 3 out of 62 times, and all other combinations that get the maximum value 2 out of 62 times are related to places: ‘South Africa” and “United States”, “China” and “Japan”, “Italy” and “United States”, “India” and “United States”, “India” and “Iran”, “China” and “Russia”, “Asia” and “Europe”, and “Italy” and “Russia”. The average maximum value is 0.9972, and the lowest maximum is 0.9870.

The minimum cosine similarity per language has a mean of 0.7401, and a maximum value of 0.9458. The total minimum also happens within the Wolof language: its value is 0.2582, when comparing the topics “English Language” and “Water” (in yellow, in the table). The topics which were compared more often when achieving the lowest cosine similarity measure per language were: “Osama Bin Laden” and “Water”, “Asia” and “Barack Obama”, and “English Language” and “Pope John Paul II”. Each one of these get the lowest value 2 out of 62 times. The average cosine similarity per language is a range between 0.6791 - 0.9896, and the average cosine similarity among all of them is 0.9429.

4.3 Pattern-based cosine similarity

The linguistic patterns were extracted for every language using the methodology explained in section 3.2. They contain both the POS tags of the tokens and their dependency relations, and are available at the project’s GitHub repository. Based on the patterns and their frequency, a vector was calculated that represented each language. The vector is of size 241 998, which is the number of patterns found across all languages. The number in each dimension represents the amount of times a language presents this pattern. Each language vector was compared to the other language vectors using PyTorch’s cosine similarity measure, creating a matrix whose full data is available here. We use another measure of comparison because GF-UD’s cosine similarity measure is not ready to be used with the patterns extracted previously.

Table 6 presents a summarised version of the full matrix. Every line shows the name of the language, the language family to which it belongs (based on Ager (2022)), the three languages compared to which it got the highest cosine similarity, the three languages compared to which it got the lowest, and its average cosine similarity. For instance, Afrikaans is a Germanic Language that got its highest cosine similarity value with Italian, Dutch and Danish; its lowest, with Classical Chinese, Wolof and Arabic; and it has an average cosine similarity among all languages of 0.5881.

Language	Lang fam	Max sim	Min sim	Avg sim
Afrikaans	Indo-European languages, Germanic languages	0.7843 Italian	0.1087 Cl. Chinese	0.5881
		0.7703 Dutch	0.1742 Wolof	
		0.7636 Danish	0.187 Arabic	
Arabic	Afroasiatic languages, Semitic languages	0.2361 Lithuanian	0.0156 Cl. Chinese	0.1833
		0.224 Czech	0.0513 Kazakh	
		0.222 Slovak	0.0567 Wolof	
Armenian	Indo-European languages, Armenian languages	0.7326 Serbian	0.0348 Cl. Chinese	0.4766
		0.6894 Danish	0.1176 Wolof	
		0.6818 Slovenian	0.1392 Arabic	
Basque	Language isolates, Language isolates	0.8002 Latin	0.0664 Cl. Chinese	0.6028
		0.7862 Swedish	0.1858 Arabic	
		0.7703 Danish	0.2285 Kazakh	
Belarusian	Indo-European languages, Slavic languages	0.9387 Ukrainian	0.041 Cl. Chinese	0.5893
		0.8828 Russian	0.1553 Sanskrit	
		0.8706 Latvian	0.1678 Wolof	
Bulgarian	Indo-European languages, Slavic languages	0.8454 Italian	0.068 Cl. Chinese	0.6088
		0.8177 Slovak	0.1881 Kazakh	
		0.8065 Spanish	0.1892 Wolof	
Catalan	Indo-European languages, Romance languages	0.9238 Spanish	0.0539 Cl. Chinese	0.6342
		0.8673 Romanian	0.1981 Kazakh	
		0.8602 Italian	0.2069 Arabic	
Chinese	Sino-Tibetan languages, Sinitic (Chinese) languages	0.7192 Latin	0.162 Cl. Chinese	0.5306
		0.7149 Basque	0.1652 Kazakh	
		0.6826 Swedish	0.1756 Arabic	
Classical Chinese	Sino-Tibetan languages, Sinitic (Chinese) languages	0.2289 Vietnamese	0.0156 Arabic	0.0801
		0.162 Chinese	0.021 Kazakh	
		0.1182 Marathi	0.0286 Indonesian	
Croatian	Indo-European languages, Slavic languages	0.9458 Serbian	0.0493 Cl. Chinese	0.6271
		0.8772 Slovenian	0.1951 Kazakh	
		0.8763 Italian	0.2047 Arabic	
Czech	Indo-European languages, Slavic languages	0.9457 Slovak	0.0406 Cl. Chinese	0.6343
		0.911 Polish	0.1654 Wolof	
		0.87 Ukrainian	0.1822 Kazakh	
Danish	Indo-European languages, Germanic languages	0.9208 Swedish	0.0572 Cl. Chinese	0.6804
		0.8995 Dutch	0.1736 Wolof	
		0.8966 Finnish	0.1992 Kazakh	
Dutch	Indo-European languages, Germanic languages	0.8995 Danish	0.0595 Cl. Chinese	0.6747
		0.8991 Swedish	0.2045 Arabic	
		0.8983 Italian	0.2045 Kazakh	
English	Indo-European languages, Germanic languages	0.8621 Italian	0.0474 Cl. Chinese	0.6132
		0.8525 Dutch	0.1689 Kazakh	
		0.8229 German	0.1747 Wolof	
Estonian	Uralic languages, Finnic languages	0.9302 Finnish	0.0617 Cl. Chinese	0.662

		0.8931 Danish	0.1963 Kazakh	
		0.8815 Swedish	0.2036 Arabic	
Finnish	Uralic languages, Finnic languages	0.9302 Estonian	0.0598 Cl. Chinese	0.6498
		0.8966 Danish	0.1693 Wolof	
		0.8668 Swedish	0.1966 Arabic	
French	Indo-European languages, Romance languages	0.8263 Italian	0.0527 Cl. Chinese	0.5976
		0.811 Catalan	0.195 Kazakh	
		0.8007 Dutch	0.2135 Arabic	
Galician	Indo-European languages, Romance languages	0.8194 Italian	0.0572 Cl. Chinese	0.6224
		0.8034 Swedish	0.192 Kazakh	
		0.802 Danish	0.1981 Wolof	
German	Indo-European languages, Germanic languages	0.8953 Dutch	0.0497 Cl. Chinese	0.6262
		0.8515 Italian	0.172 Kazakh	
		0.8287 Danish	0.1906 Wolof	
Gothic	Indo-European languages, Germanic languages	0.5242 Old Church Slavonic	0.0717 Arabic	0.2782
		0.4558 Latin	0.0859 Indonesian	
		0.4057 Chinese	0.0882 Wolof	
Greek	Indo-European languages, Hellenic languages	0.8399 Swedish	0.0495 Cl. Chinese	0.6209
		0.837 Ukrainian	0.1805 Wolof	
		0.8367 Russian	0.2062 Sanskrit	
Hebrew	Afroasiatic languages, Semitic languages	0.8424 Russian	0.0337 Cl. Chinese	0.4554
		0.8334 Belarusian	0.0954 Sanskrit	
		0.8078 Ukrainian	0.0985 Wolof	
Hindi	Indo-European languages, Indo-Iranian languages	0.7134 Urdu	0.0503 Cl. Chinese	0.4927
		0.7112 Basque	0.1469 Arabic	
		0.6261 Croatian	0.1922 Kazakh	
Hungarian	Uralic languages, Finno-Ugric languages	0.4685 Hindi	0.0305 Cl. Chinese	0.3487
		0.4486 Indonesian	0.0847 Arabic	
		0.4368 Lithuanian	0.1472 Kazakh	
Indonesian	Austronesian languages, Malayo-Polynesian languages	0.6222 English	0.0286 Cl. Chinese	0.3597
		0.5702 Spanish	0.0618 Sanskrit	
		0.5514 Greek	0.0828 Kazakh	
Irish	Indo-European languages, Celtic languages	0.7327 Swedish	0.0598 Cl. Chinese	0.5504
		0.7305 Turkish	0.1663 Arabic	
		0.7262 Estonian	0.2378 Kazakh	
Italian	Indo-European languages, Romance languages	0.8983 Dutch	0.0518 Cl. Chinese	0.6727
		0.8763 Croatian	0.2137 Kazakh	
		0.8691 Swedish	0.2191 Arabic	
Japanese	Japonic languages, Japonic languages	0.7693 Korean	0.055 Cl. Chinese	0.5224
		0.7279 Basque	0.1566 Arabic	
		0.7053 Estonian	0.2256 Wolof	
Kazakh	Turkic languages, Turkic languages	0.3136 Uyghur	0.021 Cl. Chinese	0.2054
		0.3117 Turkish	0.0507 Wolof	
		0.2851 Marathi	0.0513 Arabic	
Korean	Koreanic languages, Koreanic languages	0.7693 Japanese	0.0654 Cl. Chinese	0.5269
		0.7442 Basque	0.1511 Arabic	
		0.7321 Latin	0.187 Kazakh	
Latin	Indo-European languages, Italic languages	0.8456 Swedish	0.0969 Cl. Chinese	0.6226
		0.8002 Basque	0.1805 Wolof	
		0.7922 Dutch	0.1998 Arabic	

