Investigating Dominant Word Order on Universal Dependencies with Graph Rewriting

Hee-Soo Choi, Bruno Guillaume, Karën Fort, Guy Perrier

Investigating Dominant Word Order on Universal Dependencies with Graph Rewriting

Hee-Soo Choi
Sorbonne Université and
Université de Lorraine, CNRS, Inria,
LORIA, F-54000 Nancy, France
hee-soo.choi@loria.fr

Bruno Guillaume
Université de Lorraine, CNRS, Inria,
LORIA, F-54000 Nancy, France
Bruno.Guillaume@loria.fr

Karën Fort
Sorbonne Université / STIH and
Université de Lorraine, CNRS, Inria,
LORIA, F-54000 Nancy, France
karen.fort@loria.fr

Guy Perrier
Université de Lorraine, CNRS, Inria,
LORIA, F-54000 Nancy, France
Guy.Perrier@loria.fr

Abstract

This paper details experiments we performed on the Universal Dependencies 2.7 corpora in order to investigate the dominant word order in the available languages. For this purpose, we used a graph rewriting tool, GREW, which allowed us to go beyond the surface annotations and identify the implicit subjects. We first measured the distribution of the six different word orders (SVO, SOV, VSO, VOS, OVS, OSV) in the corpora and investigated when there was a significant difference in the corpora within a given language. Then, we compared the obtained results with information provided in the WALS database (Dryer and Haspelmath, 2013) and in Östling (2015). Finally, we examined the impact of using a graph rewriting tool for this task. The tools and resources used for this research are all freely available.

1 Introduction

Language typology has proven to be useful in natural language processing (NLP) (Bender, 2016; O'Horan et al., 2016), for example for improving performance in language transfer (Naseem et al., 2012; Ahmad et al., 2019) and joint learning.

As noted by O'Horan et al. (2016) "WALS is currently by far the most commonly-used typological resource in NLP due to its broad coverage of features and languages". However, the WALS database (Dryer and Haspelmath, 2013) has been compiled from the work of 55 linguists[1] and is not systemically based on a large quantity of data. Moreover, it does not provide all the considered features for all the languages it covers.

On the other hand, the Universal Dependencies (UD) framework (Nivre et al., 2016) provides a large number of corpora annotated in dependency syntax (in version 2.7, there are 183 corpora for 104 languages).

We decided to automatically extract from the UD corpora one of the most used features in NLP, the dominant word order, i.e. the way the subject (S), verb (V) and object (O) are ordered in a language (feature 81A in WALS). To do so, we use a freely available graph rewriting tool, which allows us to perform complex searches, to take into account the context of the construction and to add or modify the existing annotations to expose relations which are not directly accessible in the corpora.

These experiments led us to define what is a dominant word order, to observe the distribution in word orders within the corpora of a given language, to determine the frequency of the different word orders in all the considered corpora, and to compare the obtained results with those of existing databases, including WALS.

2 Previous Work

2.1 UD-based Typology

Dependency treebanks have already been used to investigate the order of subject and object in different languages. Liu (2010) presented a statistical overview of several binary parameters including SV vs VS, OV vs VO on 20 languages and compared their results with WALS'. However, their experiments were conducted before the UD framework, on treebanks with different annotation schemes.

To our knowledge, the closest work to ours is that of Östling (2015). He considered word order typology based upon the translated and aligned new testament in almost 1,000 languages and compared his results with WALS data. The main difference

[1] See: https://wals.info/author.

with our work is that for us, identifying the dominant word order is our goal and not just a production allowing the evaluation of a system. Besides, the data used in his experiment was generated automatically, rather than (at least partially) manually annotated. The produced data is available, so we were able to compare our results with his.

UD treebanks were also used to study word order freedom. Futrell et al. (2015) and Berdicevskis and Piperski (2020) examined the word order freedom of subjects and objects, focusing on the correlation with case marking.

More recently, Alzetta et al. (2018) applied a plausibility assessment algorithm to the UD treebanks to assess its usability in identifying typological features. They focused on the subject-verb and adjective-noun orders and experimented with three languages, English, Italian and Spanish. While their analysis is quite thorough, the algorithm they employed is not available, so their work cannot be extended.