Latvian	Indo-European languages, Baltic languages	0.8984 Ukrainian 0.8795 Russian 0.8706 Belarusian	0.0515 Cl. Chinese 0.1868 Kazakh 0.1901 Sanskrit	0.6309
Lithuanian	Indo-European languages, Baltic languages	0.8676 Swedish 0.8455 Danish 0.8361 Dutch	0.0608 Cl. Chinese 0.2073 Kazakh 0.2297 Wolof	0.6549
Maltese	Afroasiatic languages, Semitic languages	0.7788 Belarusian 0.7644 Ukrainian 0.7607 Greek	0.0494 Cl. Chinese 0.1358 Wolof 0.1607 Sanskrit	0.5582
Marathi	Indo-European languages, Indo-Iranian languages	0.7654 Latin 0.751 Swedish 0.7034 Danish	0.1182 Cl. Chinese 0.1364 Wolof 0.171 Arabic	0.5598
North Sami	Uralic languages, Sámi languages	0.6435 Swedish 0.6381 Latin 0.6308 Estonian	0.1037 Cl. Chinese 0.1306 Arabic 0.1395 Wolof	0.4896
Old Church Slavonic	Indo-European languages, Slavic languages	0.6537 Sanskrit 0.5568 Korean 0.5404 Latin	0.1002 Arabic 0.1078 Cl. Chinese 0.1733 Hebrew	0.3995
Persian	Indo-European languages, Indo-Iranian languages	0.7438 Swedish 0.7064 Latin 0.6999 Danish	0.0462 Cl. Chinese 0.1417 Wolof 0.1797 Arabic	0.5438
Polish	Indo-European languages, Slavic languages	0.911 Czech 0.8912 Slovak 0.8092 Ukrainian	0.0384 Cl. Chinese 0.1486 Wolof 0.1688 Kazakh	0.6075
Portuguese	Indo-European languages, Romance languages	0.675 English 0.6615 French 0.6597 Catalan	0.0397 Cl. Chinese 0.134 Kazakh 0.1522 Arabic	0.4898
Romanian	Indo-European languages, Romance languages	0.8673 Catalan 0.8673 Italian 0.8441 Serbian	0.053 Cl. Chinese 0.2113 Arabic 0.2124 Kazakh	0.6428
Russian	Indo-European languages, Slavic languages	0.901 Ukrainian 0.8828 Belarusian 0.8795 Latvian	0.0338 Cl. Chinese 0.1379 Wolof 0.1505 Sanskrit	0.5564
Sanskrit	Indo-European languages, Indo-Iranian languages	0.6537 Old Church Slavonic 0.3507 Marathi 0.2914 Latin	0.0505 Cl. Chinese 0.0609 Arabic 0.0618 Indonesian	0.2367
Scottish Gaelic	Indo-European languages, Celtic languages	0.6702 Latin 0.6658 Swedish 0.6285 Basque	0.0912 Cl. Chinese 0.1541 Arabic 0.1777 Kazakh	0.4993
Serbian	Indo-European languages, Slavic languages	0.9458 Croatian 0.862 Slovenian 0.8614 Italian	0.0576 Cl. Chinese 0.1826 Kazakh 0.1829 Wolof	0.6299
Slovak	Indo-European languages, Slavic languages	0.9457 Czech 0.8912 Polish 0.8482 Italian	0.0618 Cl. Chinese 0.1962 Kazakh 0.21 Sanskrit	0.6412
Slovenian	Indo-European languages, Slavic languages	0.8772 Croatian 0.862 Serbian 0.8479 Italian	0.054 Cl. Chinese 0.1994 Wolof 0.2037 Arabic	0.6318
Spanish	Indo-European languages, Romance languages	0.9238 Catalan	0.0455 Cl. Chinese	0.6198

		0.847 Italian	0.173 Kazakh	
		0.8182 Romanian	0.1916 Sanskrit	
Swedish	Indo-European languages, Germanic languages	0.9208 Danish	0.0647 Cl. Chinese	0.6841
		0.8991 Dutch	0.1779 Wolof	
		0.8815 Estonian	0.215 Arabic	
Tamil	Dravidian languages, Dravidian languages	0.7979 Estonian	0.0679 Cl. Chinese	0.5932
		0.7787 Telugu	0.1832 Arabic	
		0.7487 Latvian	0.2149 Sanskrit	
Telugu	Dravidian languages, Dravidian languages	0.7787 Tamil	0.0714 Cl. Chinese	0.5702
		0.7691 Estonian	0.1696 Arabic	
		0.7548 Finnish	0.1937 Wolof	
Turkish	Turkic languages, Turkic languages	0.7305 Irish	0.0656 Cl. Chinese	0.547
		0.7053 Swedish	0.1607 Arabic	
		0.7045 Finnish	0.162 Wolof	
Ukrainian	Indo-European languages, Slavic languages	0.9387 Belarusian	0.0409 Cl. Chinese	0.6193
		0.901 Russian	0.1726 Wolof	
		0.8984 Latvian	0.1778 Sanskrit	
Urdu	Indo-European languages, Indo-Iranian languages	0.7134 Hindi	0.0568 Cl. Chinese	0.4873
		0.6504 Basque	0.1504 Arabic	
		0.6296 Slovenian	0.2068 Kazakh	
Uyghur	Turkic languages, Turkic languages	0.6243 Telugu	0.1005 Cl. Chinese	0.4485
		0.5712 Irish	0.1421 Arabic	
		0.5643 Turkish	0.1561 Wolof	
Vietnamese	Austroasiatic languages, Vietic languages	0.6718 Swedish	0.1387 Wolof	0.5233
		0.668 Danish	0.1471 Kazakh	
		0.6676 Lithuanian	0.1569 Arabic	
Welsh	Indo-European languages, Celtic languages	0.6284 Basque	0.0617 Cl. Chinese	0.4817
		0.621 Estonian	0.1476 Arabic	
		0.6209 Swedish	0.1614 Wolof	
Wolof	Niger-Congo languages, Senegambian languages	0.3377 Tamil	0.0507 Kazakh	0.2004
		0.3067 Hindi	0.0567 Arabic	
		0.2912 Chinese	0.0801 Cl. Chinese	

Table 6: Cosine similarity based on pattern

Based on the previous table, we can see that the languages pairs that achieve the highest similarity score are Croatian and Serbian, with a cosine similarity value of 0.9458 (in blue, in the table). The language that gets the highest cosine similarity most often, when comparing it with other languages, is Swedish (8 out of 58 times), followed by Italian (5 out of 58 times) and Latin (4 out of 58 times). The minimum cosine similarity value is 0.0156, found when comparing Classical Chinese to Arabic (marked in yellow in the table). In fact, 53 out of 58 times, Classical Chinese gets the worst cosine similarity value with the other languages. Arabic is the next worst cosine similarity value language with whom to pair, in 3 out of 58 cases. The total average similarity based on the average of every language is 0.5325.

4.4 Recurring patterns

After gathering the recurring patterns for each language, the 20 most frequent structures per language were saved and compared, so as to obtain the most-common language structures overall. Before comparing, all non-root elements were sorted alphabetically, ensuring that language-specific word order becomes irrelevant.

Table 7 contains the most common top-level structures among all languages, in descending order. The first column shows a language structure, the second column shows the number of languages that contain these structures, and the third column, the percentages of the languages they represent. For example, the structure of NOUN as a root with an ADJ as amod was found in 45 out of 58 languages, which corresponds to 77.59% of the analysed languages. For visualisation purposes, only the structures shared among at least 10 languages are shown in table 7. The full table is available here.

Structure	**Frequency**	**% of Lang**
NOUN()	57	98.28%
NOUN(ADJ-amod)	45	77.59%
VERB(NOUN-nsubj, NOUN-obl, PUNCT-punct)	45	77.59%
VERB(NOUN-nsubj, NOUN-obj, PUNCT-punct)	43	74.14%
NOUN(NOUN-nmod)	40	68.97%
VERB(NOUN-nsubj, NOUN-obj, NOUN-obl, PUNCT-punct)	37	63.79%
NOUN(NOUN-conj)	36	62.07%
VERB(NOUN-nsubj, NOUN-obl, NOUN-obl, PUNCT-punct)	28	48.28%
PROPN()	27	46.55%
VERB(NOUN-nsubj, NOUN-obj, PUNCT-punct, VERB-conj)	22	37.93%
VERB(NOUN-nsubj, PUNCT-punct, VERB-ccomp)	21	36.21%
VERB(NOUN-obj, PROPN-nsubj, PUNCT-punct)	20	34.48%
VERB(NOUN-nsubj, PUNCT-punct)	19	32.76%
PUNCT()	19	32.76%
VERB(NOUN-obj, NOUN-obl, PROPN-nsubj, PUNCT-punct)	18	31.03%
VERB(ADV-advmod, NOUN-nsubj, NOUN-obj, PUNCT-punct)	16	27.59%
VERB(ADV-advmod, NOUN-nsubj, NOUN-obl, PUNCT-punct)	15	25.86%
VERB(NOUN-obj, NOUN-obl, PUNCT-punct)	15	25.86%
VERB(NOUN-nsubj, PUNCT-punct, VERB-xcomp)	13	22.41%
VERB(AUX-aux, NOUN-nsubj, NOUN-obl, PUNCT-punct)	12	20.69%
VERB(NOUN-nsubj, NOUN-obl, PUNCT-punct, VERB-conj)	11	18.97%
VERB(AUX-aux:pass, NOUN-nsubj:pass, NOUN-obl, PUNCT-punct)	10	17.24%
VERB(AUX-aux, NOUN-nsubj, NOUN-obj, PUNCT-punct)	10	17.24%
NOUN(PUNCT-punct)	10	17.24%
NOUN(PROPN-nmod)	10	17.24%

Table 7: Recurring patterns

There are 381 different structures, based on the full table. The most common elements are verbs, in 217 out of 381 structures, nouns, in 72 structures, and adjectives, in 34 structures. Table 7 shows that the first two most-common structures have a noun as its root: they consist of a noun, or a noun modified by an adjective. The first is shared among the vast majority of the languages, and the second one, in 45 out of the 58 languages. The next two most common structures have a verb as a root. They are shared in more than 74% of the languages and consist of sentence with an oblique (in English, subject + verb + oblique) and a prototypical transitive sentence (subject + verb + object). The next example illustrates the four most-common constructions with sentences in English, Spanish and Finnish taken from the topic “Earth”. Finnish is chosen as one of the example languages because its grammar is exceptionally different to the English grammar.

- **NOUN()**
  - **Eng** Etymology
  - **Spa** Cronología (‘Cronology’)
  - **Fin** Rakenne (‘Structure’)
- **NOUN(ADJ-amod)**
  - **Eng** Geological history
  - **Spa** Composición química (‘Chemical composition’)
  - **Fin** Transneptuniset kohteet (‘Transneptune targets’)
- **VERB(NOUN-nsubj, NOUN-obl, PUNCT-punct)**
  - **Eng** The amount of solar energy that reaches the Earth’s surface decreases with increasing latitude.
  - **Spa** En la década de 1960 surgió una hipótesis que afirmaba que durante el período Neoproterozoico, desde 750 hasta los 580 Ma, se produjo una intensa glaciación en la que gran parte del planeta fue cubierto por una capa de hielo.
    (‘In the 60s, a hypothesis emerged that stated that during the Neoproterozoic period, from 750 to 580 Ma, there was an intense glaciation where much of the planet was covered by an ice cap.’)
  - **Fin** Ydin koostuu pääosin raudasta ja nikkelistä. (‘The core consists mainly of iron and nickel.’)
- **VERB(NOUN-nsubj, NOUN-obj, PUNCT-punct)**
  - **Eng** The most abundant silicate minerals on Earth’s surface include quartz, feldspars, amphibole, mica, pyroxene and olivine.
  - **Spa** La atracción gravitatoria entre la Tierra y la Luna causa las mareas en la Tierra.
    (‘The gravitational attraction between the Earth and the Moon causes the tides on Earth’)
  - **Fin** Maassa esiintyy runsaasti elämää. (‘There is a lot of life on Earth.’)

5 Discussion

5.1 Cosine similarity

The results of the DEPREL-based cosine similarity are quite encouraging. Looking at the differences among languages in the same topic, the average DEPREL-based cosine similarity is 0.7213 and, among the same language, 0.9429. These results seem to point out that there are syntactic similarities not only intralingually, which is expected, but also interlingually.

Within the same topic, Catalan and Spanish get the highest DEPREL-based cosine similarity 43 out of 62 times. This could indicate that one of the articles has been created via translation, perhaps using the APERTIUM software (Forcada et al., 2011). However, after looking at some articles in which the cosine similarity has a high value between Catalan and Spanish, this theory is discarded. The content of both articles is quite similar because they are concerned with the same idea, but not identical, thus they have not been very likely created through translation.

Catalan and Spanish are two languages that are very closely related in their computational advances. In fact, there are many resources for both languages that have been created simultaneously. This is the case of the AnCora Treebank (Taulé et al., 2008), which is the treebank used to train the Catalan (AnCora-Ca) and Spanish (AnCora-Es) parser. Even though it has been created together, one of them is not a translation of the other. It is, therefore, possible that the high DEPREL-based cosine similarity achieves a high value because the treebanks used for the parser were based on the same annotation rules.

There are two language combinations within topics that achieve rather low results: comparing Gothic and Hungarian has achieved the lowest cosine similarity per topic in 23 cases, and comparing Japanese and Sanskrit, 18. One possible explanation for their low results is their disparity in origin: Gothic is a Germanic (Indo-European family) language, while Hungarian is a Finno-Ugric language (Uralic family), and Japanese is a Japonic language, while Sanskrit is an Indo-Iranian language (Indo-European). Nonetheless, there are other language combinations within topics that achieve a high result in spite of coming from different languages. For instance, Belarussian (a Slavic language from the Indo-European family) and Arabic (a Semitic language, from the Afroasiatic family) get a high result in the topic of “Africa” (0.8455).

Another possibility of their systematic differences is the length of the articles analysed. The Gothic article about “Money” contains 5 sentences, and the Hungarian article on the same topic, 265. Then again, Gothic is an extinct language that has not been spoken for many years. Sanskrit follows the same pattern: it is a language that is not currently spoken and contains a very low number of sentences per article. The low number of sentences should not immediately yield a bad result, because the cosine similarity measure takes size into account. However, articles with a few sentences tend to have different structures, for instance only noun phrases. It is possible that the effects of the different size of the articles in addition to the different origins has caused these languages to yield low results. If we look at the average number of sentences per article in figure 2, we can see that Gothic and Sanskrit are some of the languages with the lowest number of sentences per article of Wikipedia.

If we look at the use of dependencies used in the analysis of the different languages (available in the GitHub repository) we can see that there are some dependencies which are rather language-specific, which could cause a low similarity measure. For instance, *flat:vv*, used in serial verbs in Classical Chinese (de Marneffe et al., 2021b), or *discourse:sp*, sentence-final particles in Chinese and Classical Chinese (de Marneffe et al., 2021a). However, the most interesting dependency is the one we cannot find: *punct*. There are no punctuation dependencies in Classical Chinese, Gothic, Old Church Slavonic and Sanskrit (and no punctuation POS tags). This is probably one of the reasons for the low similarity values found when comparing the other languages to these.

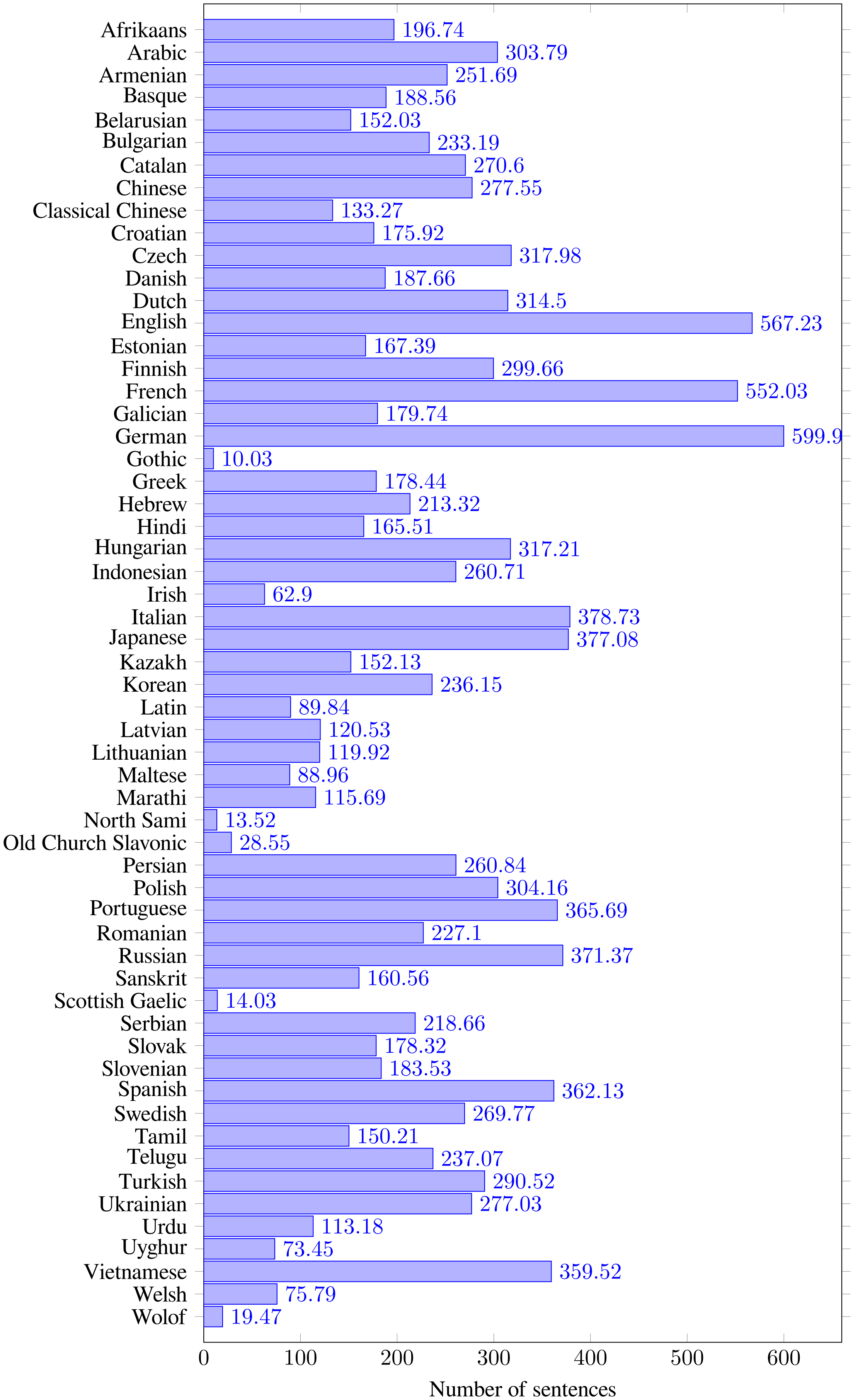


Figure 2: Average number of sentences per Wikipedia article

The DEPREL-based cosine similarity is higher when comparing different articles in the same language. This result was expected, given that the articles usually have the same length in one language, and given that the parser, UDPipe 2.0, uses the same model to analyse both articles. In addition, the articles with a higher cosine similarity value seem to be semantically related, and the ones with the lowest do not. The topics that achieved the highest cosine similarity value when compared were two celestial bodies (e.g. "Moon" and "Sun"), metals (e.g. "Gold" and "Silver"), religions (e.g. "Christianity" and "Judaism"), people (e.g. "Barack Obama" and "George W. Bush"), but mostly, places (e.g. "Italy" and "United Kingdom"). Perhaps the most striking is the high cosine similarity value in Catalan between the topics "Eye" and "Milk". Nonetheless, the higher similarity values in the semantically related topics is encouraging for the Abstract Wikipedia project, because it does reflect how similar topics are expressed similarly within a language.

The pattern-based cosine similarity is generally lower than the DEPREL-based cosine similarity: its average is 0.5325 among all languages. This is based in the comparison of all the sentences in one language, to all the sentences in another language. The patterns contain the root and the elements that directly depend on it, stating the POS tag and their dependency relation. It therefore is not surprising that the pattern-based cosine similarity is lower than the DEPREL-based one, because there is more information added: the POS tag of each element. In total, there are 241 998 patterns among all languages.

The language that gets the worst cosine similarity value when pairing it with others is Classical Chinese (91.4% of the times). Even though it does not have as many sentences as other languages, it does not have a particularly low number of sentences (average of 133 sentences per topic). It is from the same family as Chinese, but the latter does not yield such low results. Therefore, the low cosine similarity must come from elsewhere. We believe that the lack of punctuation in the language is the cause of this; the majority of the patterns found, 89% of them, do contain punctuation.

The three languages that are compared most often when achieving high pattern-based cosine similarity are Swedish, Italian and Latin. A possible reason for this phenomenon is the overrepresentation of Indo-European languages in the data: 38 languages out of 58 are from this family.

5.2 Recurring patterns

There is an extremely large number of top-level patterns, close to a quarter of a million. In order to get the most representative of each language, only the 20 most-common patterns per language were saved and compared. That leaves 381 different structures, 25 of which are shared among 10 languages or more (table 7).

There are 72 structures out of 381 which have a noun as their root. From the ones at the table 7, we know that these can be isolated nouns, nouns with adjectives or another noun working as a nominal modifier, two nouns, proper nouns, nouns with a punctuation mark or nouns modified by a proper noun. These structures are typical of titles, rather than the content of articles. Separating the original data into titles and content could have been a way to improve the analysis of the actual semantic content of the articles and allow for a more accurate separate analysis.

Curiously, the structure of nouns with a determiner are not represented in at least 10 languages. This is probably due to the nature of the patterns: we are representing only the elements that directly depend on the root. If we analysed subtrees too, determiners would probably be more represented.