Finally, Gerdes et al. (2019b) tested some of Greenberg's universals[2] on UD. Their work does include word order information but focuses on only two classes (the verb is before or after the object/subject). Besides, they decided to merge the treebanks for multi-corpora languages.

2.2 Enriching UD Annotations

There is no easy way to decide which dependency relations should be taken into account in order to observe word order dominance. In basic UD annotations, the tree restriction of usual dependency annotation frameworks impose some arbitrary choices: it is not possible to consider that the same token can be used twice as subject of different verbs. In our study, we try to overcome this limitation by making explicit some "syntactic" relations which cannot be expressed in UD (see Section 4.3). In Section 7, we compare what we observe using what we call *implicit subjects* with the same analysis on basic UD annotations only.

Similar types of enrichment have been proposed before, namely the Enhanced Universal Dependencies and the Deep Universal Dependencies.

Enhanced Universal Dependencies (EUD) were proposed in Schuster and Manning (2016). The goal of this work is to create an annotation which is more suitable for natural language understanding tasks, by making some of the implicit relations between words more explicit. Five kinds of new annotations are considered in this framework[3]: adding null nodes for elided predicates, propagating relations over conjuncts, adding subject relations for control and raising constructions, adding coreference in relative clause constructions and modifier labels that contain the preposition or other case-marking information. Unfortunately, adding these annotations requires manual annotation, therefore the EUD annotation layer is available in only 34 of the 183 treebanks in version 2.7. Moreover some of these 34 treebanks have only a subpart of the five extensions mentioned above.

The goal of the Deep Universal Dependencies (DUD) (Droganova and Zeman, 2019) is also to provide annotations adapted to natural language understanding. DUD expresses relations that are closer to predicate-argument structure than the annotations of EUD, using relations names (`arg1`, `arg2`, ...) borrowed from semantic frameworks like the Abstract Meaning Representation (AMR) (Banarescu et al., 2013). DUD is built automatically from EUD when annotations are available or with an automatic production of EUD for other corpora.[4]

3 Methodology

3.1 Taking the Corpora as Basis

Our study is based on the version 2.7 of UD, with 183 corpora and 104 languages available. Since our experiments consist in the extraction of statistics from data in corpora, we chose to eliminate corpora with fewer than 1,000 sentences, since we consider them too small to be representative of the language. Once this filter was applied, we obtained 141 corpora in 74 languages, which constitute the UD 2.7_{1K} corpus.

We decided to compile statistics at the corpus level rather than the language level, in order to observe variations between corpora of a given language and to compare the significance between them. 29 languages are represented by more than one corpus and for the 45 remaining ones we consider that "corpus equals language".

[2] Greenberg's universals are 45 linguistic universals dealing with basic word order, morphology and syntax based on 30 languages (Greenberg, 1963).

[3] See: https://universaldependencies.org/u/overview/enhanced-syntax.html.

[4] At the time of writing, DUD annotations are not available for version 2.7.

3.2 Defining a Dominant Word Order

Describing an order as a language's dominant order can have two meanings: either the order is the only possible one for the language, or the language exhibits several different orders and one is more frequently used.

In our experiments, we count the occurrences of the six possible orders in UD 2.7_{1K}, which means that our results are based only on the occurrence frequencies of the orders and therefore depend heavily on the composition of the corpora. Although we are aware of possible biases due to the corpus's degree of representativeness, our purpose is to determine a dominant order per corpus from raw data and to check whether we obtain results which are consistent with those of descriptive grammars.

Inspired by Dryer (2013), we consider the most frequent order as the dominant order for a given corpus provided that it is at least twice as frequent as the next most frequent. This means that for each corpus, we observe the ratio between the number of occurrences of the most frequent order with respect to the number of occurrences of the second most frequent order; if the ratio is greater than or equal to 2, the most frequent order is the dominant order, else we consider the corpus to be NDO (No Dominant Order). This allows us to classify as NDO corpora exhibiting little differences between two orders (for example, GERMAN-GSD with implicit subjects shows 35.7% SOV and 34.8% SVO). When the results differ among corpora of a given language, we study the corpora on a case by case basis.