The presence of structures consisting only of a punctuation mark in almost half of the languages suggests that the data has not been gathered perfectly or that the parser has not tokenised sentences correctly. For instance, when looking at the data in English, we can find some formulas within the text that have been separated during the parsing, as well as some punctuation tokens that appear in the raw text but do not seem visible in the articles of Wikipedia, such as "⁂".

The are many structures with verbs as roots (217), a lot of which have nouns (or proper nouns) as subjects, nouns as obliques or objects, or both, and a punctuation mark. They also can be coordinated or have subordinated clauses. An abundance of verbs as roots is expected, because most of the text in a Wikipedia article are sentences, and most sentences tend to have verbs as their root.

6 Conclusion

In this thesis, we have computationally analysed the language used in Wikipedia, from a multilingual perspective. First, we have presented a new syntactically analysed dataset based on Wikipedia articles. The articles have been parsed using the pre-trained models of UDPipe 2.0, which is based on Universal Dependencies. We have evaluated the parser by creating a gold standard in three languages: English, Spanish, and French. Moreover, we have created a new measure of evaluation for the parser made to improve the visualisation of missalignments, with which we have learnt that UDPipe 2.0 achieves a high score when parsing the analysed languages.

Then, we have gathered the syntactic patterns of every language and their distribution using GF-UD, a powerful framework that supports the interlingual perspective, and compared them among each other. The distribution of syntactic patterns per language can be a good foundation for the Abstract Wikipedia project, whose goal is to make more knowledge available in more languages. They want to do so by creating abstract representations of the content which can be rendered into different languages on request. The patterns found of each language can make up the Renderers of the language, perhaps using Grammatical Framework.

Finally, we have compared the syntactic analyses using cosine similarity, first based on their dependencies and then on their syntactic patterns. In doing so, we have found that the articles do have some similarities, not only within the same language, which is to be expected, but also among different languages. Furthermore, we have seen that semantically related articles tend to be more similar that those which are not. These results can be taken as encouraging for the Abstract Wikipedia project, because they support the theory that languages express the same concepts in a similar manner.

This is, to our knowledge, the first computational analysis of the language used in Wikipedia based on dependency relations. It is, in addition, one of the first ones to use dependency relations as its base of analysis in such a big scale. We believe that the dependency relations can represent the linguistic characteristics of a language, especially using Universal Dependencies, and be the base for future cross-linguistic analyses.

6.1 Critique and Future Work

The work presented heavily relies on the analysis done by UDPipe 2.0. Even though there are many pre-trained models available for a variety of languages, it is necessary that this (or other parsers) are further developed to work with multiple languages. Especially languages that are not from the Indo-European family, which are over-represented in this thesis.

Some of the analysed languages are extinct or used in very specific contexts, like Gothic or Sanskrit. Their low resources and, in some cases, their distinctive characteristics (like the lack of punctuation) has made them obtain low results. Since Abstract Wikipedia is interested in sharing knowledge so people can read it, it is probably not a priority to develop resources for languages with an extremely low or non-existent number of speakers. For this reason, we think that future work should prioritise the development of living languages before focusing on these ones.

Separating the titles from the content of the articles could have been a better resource from which to gather the linguistic patterns. This would have allowed to confirm that noun phrases are as common within the text as the results show, and see what structures they reflect more often. There are, in addition, some mistakes in the gathering of the data, such as the isolated punctuation marks that show no linguistic purpose.

Overall, the most important work to follow is the development of rules and functions in Grammatical Framework that can convert the frequent patterns of each language into concrete syntaxes. Doing so could be part of the solution for the Abstract Wikipedia project to answer the challenge posed by Wikipedia.

6.2 Ethical considerations

The environmental impact of this thesis is almost negligible because we did not train any machine learning model. The most computationally-intensive part has been the parsing of the texts using UDPipe 2.0, which has was made lower by using the pre-trained models available. We are aware of the possible biases that the articles of Wikipedia may carry, such as the gender inequalities of the editors of Wikipedia. However, these are more likely to be reflected on the choice of articles and their development, rather than the language used. We do not consider this thesis to have further ethical implications.

References

Ager, S. (2022). Omniglot. https://omniglot.com/writing/langfam.htm. Retrieved 2nd of May 2022.

Antin, J., Yee, R., Cheshire, C., & Nov, O. (2011). Gender differences in Wikipedia editing. *WikiSym 2011 Conference Proceedings - 7th Annual International Symposium on Wikis and Open Collaboration*, (pp. 11–14).

Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M., & Gergle, D. (2012). Omnipedia: Bridging the Wikipedia Language Gap. (pp. 1075–1084).

Crystal, D. (2008). *A dictionary of linguistics and phonetics.*

de Marneffe, M., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., McDonald, R., Nivre, J., Petrov, S., Pyysalo, S., Schuster, S., Silveira, N., Tsarfaty, R., Tyers, F., & Zeman, D. (2021a). UD discourse:sp. https://universaldependencies.org/lzh/dep/discourse-sp.html. Retrieved 15 May 2022.

de Marneffe, M., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., McDonald, R., Nivre, J., Petrov, S., Pyysalo, S., Schuster, S., Silveira, N., Tsarfaty, R., Tyers, F., & Zeman, D. (2021b). UD flat:vv. https://universaldependencies.org/lzh/dep/flat-vv.html. Retrieved 15 May 2022.

de Marneffe, M.-C., Guillaume, B., Grioni, M., Dickerson, C., & Perrier, G. (2021c). UD French GSD. https://universaldependencies.org/treebanks/fr_gsd/. Retrieved 25 April 2022.

de Marneffe, M. C., Manning, C. D., Nivre, J., & Zeman, D. (2021d). Universal dependencies. *Computational Linguistics*, 47(2), 255–308.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1 (pp. 4171–4186).

Erkan, G., Özgür, A., & Radev, D. R. (2007). Semi-supervised classification for extracting protein interaction sentences using dependency parsing. *EMNLP-CoNLL 2007 - Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, (June), 228–237.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., & Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2), 127–144.

Harris, Z. S. (1954). Distributional Structure. *<i>WORD</i>*, 10(2-3), 146–162.

Hill, B. M. & Shaw, A. (2013). The Wikipedia Gender Gap Revisited: Characterizing Survey Response Bias with Propensity Score Estimation. *PLoS ONE*, 8(6), 1–5.

Inan, E. (2020). SimiT: A Text Similarity Method Using Lexicon and Dependency Representations. *New Generation Computing*, 38(3), 509–530.

Joo, S. (2020). Exploring the domain of information "users": Semantic analysis of wikipedia articles. *Journal of Library and Information Studies*, 18(1), 1–23.

Kolachina, P. & Ranta, A. (2016). From Abstract Syntax to Universal Dependencies. *Literature in Language Teaching*, 13(3), 1–57.

Levy, O. & Goldberg, Y. (2014). Dependency-Based Word Embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 302–308). Baltimore, Maryland, USA.

Lindat & CLARIAH-CZ (2022). UDPipe API. `http://lindat.mff.cuni.cz/services/udpipe/`. Retrieved 25 April 2022.

Liu, K. & El-Gohary, N. (2017). Similarity-Based Dependency Parsing for Extracting Dependency Relations from Bridge Inspection Reports. *American Society of Civil Engineers*, 5.

Massa, P. & Scrinzi, F. (2012). Manypedia: Comparing language points of view of Wikipedia communities. *WikiSym 2012 Conference Proceedings - 8th Annual International Symposium on Wikis and Open Collaboration*.

Nivre, J., De Marneffe, M. C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., & Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016* (pp. 1659–1666).

Ortega, F., Gonzalez-Barahona, J. M., & Robles, G. (2008). On the inequality of contributions to wikipedia. In *Annual Hawaii International Conference on System Sciences* (pp. 1–7).

Ortega Soto, J. F. (2009). *Wikipedia: A quantitative analysis*. PhD thesis, Universidad Rey Juan Carlos.

Radev, D. R., Özgür, A., & Özateş, �. B. (2008). Sentence Similarity based on Dependency Tree Kernels for Multi-document Summarization. (pp. 2833–2838).

Ranta, A. (2020). *Computational Grammar. An Interlingual Perspective*.

Ranta, A., Angelov, K., Gruzitis, N., & Kolachina, P. (2020). Abstract syntax as interlingua: Scaling up the grammatical framework from controlled languages to robust pipelines. *Computational Linguistics*, 46(2), 425–486.

Ranta, A. & Kolachina, P. (2017). From Universal Dependencies to Abstract Syntax. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies, UDW 2017*, number May (pp. 107–116).

Ranta, A., Masciolini, A., Källberg, A., Listenmaa, I., & Lange, H. (2022). GF-UD. `https://github.com/GrammaticalFramework/gf-ud`. Retrieved 25 April 2022.

Richardson, L. (2007). Beautiful soup documentation. *April*.

Samoilenko, A. & Yasseri, T. (2014). The distorted mirror of wikipedia: A quantitative analysis of wikipedia coverage of academics. *EPJ Data Science*, 3(1), 1–11.

Silveira, N., Dozat, T., Schuster, S., Connor, M., de Marneffe, M.-C., Schneider, N., Chi, E., Bowman, S., Manning, C., Zhu, H., Galbraith, D., & Bauer, J. (2021). UD English EWT. `https://universaldependencies.org/treebanks/en_ewt/`. Retrieved 25 April 2022.

Straka, M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In *CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 197–207).

Straka, M. & Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *CoNLL 2017 - SIGNLL Conference on Computational Natural Language Learning, Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, volume 2 (pp. 88–99).

Taulé, M., Martí, M. A., & Recasens, M. (2008). AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)* Marrakech, Morocco: European Language Resources Association (ELRA).

Tran, Q. H., Tran, V. D., Vu, T. T., Le Nguyen, M., & Pham, S. B. (2015). JAIST: Combining multiple features for Answer Selection in Community Question Answering. *SemEval 2015 - 9th International Workshop on Semantic Evaluation, co-located with the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2015 - Proceedings*, (SemEval), 215–219.

Vrandečić, D. (2018a). Capturing meaning: Toward an abstract Wikipedia. *CEUR Workshop Proceedings*, 2018.

Vrandečić, D. (2018b). Toward an abstract Wikipedia. In *CEUR Workshop Proceedings*, volume 2211.

Vrandečić, D. (2021). *Abstract Wikipedia*, 7th Grammatical Framework Summer School. `https://youtu.be/if5TeJ8N2p8`.

Wikimedia (2022a). Gender bias on Wikipedia. `https://en.wikipedia.org/wiki/Gender_bias_on_Wikipedia`. Retrieved 8 June 2022.