3.3 Dealing with UD Specifics

In UD (Nivre et al., 2016), a given label can be ambiguous with respect to syntactic relations. For example, the labels `xcomp` and `ccomp` are used for both direct and indirect objects. Because of this limitation, we restrict our study to nominal objects, i.e. to `obj` relations. A similar difficulty arises with subjects. In UD, a personal subject is annotated with a relation `subj`, while an impersonal subject is annotated with the relation `expl`, which is also used for other relations with expletives. This ambiguity leads us to ignore impersonal subjects in our study[5].

Due to the tree constraint, some relations are not explicitly given in the data. In our study, this can affect subjects that can be shared by several verbs in coordination or through control of raising verbs. We call these hidden relations *implicit* relations.

For instance, consider the Polish sentence:

Kuba	*tego*	*nie*	*potrzebuje*	*ale*	*ma*	*to*	*od*
Kuba	this	not	need	but	has	this	from
mamy							
mom							

Kuba does not need this, but has it from her mother

There is an implicit subject relation between *ma* and *Kuba* which is not represented in the UD annotation. In our experiments, we ran an extended search on UD data with implicit subjects that can be predicted from surface syntax. Implicit objects also exist but it is not possible to recover them automatically from surface syntax. In the previous example, *tego* is the object of *potrzebuje* but it is not possible to determine if *tego* is an implicit object of *ma*.

Besides these issues, UD 2.7 includes two code-switching corpora: Turkish-German and Hindi-English. They were added as new "languages" and we therefore consider them here as such.

4 Going Deeper with Graph Rewriting

4.1 GREW

GREW is a graph rewriting tool dedicated to NLP applications, which can be used to query treebanks using graph patterns written with a specified syntax. Given a set of queries and a set of corpora, a script produces a table with the number of occurrences of each query in each corpus (see Section 4.2, for examples). An online interface to the tool is available[6], which enables users to observe examples in context within corpora and to interactively design and debug the patterns before running the script.

GREW also allows users to describe a set of transformations and to apply them to each item in a corpus. In this paper, we use this feature to enrich the available annotations (see Section 4.3).

4.2 Extraction Patterns

The patterns we use to extract data in UD 2.7_{1K} include two syntactic relations: subject[7] and object. As explained in Section 3.3, only nominal objects can be reliably recovered from UD annotations. To be consistent, we use the same restriction

[5] This is a limitation, in particular for some languages like French with impersonal redistribution.

[6] See: `http://match.grew.fr/`.

[7] A limitation of our experiment is that we cannot take into account cases in which the pronoun is not explicit (in pro-drop languages).

for the subject relation and focus only on nominal subjects (nsubj), without considering clausal subjects (csubj). For instance, the GREW pattern for SVO is presented in Figure 1.

```
pattern {
 V [upos=VERB];
 V -[1=nsubj|isubj]-> S; V -[1=obj]-> O;
 S << V; V << O
}
```

Figure 1: GREW pattern for SVO.

In UD, it is possible to include subtypes in relations, for instance the relation nsubj:pass can be used for a regular nominal subject in a passive construction. However, as these extensions are defined at the language level, we do not consider them here. The GREW syntax 1=subj|isubj allows to capture all relations that are either subj or isubj with or without subtypes.

4.3 Enriching UD Annotations

When used on UD annotations, the aforementioned extraction patterns present some limits as they only identify cases where the subject and the object are syntactically directly related to the same verb. However, there exist constructions admitting a subject and an object with two different governors. In our study, we consider two cases where the information can safely be recovered from surface annotations by adding implicit subjects, isubj, in an enriched UD annotation (see Section 3.3).[8]

The first case is coordination: when two clauses involving the same subject are linked by a coordinating conjunction with an ellipsis of the subject of the second clause, we add a new implicit subject to the head of the second clause. More technically, this is described by the rule in Figure 2: if two verbs V1 and V2 are linked by a conj relation and V2 does not have its own subject; then add the subject S1 of V1 as an isubj of V2. For instance in a sentence "*He obtains these things, but loses the ability to manage them.*" a relation isubj will be added from *loses* to *He*.

The second case we consider is control or raising. In UD, this is annotated with the relation xcomp between the two verbs. We can use a rule that is similar to the one in Figure 2 with xcomp instead of conj. In the sentence "*I should like to address one final point.*", the enriched annotation will show a relation isubj from *address* to *I*.

[8]Note that, with the new annotations, we obtain a graph.

```
rule conj {
  pattern {
    V1 [upos=VERB]; V2 [upos=VERB];
    V1 -[1=conj]-> V2;
    V1 -[1=nsubj]-> S1;
  }
  without { V2 -[1=nsubj]-> S2; }
  commands { add_edge V2 -[isubj]-> S1;}
}
```

Figure 2: GREW rule adding the isubj relation.

5 Determining Dominant Word Order in Multi-Corpora Languages

We detail here the results obtained for multi-corpora languages. For the mono-corpus languages, we examine our results as compared to WALS' and Östling (2015) in Section 6.

5.1 Intra-language Consistency

We obtain the number of occurrences of each of the six possible orders for each corpus in UD 2.7_{1K}. This data can be used to determine whether different corpora of a given language exhibit similar distributions. For this purpose, we compute the cosine between the 6-dimensional vectors for each corpus. This technique of comparing two feature vectors as a means of comparing two languages has already been used in several works on language typology (Georgi et al., 2010; Berzak et al., 2014). We expect two corpora of the same language to display similar distributions and therefore expect a cosine value close to 1.

The lowest value we observe is for two corpora of Romanian. Table 1 illustrates the vectors describing the distribution of the six possible orders for the three Romanian treebanks and Figure 3 represents as a heatmap the cosine values between these vectors.

Figure 4 reports the minimum cosine value among all possible pairs of corpora for the 29 multi-corpora languages.

Ten languages have a value below 0.95 and three have a value below 0.8 (Romanian, Hindi, Arabic). We present below a basic analysis of these results, either by seeking an explanation in the description of the corpora on the UD website or by asking language experts to examine the data.

Different text genres Four languages present corpora in different text genres, which could

	SVO	SOV	VSO	VOS	OSV	OVS
ROMANIAN_NONSTANDARD	38.07%	31.87%	9.66%	3.97%	1.71%	14.72%
ROMANIAN_RRT	85.32%	7.76%	1.12%	0.70%	1.18%	3.91%
ROMANIAN_SIMONERO	97.61%	0.97%	0.09%	0.09%	0.13%	1.10%

Table 1: Distribution vectors for the Romanian treebanks.

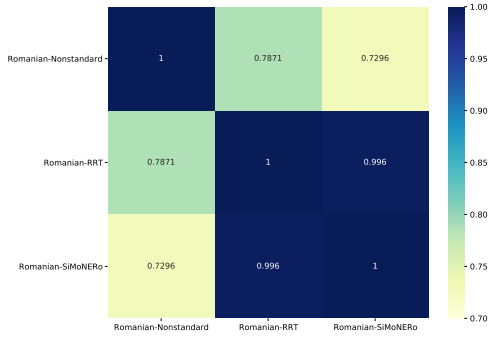

Figure 3: Cosine values between the three Romanian corpora in UD 2.7_{1K}.

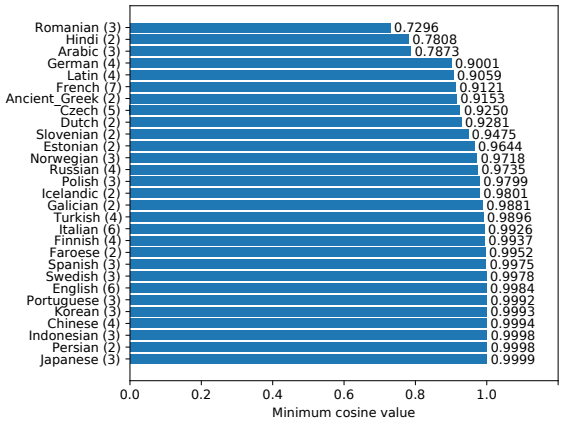

Figure 4: Multi-corpora (number in parenthesis) languages ordered by minimum cosine value.

explain the low cosine value: Dutch (0.928), French (0.912), Romanian (0.729) and Slovenian (0.947). One corpus in Romanian (ROMANIAN-NONSTANDARD) is dedicated to non-standard usage of that language (see Figure 3). Some corpora focus on specific types of texts: questions (FRENCH-FQB, which clearly stands out in Figure 5) and/or material from test suites and sentences from a reference grammar (DUTCH-ALPINO). French and Slovenian present corpora of spoken language (FRENCH-SPOKEN and SLOVENIAN-SSL). In Czech (0.925), one of the corpora (CZECH-FICTREE) contains only fiction