Wikimedia (2022b). List of most visited websites. `https://en.wikipedia.org/wiki/List_of_most_visited_websites`. Retrieved 8 March 2022.

Wikimedia (2022c). Wikipedia: About. `https://en.wikipedia.org/wiki/Wikipedia:About`. Retrieved 8 March 2022.

Wikimedia (2022d). Wikipedia articles written in the greatest number of languages. `https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_articles_written_in_the_greatest_number_of_languages`. Retrieved 8 March 2022.

Yasseri, T., Kornai, A., & Kertész, J. (2012). A Practical Approach to Language Complexity: A Wikipedia Case Study. *PloS ONE*, 7(11).

Appendices

A Wikipedia Topics

- Adolf Hitler
- Africa
- Asia
- Association football
- Barack Obama
- Bible
- Buddha
- Buddhism
- China
- Christianity
- Christmas
- Dog
- Earth
- English Language
- Europe
- Eye
- George W. Bush
- Ghana
- Gold
- Hinduism
- Human
- India
- Internet
- Iran
- Iraq
- Iron
- Islam
- Italy
- Japan
- Jesus
- Judaism
- Julius Caesar
- Koran
- Maize
- Milk
- Mohandas Karamchand Gandhi
- Money
- Moon
- Moses
- Muhammad
- New York City
- Niger
- Osama Bin Laden
- Paris
- Periodic table
- Pope Benedict XVI
- Pope John Paul II
- Religion
- Rice
- Roman Catholic Church
- Rome
- Russia
- Silver
- South Africa
- South America
- Soviet Union
- Sun
- United Kingdom
- United States
- Water
- Wikipedia
- World War II

B UDPipe Models

This section shows all the available pre-trained models in UDPipe, separated by language. The models used in this thesis have been marked in italics.

1. **Afrikaans**
   - *afrikaans-afribooms-ud-2.6-200830*
   - afrikaans-afribooms-ud-2.5-191206
   - afrikaans-afribooms-ud-2.4-190531
2. **Ancient Greek**
   - ancient_greek-perseus-ud-2.6-200830
   - ancient_greek-proiel-ud-2.6-200830
   - ancient_greek-perseus-ud-2.5-191206
   - ancient_greek-proiel-ud-2.5-191206
   - ancient_greek-perseus-ud-2.4-190531
   - ancient_greek-proiel-ud-2.4-190531
   - ancient_greek-ud-2.0-170801
   - ancient_greek-proiel-ud-2.0-170801
   - ancient-greek-ud-1.2-160523
   - ancient-greek-proiel-ud-1.2-160523
3. **Arabic**
   - *arabic-padt-ud-2.6-200830*
   - arabic-padt-ud-2.5-191206
   - arabic-padt-ud-2.4-190531
   - arabic-ud-2.0-170801
   - arabic-ud-1.2-160523
4. **Armenian**
   - *armenian-armtdp-ud-2.6-200830*
   - armenian-armtdp-ud-2.5-191206
   - armenian-armtdp-ud-2.4-190531
5. **Basque**
   - *basque-bdt-ud-2.6-200830*
   - basque-bdt-ud-2.5-191206
   - basque-bdt-ud-2.4-190531
   - basque-ud-2.0-170801
   - basque-ud-1.2-160523
6. **Belarusian**
   - *belarusian-hse-ud-2.6-200830*
   - belarusian-hse-ud-2.5-191206
   - belarusian-hse-ud-2.4-190531
   - belarusian-ud-2.0-170801
7. **Bulgarian**
   - *bulgarian-btb-ud-2.6-200830*
   - bulgarian-btb-ud-2.5-191206
   - bulgarian-btb-ud-2.4-190531
   - bulgarian-ud-2.0-170801
   - bulgarian-ud-1.2-160523
8. **Catalan**
   - *catalan-ancora-ud-2.6-200830*
   - catalan-ancora-ud-2.5-191206
   - catalan-ancora-ud-2.4-190531
   - catalan-ud-2.0-170801
9. **Chinese**
   - *chinese-gsdsimp-ud-2.6-200830*
   - chinese-gsd-ud-2.6-200830
   - chinese-gsdsimp-ud-2.5-191206
   - chinese-gsd-ud-2.5-191206
   - chinese-gsd-ud-2.4-190531
   - chinese-ud-2.0-170801
10. **Classical Chinese**
    - *classical_chinese-kyoto-ud-2.6-200830*
    - classical_chinese-kyoto-ud-2.5-191206
    - classical_chinese-kyoto-ud-2.4-190531
11. **Coptic**
    - coptic-scriptorium-ud-2.6-200830
    - coptic-scriptorium-ud-2.5-191206
    - coptic-scriptorium-ud-2.4-190531
    - coptic-ud-2.0-170801
12. **Croatian**
    - *croatian-set-ud-2.6-200830*
    - croatian-set-ud-2.5-191206
    - croatian-set-ud-2.4-190531
    - croatian-ud-2.0-170801

- croatian-ud-1.2-160523

13. **Czech**
    - *czech-pdt-ud-2.6-200830*
    - czech-cac-ud-2.6-200830
    - czech-fictree-ud-2.6-200830
    - czech-cltt-ud-2.6-200830
    - czech-pdt-ud-2.5-191206
    - czech-cac-ud-2.5-191206
    - czech-fictree-ud-2.5-191206
    - czech-cltt-ud-2.5-191206
    - czech-pdt-ud-2.4-190531
    - czech-cac-ud-2.4-190531
    - czech-fictree-ud-2.4-190531
    - czech-cltt-ud-2.4-190531
    - czech-ud-2.0-170801
    - czech-cac-ud-2.0-170801
    - czech-cltt-ud-2.0-170801
    - czech-ud-1.2-160523
14. **Danish**
    - *danish-ddt-ud-2.6-200830*
    - danish-ddt-ud-2.5-191206
    - danish-ddt-ud-2.4-190531
    - danish-ud-2.0-170801
    - danish-ud-1.2-160523
15. **Dutch**
    - *dutch-alpino-ud-2.6-200830*
    - dutch-lassysmall-ud-2.6-200830
    - dutch-alpino-ud-2.5-191206
    - dutch-lassysmall-ud-2.5-191206
    - dutch-alpino-ud-2.4-190531
    - dutch-lassysmall-ud-2.4-190531
    - dutch-ud-2.0-170801
    - dutch-lassysmall-ud-2.0-170801
    - dutch-ud-1.2-160523
16. **English**
    - *english-ewt-ud-2.6-200830*
    - english-gum-ud-2.6-200830
    - english-lines-ud-2.6-200830
    - english-partut-ud-2.6-200830
    - english-ewt-ud-2.5-191206
    - english-gum-ud-2.5-191206
    - english-lines-ud-2.5-191206
    - english-partut-ud-2.5-191206
    - english-ewt-ud-2.4-190531
    - english-gum-ud-2.4-190531
    - english-lines-ud-2.4-190531
    - english-partut-ud-2.4-190531
    - english-ud-2.0-170801
    - english-lines-ud-2.0-170801
    - english-partut-ud-2.0-170801
    - english-ud-1.2-160523
17. **Estonian**
    - *estonian-edt-ud-2.6-200830*
    - estonian-ewt-ud-2.6-200830
    - estonian-edt-ud-2.5-191206
    - estonian-ewt-ud-2.5-191206
    - estonian-edt-ud-2.4-190531
    - estonian-ewt-ud-2.4-190531
    - estonian-ud-2.0-170801
    - estonian-ud-1.2-160523
18. **Finnish**
    - *finnish-tdt-ud-2.6-200830*
    - finnish-ftb-ud-2.6-200830
    - finnish-tdt-ud-2.5-191206
    - finnish-ftb-ud-2.5-191206
    - finnish-tdt-ud-2.4-190531
    - finnish-ftb-ud-2.4-190531
    - finnish-ud-2.0-170801
    - finnish-ftb-ud-2.0-170801
    - finnish-ud-1.2-160523
    - finnish-ftb-ud-1.2-160523
19. **French**
    - *french-gsd-ud-2.6-200830*
    - french-sequoia-ud-2.6-200830
    - french-partut-ud-2.6-200830
    - french-spoken-ud-2.6-200830
    - french-gsd-ud-2.5-191206
    - french-sequoia-ud-2.5-191206
    - french-partut-ud-2.5-191206

- french-spoken-ud-2.5-191206
- french-gsd-ud-2.4-190531
- french-sequoia-ud-2.4-190531
- french-partut-ud-2.4-190531
- french-spoken-ud-2.4-190531
- french-ud-2.0-170801
- french-partut-ud-2.0-170801
- french-sequoia-ud-2.0-170801
- french-ud-1.2-160523

20. **Galician**
    - *galician-ctg-ud-2.6-200830*
    - galician-treegal-ud-2.6-200830
    - galician-ctg-ud-2.5-191206
    - galician-treegal-ud-2.5-191206
    - galician-ctg-ud-2.4-190531
    - galician-treegal-ud-2.4-190531
    - galician-ud-2.0-170801
    - galician-treegal-ud-2.0-170801

21. **German**
    - *german-hdt-ud-2.6-200830*
    - german-gsd-ud-2.6-200830
    - german-hdt-ud-2.5-191206
    - german-gsd-ud-2.5-191206
    - german-gsd-ud-2.4-190531
    - german-ud-2.0-170801
    - german-ud-1.2-160523

22. **Gothic**
    - *gothic-proiel-ud-2.6-200830*
    - gothic-proiel-ud-2.5-191206
    - gothic-proiel-ud-2.4-190531
    - gothic-ud-2.0-170801
    - gothic-ud-1.2-160523

23. **Greek**
    - *greek-gdt-ud-2.6-200830*
    - greek-gdt-ud-2.5-191206
    - greek-gdt-ud-2.4-190531
    - greek-ud-2.0-170801
    - greek-ud-1.2-160523

24. **Hebrew**
    - *hebrew-htb-ud-2.6-200830*
    - hebrew-htb-ud-2.5-191206
    - hebrew-htb-ud-2.4-190531
    - hebrew-ud-2.0-170801
    - hebrew-ud-1.2-160523

25. **Hindi**
    - *hindi-hdtb-ud-2.6-200830*
    - hindi-hdtb-ud-2.5-191206
    - hindi-hdtb-ud-2.4-190531
    - hindi-ud-2.0-170801
    - hindi-ud-1.2-160523