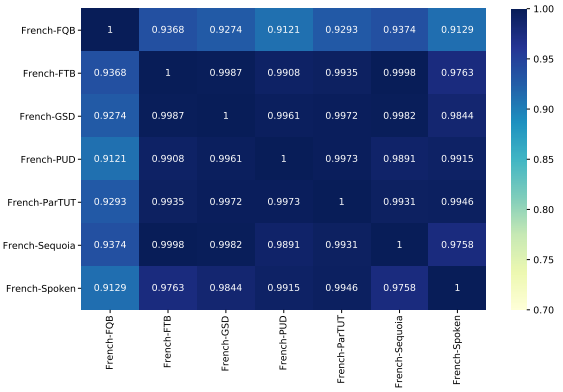

Figure 5: Cosine values between the seven French corpora in UD 2.7_{1K}.

and shows a higher proportion of SOV, while the four other Czech treebanks are clearly SVO.

Different text periods For two dead languages (Latin and Ancient Greek), corpora gather texts from very different historical periods, which could explain the differences. Latin texts range from 1^{st} century BC (LATIN-PERSEUS) to 13^{th} century Medieval Latin (LATIN-ITTB). For Ancient Greek, very different kinds of text are mixed: ANCIENT_GREEK-PROIEL contains both Herodotus (5^{th} century BC) and Bible texts[9]; the other corpus (ANCIENT_GREEK-PERSEUS) is a larger mix of several periods from Homer (8^{th} century BC) to Athenaeus (late second century). Undoubtedly, the fact that Latin and Ancient Greek are considered free word order languages amplifies the phenomenon. As for German, two corpora are NDO and two are SOV, GERMAN-HDT with a low ratio (2.01) and GERMAN-LIT. The latter is the only corpus to be composed of 18^{th} century texts.

Non-standard annotations In one of the two Hindi corpora (HINDI-HDTB), there is a large percentage of SVO cases (82.5%) where the object is a verb, in contradiction of the UD guidelines. If

[9]The cosine between these two subcorpora is 0.907.

we consider only nominal subjects and objects in the patterns, the cosine value rises to 0.993.

Language specifics In Arabic (modern standard), the basic order is VSO. However, SVO is used in cases of topicalization of the subject and in completives. The PADT corpus contains many titles of news articles featuring topicalization, which could explain the prevalence of SVO.

5.2 Dominant Word Order in Multi-Corpora Languages

For all multi-corpora languages with a minimum cosine value above 0.95, the dominant word order ratio consistently produces the same dominant order for all corpora of the language, except for Estonian which presents a SVO corpus (ESTONIAN-EDT) and a NDO corpus (ESTONIAN-EWT), corresponding to different text genres (fiction, news, nonfiction, academic *vs* blog, web, social). 14 multi-corpora languages are thus identified as SVO (Chinese, English, Faroese, Finnish, Galician, Icelandic, Indonesian, Italian, Norwegian, Polish, Portuguese, Russian, Spanish and Swedish) and four multi-corpora languages as SOV (Japanese, Korean, Persian and Turkish).

Out of the 10 multi-corpora languages with a minimum cosine value below 0.95, two present a clear dominant order SVO: French and Czech. As for Dutch and Ancient Greek, they both are NDO, but with inconsistent main orders: SOV/SVO and SVO/SOV. The six other languages present inconsistent dominant word orders. However, Romanian and Slovenian are both SVO in their standard or written form, even though one of their corpora is NDO (SVO/SOV). As for German, two out of four corpora are NDO (SOV/SVO) and two corpora are SOV. However, this result can be attributed to a threshold effect, since these two corpora present a SOV order at low ratios (2.01 for GERMAN-HDT, 2.53 for GERMAN-LIT).

Two corpora of Arabic are NDO (one is SVO/VSO and the other VSO/SVO) and one corpus is VSO. For the reasons explained in Section 5.1, we consider that the dominant order is most probably VSO. Hindi has one SOV and one NDO (SOV/SVO) corpora, but if we remove SVO occurrences probably due to annotation errors (i.e. O is a verb) in the latter, both corpora are clearly SOV. Latin is the language with the most heterogeneous corpora, with three NDO corpora (one SVO/SOV, two SOV/SVO) and one SOV corpus. These differences can probably be explained by the time range among texts.