26. **Hungarian**
    - *hungarian-szeged-ud-2.6-200830*
    - hungarian-szeged-ud-2.5-191206
    - hungarian-szeged-ud-2.4-190531
    - hungarian-ud-2.0-170801
    - hungarian-ud-1.2-160523

27. **Indonesian**
    - *indonesian-gsd-ud-2.6-200830*
    - indonesian-gsd-ud-2.5-191206
    - indonesian-gsd-ud-2.4-190531
    - indonesian-ud-2.0-170801
    - indonesian-ud-1.2-160523

28. **Irish**
    - *irish-idt-ud-2.6-200830*
    - irish-idt-ud-2.5-191206
    - irish-idt-ud-2.4-190531
    - irish-ud-2.0-170801
    - irish-ud-1.2-160523

29. **Italian**
    - *italian-isdt-ud-2.6-200830*
    - italian-partut-ud-2.6-200830
    - italian-postwita-ud-2.6-200830
    - italian-twittiro-ud-2.6-200830
    - italian-vit-ud-2.6-200830
    - italian-isdt-ud-2.5-191206
    - italian-partut-ud-2.5-191206
    - italian-postwita-ud-2.5-191206

- italian-twittiro-ud-2.5-191206
- italian-vit-ud-2.5-191206
- italian-isdt-ud-2.4-190531
- italian-partut-ud-2.4-190531
- italian-postwita-ud-2.4-190531
- italian-vit-ud-2.4-190531
- italian-ud-2.0-170801
- italian-ud-1.2-160523

30. **Japanese**
    - *japanese-gsd-ud-2.6-200830*
    - japanese-gsd-ud-2.5-191206
    - japanese-gsd-ud-2.4-190531
    - japanese-ud-2.0-170801

31. **Kazakh**
    - *kazakh-ud-2.0-170801*

32. **Korean**
    - *korean-kaist-ud-2.6-200830*
    - korean-gsd-ud-2.6-200830
    - korean-kaist-ud-2.5-191206
    - korean-gsd-ud-2.5-191206
    - korean-kaist-ud-2.4-190531
    - korean-gsd-ud-2.4-190531
    - korean-ud-2.0-170801

33. **Latin**
    - *latin-ittb-ud-2.6-200830*
    - latin-llct-ud-2.6-200830
    - latin-proiel-ud-2.6-200830
    - latin-perseus-ud-2.6-200830
    - latin-evalatin20-200830
    - latin-ittb-ud-2.5-191206
    - latin-proiel-ud-2.5-191206
    - latin-perseus-ud-2.5-191206
    - latin-ittb-ud-2.4-190531
    - latin-proiel-ud-2.4-190531
    - latin-perseus-ud-2.4-190531
    - latin-ud-2.0-170801
    - latin-ittb-ud-2.0-170801
    - latin-proiel-ud-2.0-170801
    - latin-ud-1.2-160523
    - latin-itt-ud-1.2-160523
    - latin-proiel-ud-1.2-160523

34. **Latvian**
    - *latvian-lvtb-ud-2.6-200830*
    - latvian-lvtb-ud-2.5-191206
    - latvian-lvtb-ud-2.4-190531
    - latvian-ud-2.0-170801

35. **Lithuanian**
    - *lithuanian-alksnis-ud-2.6-200830*
    - lithuanian-hse-ud-2.6-200830
    - lithuanian-alksnis-ud-2.5-191206
    - lithuanian-hse-ud-2.5-191206
    - lithuanian-alksnis-ud-2.4-190531
    - lithuanian-hse-ud-2.4-190531
    - lithuanian-ud-2.0-170801

36. **Maltese**
    - *maltese-mudt-ud-2.6-200830*
    - maltese-mudt-ud-2.5-191206
    - maltese-mudt-ud-2.4-190531

37. **Marathi**
    - *marathi-ufal-ud-2.6-200830*
    - marathi-ufal-ud-2.5-191206
    - marathi-ufal-ud-2.4-190531

38. **Naija**
    - naija-nsc-ud-2.6-200830

39. **North Sami**
    - *north_sami-giella-ud-2.6-200830*
    - north_sami-giella-ud-2.5-191206
    - north_sami-giella-ud-2.4-190531

40. **Norwegian Bokmaal**
    - norwegian-bokmaal-ud-2.6-200830
    - norwegian-bokmaal-ud-2.5-191206
    - norwegian-bokmaal-ud-2.4-190531
    - norwegian-bokmaal-ud-2.0-170801

41. **Norwegian Nynorsk**
    - norwegian-nynorsk-ud-2.6-200830
    - norwegian-nynorsk-ud-2.5-191206

- norwegian-nynorsk-ud-2.4-190531
- norwegian-nynorsk-ud-2.0-170801
- norwegian-ud-1.2-160523

42. **Norwegian Nynorsklia**
    - norwegian-nynorsklia-ud-2.6-200830
    - norwegian-nynorsklia-ud-2.5-191206
    - norwegian-nynorsklia-ud-2.4-190531

43. **Old Church Slavonic**
    - *old_church_slavonic-proiel-ud-2.6-200830*
    - old_church_slavonic-proiel-ud-2.5-191206
    - old_church_slavonic-proiel-ud-2.4-190531
    - old_church_slavonic-ud-2.0-170801
    - old-church-slavonic-ud-1.2-160523

44. **Old French**
    - old_french-srcmf-ud-2.6-200830
    - old_french-srcmf-ud-2.5-191206
    - old_french-srcmf-ud-2.4-190531

45. **Old Russian**
    - old_russian-torot-ud-2.6-200830
    - old_russian-rnc-ud-2.6-200830
    - old_russian-torot-ud-2.5-191206
    - old_russian-torot-ud-2.4-190531

46. **Persian**
    - *persian-seraji-ud-2.6-200830*
    - persian-seraji-ud-2.5-191206
    - persian-seraji-ud-2.4-190531
    - persian-ud-2.0-170801
    - persian-ud-1.2-160523

47. **Polish**
    - *polish-pdb-ud-2.6-200830*
    - polish-lfg-ud-2.6-200830
    - polish-pdb-ud-2.5-191206
    - polish-lfg-ud-2.5-191206
    - polish-pdb-ud-2.4-190531
    - polish-lfg-ud-2.4-190531
    - polish-ud-2.0-170801
    - polish-ud-1.2-160523

48. **Portuguese**
    - *portuguese-gsd-ud-2.6-200830*
    - portuguese-bosque-ud-2.6-200830
    - portuguese-gsd-ud-2.5-191206
    - portuguese-bosque-ud-2.5-191206
    - portuguese-gsd-ud-2.4-190531
    - portuguese-bosque-ud-2.4-190531
    - portuguese-ud-2.0-170801
    - portuguese-br-ud-2.0-170801
    - portuguese-ud-1.2-160523

49. **Romanian**
    - *romanian-rrt-ud-2.6-200830*
    - romanian-nonstandard-ud-2.6-200830
    - romanian-rrt-ud-2.5-191206
    - romanian-nonstandard-ud-2.5-191206
    - romanian-rrt-ud-2.4-190531
    - romanian-nonstandard-ud-2.4-190531
    - romanian-ud-2.0-170801
    - romanian-ud-1.2-160523

50. **Russian**
    - *russian-syntagrus-ud-2.6-200830*
    - russian-gsd-ud-2.6-200830
    - russian-taiga-ud-2.6-200830
    - russian-syntagrus-ud-2.5-191206
    - russian-gsd-ud-2.5-191206
    - russian-taiga-ud-2.5-191206
    - russian-syntagrus-ud-2.4-190531
    - russian-gsd-ud-2.4-190531
    - russian-taiga-ud-2.4-190531
    - russian-ud-2.0-170801
    - russian-syntagrus-ud-2.0-170801

51. **Sanskrit**
    - *sanskrit-vedic-ud-2.6-200830*
    - sanskrit-ud-2.0-170801

52. **Scottish Gaelic**
    - *scottish_gaelic-arcosg-ud-2.6-200830*
    - scottish_gaelic-arcosg-ud-2.5-191206

53. **Serbian**
    - *serbian-set-ud-2.6-200830*
    - serbian-set-ud-2.5-191206
    - serbian-set-ud-2.4-190531
54. **Slovak**
    - *slovak-snk-ud-2.6-200830*
    - slovak-snk-ud-2.5-191206
    - slovak-snk-ud-2.4-190531
    - slovak-ud-2.0-170801
55. **Slovenian**
    - *slovenian-ssj-ud-2.6-200830*
    - slovenian-sst-ud-2.6-200830
    - slovenian-ssj-ud-2.5-191206
    - slovenian-sst-ud-2.5-191206
    - slovenian-ssj-ud-2.4-190531
    - slovenian-sst-ud-2.4-190531
    - slovenian-ud-2.0-170801
    - slovenian-sst-ud-2.0-170801
    - slovenian-ud-1.2-160523
56. **Spanish**
    - *spanish-ancora-ud-2.6-200830*
    - spanish-gsd-ud-2.6-200830
    - spanish-ancora-ud-2.5-191206
    - spanish-gsd-ud-2.5-191206
    - spanish-ancora-ud-2.4-190531
    - spanish-gsd-ud-2.4-190531
    - spanish-ud-2.0-170801
    - spanish-ancora-ud-2.0-170801
    - spanish-ud-1.2-160523
57. **Swedish**
    - *swedish-talbanken-ud-2.6-200830*
    - swedish-lines-ud-2.6-200830
    - swedish-talbanken-ud-2.5-191206
    - swedish-lines-ud-2.5-191206
    - swedish-talbanken-ud-2.4-190531
    - swedish-lines-ud-2.4-190531
    - swedish-ud-2.0-170801
    - swedish-lines-ud-2.0-170801
    - swedish-ud-1.2-160523
58. **Tamil**
    - *tamil-ttb-ud-2.6-200830*
    - tamil-ttb-ud-2.5-191206
    - tamil-ttb-ud-2.4-190531
    - tamil-ud-2.0-170801
    - tamil-ud-1.2-160523
59. **Telugu**
    - *telugu-mtg-ud-2.6-200830*
    - telugu-mtg-ud-2.5-191206
    - telugu-mtg-ud-2.4-190531
60. **Turkish**
    - *turkish-imst-ud-2.6-200830*
    - turkish-imst-ud-2.5-191206
    - turkish-imst-ud-2.4-190531
    - turkish-ud-2.0-170801
61. **Ukrainian**
    - *ukrainian-iu-ud-2.6-200830*
    - ukrainian-iu-ud-2.5-191206
    - ukrainian-iu-ud-2.4-190531
    - ukrainian-ud-2.0-170801
62. **Urdu**
    - *urdu-udtb-ud-2.6-200830*
    - urdu-udtb-ud-2.5-191206
    - urdu-udtb-ud-2.4-190531
    - urdu-ud-2.0-170801
63. **Uyghur**
    - *uyghur-udt-ud-2.6-200830*
    - uyghur-udt-ud-2.5-191206
    - uyghur-udt-ud-2.4-190531
    - uyghur-ud-2.0-170801
64. **Vietnamese**
    - *vietnamese-vtb-ud-2.6-200830*
    - vietnamese-vtb-ud-2.5-191206
    - vietnamese-vtb-ud-2.4-190531
    - vietnamese-ud-2.0-170801
65. **Welsh**
    - *welsh-ccg-ud-2.6-200830*
66. **Wolof**
    - *wolof-wtb-ud-2.6-200830*
    - wolof-wtb-ud-2.5-191206
    - wolof-wtb-ud-2.4-190531

C Languages

Language: Number of articles in that language

- English: 62
- Afrikaans: 62
- Arabic: 62
- Belarusian: 62
- Bulgarian: 62
- Catalan: 62
- Czech: 62
- Welsh: 62
- Danish: 62
- German: 62
- Estonian: 62
- Greek: 62
- Spanish: 62
- Basque: 62
- Persian: 62
- French: 62
- Galician: 62
- Korean: 62
- Armenian: 62
- Croatian: 62
- Indonesian: 62
- Hebrew: 62
- Latin: 62
- Latvian: 62
- Lithuanian: 62
- Hungarian: 62
- Dutch: 62
- Japanese: 62
- Polish: 62
- Portuguese: 62
- Romanian: 62
- Russian: 62
- Slovak: 62
- Slovenian: 62
- Serbian: 62
- Finnish: 62
- Swedish: 62
- Tamil: 62
- Turkish: 62
- Ukrainian: 62
- Urdu: 62
- Vietnamese: 62
- Chinese: 62
- Irish: 61
- Hindi: 61
- Marathi: 61
- Italian: 60
- Kazakh: 60
- Telugu: 59
- Scottish Gaelic: 58
- Classical Chinese: 51
- Maltese: 50
- Sanskrit: 50
- Uyghur: 49
- North Sami : 42
- Wolof: 38
- Gothic: 35
- Old Church Slavonic: 29

D Further Examples of the New Evaluation Measure

Sentence number 21 in the Spanish Gold Standard is “En el siglo XVIII d.C., el país se expandió mediante la conquista, la anexión y la exploración hasta convertirse en el tercer imperio más grande de la historia, el ruso, al extenderse desde Polonia, en poniente, hasta el océano Pacífico y Alaska, en el este.”, which can be translated to ‘In the 18th century AD, the country expanded through conquest, annexation, and exploration to become the third largest empire in history, the Russian Empire, stretching from Poland in the west to the Pacific Ocean and Alaska in the East.’. In GF-UD’s *eval* function, “d.C.” is split into two (*extra split*), which causes multiple missalignments:

```
    UDScore {udScore = 5.172413793103448e-2, udMatching = 0,
        udTotalLength = 58, udSamesLength = 3, udPerfectMatch = 0}
1  En  _  ADP  _  _  3  case                 1  En  _  ADP  _  _  3  case
2  el  _  DET  _  _  3  det                  2  el  _  DET  _  _  3  det
3  siglo  _  NOUN  _  _  10  obl         |   3  siglo  _  NOUN  _  _  11  obl
4  XVIII  _  NUM  _  _  3  compound      |   4  XVIII  _  NOUN  _  _  3  compound
5  d.C.  _  PROPN  _  _  4  flat         |   5  d.  _  NOUN  _  _  3  compound
6  ,  _  PUNCT  _  _  3  punct           |   6  C.  _  NOUN  _  _  3  compound
7  el  _  DET  _  _  8  det              |   7  ,  _  PUNCT  _  _  3  punct
8  país  _  NOUN  _  _  10  nsubj:pass   |   8  el  _  DET  _  _  9  det
9  se  _  PRON  _  _  10  expl:pv        |   9  país  _  NOUN  _  _  11  nsubj
10  expandió  _  VERB  _  _  0  root     |   10  se  _  PRON  _  _  11  obj
11  mediante  _  ADP  _  _  13  case     |   11  expandió  _  VERB  _  _  0  root
12  la  _  DET  _  _  13  det            |   12  mediante  _  ADP  _  _  14  case
13  conquista  _  NOUN  _  _  10  obl    |   13  la  _  DET  _  _  14  det
14  ,  _  PUNCT  _  _  16  punct         |   14  conquista  _  NOUN  _  _  11  obl
15  la  _  DET  _  _  16  det            |   15  ,  _  PUNCT  _  _  17  punct
16  anexión  _  NOUN  _  _  13  conj     |   16  la  _  DET  _  _  17  det
17  y  _  CCONJ  _  _  19  cc            |   17  anexión  _  NOUN  _  _  14  appos
18  la  _  DET  _  _  19  det            |   18  y  _  CCONJ  _  _  20  cc
19  exploración  _  NOUN  _  _  13  conj |   19  la  _  DET  _  _  20  det
20  hasta  _  ADP  _  _  21  mark        |   20  exploración  _  NOUN  _  _  14  conj
21-22  convertirse  _  _  _  _  _  _     |   21  hasta  _  ADP  _  _  22  mark
21  convertir  _  VERB  _  _  10  advcl  |   22-23  convertirse  _  _  _  _  _  _
22  se  _  PRON  _  _  21  expl:pv       |   22  convertir  _  VERB  _  _  11  advcl
23  en  _  ADP  _  _  26  case           |   23  se  _  PRON  _  _  22  obj
24  el  _  DET  _  _  26  det            |   24  en  _  ADP  _  _  27  case
25  tercer  _  ADJ  _  _  26  amod       |   25  el  _  DET  _  _  27  det
26  imperio  _  NOUN  _  _  21  obj      |   26  tercer  _  ADJ  _  _  27  amod
27  más  _  ADV  _  _  28  advmod        |   27  imperio  _  NOUN  _  _  22  obj
28  grande  _  ADJ  _  _  26  amod       |   28  más  _  ADV  _  _  29  advmod
29  de  _  ADP  _  _  31  case           |   29  grande  _  ADJ  _  _  27  amod
30  la  _  DET  _  _  31  det            |   30  de  _  ADP  _  _  32  case
31  historia  _  NOUN  _  _  26  nmod    |   31  la  _  DET  _  _  32  det
32  ,  _  PUNCT  _  _  34  punct         |   32  historia  _  NOUN  _  _  27  nmod
33  el  _  DET  _  _  34  det            |   33  ,  _  PUNCT  _  _  35  punct
34  ruso  _  ADJ  _  _  26  appos        |   34  el  _  DET  _  _  35  det
35  ,  _  PUNCT  _  _  34  punct         |   35  ruso  _  ADJ  _  _  27  appos
36-37  al  _  _  _  _  _  _             |   36  ,  _  PUNCT  _  _  35  punct
36  a  _  ADP  _  _  38  case            |   37  al  _  ADP  _  _  38  mark
37  el  _  DET  _  _  38  det            |   38-39  extenderse  _  _  _  _  _  _
38-39  extenderse  _  _  _  _  _  _      |   38  extender  _  VERB  _  _  22  advcl
38  extender  _  VERB  _  _  21  advcl   |   39  se  _  PRON  _  _  38  obj
39  se  _  PRON  _  _  38  expl:pv       |   40  desde  _  ADP  _  _  41  case
40  desde  _  ADP  _  _  41  case        |   41  Polonia  _  PROPN  _  _  38  obl
41  Polonia  _  PROPN  _  _  38  obl     |   42  ,  _  PUNCT  _  _  44  punct
42  ,  _  PUNCT  _  _  44  punct         |   43  en  _  ADP  _  _  44  case
43  en  _  ADP  _  _  44  case           |   44  poniente  _  NOUN  _  _  41  nmod
44  poniente  _  NOUN  _  _  41  nmod    |   45  ,  _  PUNCT  _  _  44  punct
45  ,  _  PUNCT  _  _  44  punct         |   46  hasta  _  ADP  _  _  48  case
```

```
46  hasta  _  ADP  _  _  48  case          |    47  el  _  DET  _  _  48  det
47  el  _  DET  _  _  48  det              |    48  océano  _  NOUN  _  _  38  obl
48  océano  _  NOUN  _  _  38  obl         |    49  Pacífico  _  PROPN  _  _  48  appos
49  Pacífico  _  PROPN  _  _  48  appos    |    50  y  _  CCONJ  _  _  51  cc
50  y  _  CCONJ  _  _  51  cc              |    51  Alaska  _  PROPN  _  _  48  conj
51  Alaska  _  PROPN  _  _  48  conj       |    52  ,  _  PUNCT  _  _  55  punct
52  ,  _  PUNCT  _  _  55  punct           |    53  en  _  ADP  _  _  55  case
53  en  _  ADP  _  _  55  case             |    54  el  _  DET  _  _  55  det
54  el  _  DET  _  _  55  det              |    55  este  _  NOUN  _  _  48  nmod
55  este  _  NOUN  _  _  48  nmod          |    56  .  _  PUNCT  _  _  11  punct
```

The new evaluation measure fixes the missalignments, which gives a notably higher score:

```
    # UDScore {udScore = 0.8928571428571429, udMatching = 1,
        udTotalLength = 56, udSamesLength = 50, udPerfectMatch = 0}
1  En  _  ADP  _  _  3  case                    1  En  _  ADP  _  _  3  case
2  el  _  DET  _  _  3  det                     2  el  _  DET  _  _  3  det
3  siglo  _  NOUN  _  _  10  obl                3  siglo  _  NOUN  _  _  11  obl
4  XVIII  _  NUM  _  _  3  compound             4  XVIII  _  NOUN  _  _  3  compound
5  d.C.  _  PROPN  _  _  4  flat          |     5  d.  _  NOUN  _  _  3  compound
                                                6  C.  _  NOUN  _  _  3  compound
6  ,  _  PUNCT  _  _  3  punct                  7  ,  _  PUNCT  _  _  3  punct
7  el  _  DET  _  _  8  det                     8  el  _  DET  _  _  9  det
8  país  _  NOUN  _  _  10  nsubj:pass          9  país  _  NOUN  _  _  11  nsubj
9  se  _  PRON  _  _  10  expl:pv         |     10  se  _  PRON  _  _  11  obj
10  expandió  _  VERB  _  _  0  root            11  expandió  _  VERB  _  _  0  root
11  mediante  _  ADP  _  _  13  case            12  mediante  _  ADP  _  _  14  case
12  la  _  DET  _  _  13  det                   13  la  _  DET  _  _  14  det
13  conquista  _  NOUN  _  _  10  obl           14  conquista  _  NOUN  _  _  11  obl
14  ,  _  PUNCT  _  _  16  punct                15  ,  _  PUNCT  _  _  17  punct
15  la  _  DET  _  _  16  det                   16  la  _  DET  _  _  17  det
16  anexión  _  NOUN  _  _  13  conj      |     17  anexión  _  NOUN  _  _  14  appos
17  y  _  CCONJ  _  _  19  cc                   18  y  _  CCONJ  _  _  20  cc
18  la  _  DET  _  _  19  det                   19  la  _  DET  _  _  20  det
19  exploración  _  NOUN  _  _  13  conj        20  exploración  _  NOUN  _  _  14  conj
20  hasta  _  ADP  _  _  21  mark               21  hasta  _  ADP  _  _  22  mark
21-22  convertirse  _  _  _  _  _  _            22-23  convertirse  _  _  _  _  _  _
21  convertir  _  VERB  _  _  10  advcl         22  convertir  _  VERB  _  _  11  advcl
22  se  _  PRON  _  _  21  expl:pv       |      23  se  _  PRON  _  _  22  obj
23  en  _  ADP  _  _  26  case                  24  en  _  ADP  _  _  27  case
24  el  _  DET  _  _  26  det                   25  el  _  DET  _  _  27  det
25  tercer  _  ADJ  _  _  26  amod              26  tercer  _  ADJ  _  _  27  amod
26  imperio  _  NOUN  _  _  21  obj             27  imperio  _  NOUN  _  _  22  obj
27  más  _  ADV  _  _  28  advmod               28  más  _  ADV  _  _  29  advmod
28  grande  _  ADJ  _  _  26  amod              29  grande  _  ADJ  _  _  27  amod
29  de  _  ADP  _  _  31  case                  30  de  _  ADP  _  _  32  case
30  la  _  DET  _  _  31  det                   31  la  _  DET  _  _  32  det
31  historia  _  NOUN  _  _  26  nmod           32  historia  _  NOUN  _  _  27  nmod
32  ,  _  PUNCT  _  _  34  punct                33  ,  _  PUNCT  _  _  35  punct
33  el  _  DET  _  _  34  det                   34  el  _  DET  _  _  35  det
34  ruso  _  ADJ  _  _  26  appos               35  ruso  _  ADJ  _  _  27  appos
35  ,  _  PUNCT  _  _  34  punct                36  ,  _  PUNCT  _  _  35  punct
36-37  al  _  _  _  _  _  _              |     37  al  _  ADP  _  _  38  mark
36  a  _  ADP  _  _  38  case
37  el  _  DET  _  _  38  det

38-39  extenderse  _  _  _  _  _  _             38-39  extenderse  _  _  _  _  _  _
38  extender  _  VERB  _  _  21  advcl          38  extender  _  VERB  _  _  22  advcl
39  se  _  PRON  _  _  38  expl:pv       |      39  se  _  PRON  _  _  38  obj
40  desde  _  ADP  _  _  41  case               40  desde  _  ADP  _  _  41  case
41  Polonia  _  PROPN  _  _  38  obl            41  Polonia  _  PROPN  _  _  38  obl
42  ,  _  PUNCT  _  _  44  punct                42  ,  _  PUNCT  _  _  44  punct
```

```
43  en  _  ADP  _  _  44  case                     43  en  _  ADP  _  _  44  case
44  poniente  _  NOUN  _  _  41  nmod            44  poniente  _  NOUN  _  _  41  nmod
45  ,  _  PUNCT  _  _  44  punct                  45  ,  _  PUNCT  _  _  44  punct
46  hasta  _  ADP  _  _  48  case                 46  hasta  _  ADP  _  _  48  case
47  el  _  DET  _  _  48  det                     47  el  _  DET  _  _  48  det
48  océano  _  NOUN  _  _  38  obl               48  océano  _  NOUN  _  _  38  obl
49  Pacífico  _  PROPN  _  _  48  appos          49  Pacífico  _  PROPN  _  _  48  appos
50  y  _  CCONJ  _  _  51  cc                     50  y  _  CCONJ  _  _  51  cc
51  Alaska  _  PROPN  _  _  48  conj             51  Alaska  _  PROPN  _  _  48  conj
52  ,  _  PUNCT  _  _  55  punct                  52  ,  _  PUNCT  _  _  55  punct
53  en  _  ADP  _  _  55  case                    53  en  _  ADP  _  _  55  case
54  el  _  DET  _  _  55  det                     54  el  _  DET  _  _  55  det
55  este  _  NOUN  _  _  48  nmod                 55  este  _  NOUN  _  _  48  nmod
56  .  _  PUNCT  _  _  10  punct                  56  .  _  PUNCT  _  _  11  punct
```

An example of *no split* can be found in sentence number 16 in the Spanish Gold Standard: “Fue fundado y dirigido por una clase guerrera noble de vikingos (llamados «varegos» en Europa Oriental) y sus descendientes.”, which can be translated to ‘It was founded and run by a noble warrior class of Vikings (called “Varegians” in Eastern Europe) and their descendants.’. In this case, the punctuation marks surrounding “varegos” have not been separated by UDPipe:

```
    UDScore {udScore = 0.5, udMatching = 0, udTotalLength = 22,
    udSamesLength = 11, udPerfectMatch = 0}
1  Fue  _  AUX  _  _  2  aux:pass           |    1  Fue  _  AUX  _  _  2  aux
2  fundado  _  VERB  _  _  0  root               2  fundado  _  VERB  _  _  0  root
3  y  _  CCONJ  _  _  4  cc                      3  y  _  CCONJ  _  _  4  cc
4  dirigido  _  VERB  _  _  2  conj              4  dirigido  _  VERB  _  _  2  conj
5  por  _  ADP  _  _  7  case                    5  por  _  ADP  _  _  7  case
6  una  _  DET  _  _  7  det                     6  una  _  DET  _  _  7  det
7  clase  _  NOUN  _  _  4  obj             |    7  clase  _  NOUN  _  _  2  obj
8  guerrera  _  ADJ  _  _  7  amod               8  guerrera  _  ADJ  _  _  7  amod
9  noble  _  ADJ  _  _  7  amod                  9  noble  _  ADJ  _  _  7  amod
10  de  _  ADP  _  _  11  case                   10  de  _  ADP  _  _  11  case
11  vikingos  _  NOUN  _  _  7  nmod             11  vikingos  _  NOUN  _  _  7  nmod
12  (  _  PUNCT  _  _  13  punct                 12  (  _  PUNCT  _  _  13  punct
13  llamados  _  ADJ  _  _  11  amod             13  llamados  _  ADJ  _  _  11  amod
14  «  _  PUNCT  _  _  15  punct            |    14  «varegos»  _  ADJ  _  _  13  obj
15  varegos  _  PROPN  _  _  13  obj        |    15  en  _  ADP  _  _  16  case
16  »  _  PUNCT  _  _  15  punct            |    16  Europa  _  PROPN  _  _  13  obl
17  en  _  ADP  _  _  18  case              |    17  Oriental  _  PROPN  _  _  16  flat
18  Europa  _  PROPN  _  _  13  obl         |    18  )  _  PUNCT  _  _  13  punct
19  Oriental  _  PROPN  _  _  18  flat      |    19  y  _  CCONJ  _  _  21  cc
20  )  _  PUNCT  _  _  13  punct            |    20  sus  _  DET  _  _  21  det
21  y  _  CCONJ  _  _  23  cc               |    21  descendientes  _  NOUN  _  _  7  conj
22  sus  _  PRON  _  _  23  det             |    22  .  _  PUNCT  _  _  2  punct
```

The new evaluation measure compares the sentences adding empty lines when necessary:

```
    # UDScore {udScore = 0.9166666666666666, udMatching = 1,
        udTotalLength = 24, udSamesLength = 22, udPerfectMatch = 0}
1  Fue  _  AUX  _  _  2  aux:pass               1  Fue  _  AUX  _  _  2  aux
2  fundado  _  VERB  _  _  0  root              2  fundado  _  VERB  _  _  0  root
3  y  _  CCONJ  _  _  4  cc                     3  y  _  CCONJ  _  _  4  cc
4  dirigido  _  VERB  _  _  2  conj             4  dirigido  _  VERB  _  _  2  conj
5  por  _  ADP  _  _  7  case                   5  por  _  ADP  _  _  7  case
6  una  _  DET  _  _  7  det                    6  una  _  DET  _  _  7  det
7  clase  _  NOUN  _  _  4  obj            |    7  clase  _  NOUN  _  _  2  obj
```

8 guerrera _ ADJ _ _ 7 amod		8 guerrera _ ADJ _ _ 7 amod
9 noble _ ADJ _ _ 7 amod		9 noble _ ADJ _ _ 7 amod
10 de _ ADP _ _ 11 case		10 de _ ADP _ _ 11 case
11 vikingos _ NOUN _ _ 7 nmod		11 vikingos _ NOUN _ _ 7 nmod
12 (_ PUNCT _ _ 13 punct		12 (_ PUNCT _ _ 13 punct
13 llamados _ ADJ _ _ 11 amod		13 llamados _ ADJ _ _ 11 amod
14 « _ PUNCT _ _ 15 punct	|	14 «varegos» _ ADJ _ _ 13 obj
15 varegos _ PROPN _ _ 13 obj		
16 » _ PUNCT _ _ 15 punct		
17 en _ ADP _ _ 18 case		15 en _ ADP _ _ 16 case
18 Europa _ PROPN _ _ 13 obl		16 Europa _ PROPN _ _ 13 obl
19 Oriental _ PROPN _ _ 18 flat		17 Oriental _ PROPN _ _ 16 flat
20) _ PUNCT _ _ 13 punct		18) _ PUNCT _ _ 13 punct
21 y _ CCONJ _ _ 23 cc		19 y _ CCONJ _ _ 21 cc
22 sus _ PRON _ _ 23 det		20 sus _ DET _ _ 21 det
23 descendientes _ NOUN _ _ 7 conj		21 descendientes _ NOUN _ _ 7 conj
24 . _ PUNCT _ _ 2 punct		22 . _ PUNCT _ _ 2 punct