Regarding the two code-switching languages, Hindi-English is considered NDO (SVO/SOV) which is consistent since English is SVO and Hindi SOV. As for Turkish-German, the corpus presents a SOV order, Turkish and German having this order in common.

6 Comparison with other Sources

Amongst the 74 languages available in UD 2.7_{1K}, WALS does not cover the seven dead languages, nor the two code-switching "languages". In addition, WALS does not provide Feature 81A for six languages. In Östling (2015), 22 languages are not in the database and seven are in neither sources, Galician, Hindi-English, Turkish-German and four dead languages (Old French, Old Russian, Sanskrit, Akkadian).[10]

6.1 Differences with WALS

Language	UD 2.7_{1K}	WALS
Amharic	1 NDO	SOV
Arabic	1 VSO, 2 NDO	VSO
Belarusian	1 SVO	NDO
Estonian	1 SVO, 1 NDO	SVO
German	2 SOV, 2 NDO	NDO
Greek	1 SVO	NDO
Hindi	1 SOV, 1 NDO	SOV
Mbya Guarani	1 NDO	SVO
Romanian	2 SVO, 1 NDO	SVO
Slovenian	1 SVO, 1 NDO	SVO
Urdu	1 NDO	SOV

Table 2: Differences with WALS (for UD 2.7_{1K}, we detail by corpora).

We compare our results with Feature 81A (Order of Subject, Object and Verb) in WALS[11]. We have 59 languages in common and we consistently observe the same dominant word order for 48 of these. In Table 2, we detail the remaining 11 languages where our observations are not fully consistent with WALS classification. Taking into account the explanations in Section 5.2 about multi-corpora languages, we have five languages with one corpus

[10] We are aware that WALS and Östling (2015) classifications only deal with transitive clauses, however the UD annotations do not allow us to extract them precisely.

[11] See: https://wals.info/feature/81A.

where we disagree with WALS: Amharic, Belarusian, Greek, Mbya Guarani and Urdu.

Belarusian and Greek can be considered relatively free word order languages, hence the NDO order in WALS. In our results, Belarusian is SVO with a ratio of 10.43, however the BELARUSIAN-HSE corpus is based on texts included in the Belarusian-Russian parallel subcorpus of the Russian National Corpus. Russian being a SVO language, this may explain the high proportion of SVO. Moreover, it is more common to find the SVO order as the basic order in written Belarusian. Similarly, the basic order in Greek being SVO, this may explain the ratio of 7.31 of SVO order in our results.

As for Mbya Guarani and Urdu, the most frequent order corresponds to the order in WALS. Mbya Guarani is NDO (SVO/SOV) with a ratio of 1.25 and Urdu NDO (SOV/SVO) with a ratio of 1.52. Finally, Amharic has an OVS order as the most frequent order, contrary to WALS'.

There are six languages present in WALS which do not have the Feature 81A: Galician, Faroese, Kazakh, Maltese, Naija and Slovak. Our results could therefore be used to enrich WALS' data.

6.2 Differences with Östling (2015)

Language	UD 2.7_{1K}	Östling
Amharic	1 NDO	SOV
Ancient Greek	2 NDO	SVO
Armenian	1 NDO	SVO
Basque	1 SOV	SVO
Dutch	2 NDO	SOV
Estonian	1 SVO, 1 NDO	SVO
German	2 SOV, 2 NDO	SOV
Hindi	1 SOV, 1 NDO	SOV
Hungarian	1 NDO	SVO
Latin	1 SOV, 3 NDO	SVO
Mbya Guarani	1 NDO	SVO
Romanian	2 SVO, 1 NDO	SVO
Slovenian	1 SVO, 1 NDO	SVO
Welsh	1 VSO	SVO

Table 3: Differences with Östling (2015) (for UD 2.7_{1K}, we detail by corpora).

The data presented in Östling (2015) is computed from the automatically aligned New Testament. The corpora are homogeneous and the data on which the dominant order is computed can sometimes be very small (for Hungarian, 127 structures *vs* 876 in our experiment). Moreover, Östling (2015) considers the single most prevalent order as the dominant one, without taking into account the difference with the second one.

We have 52 languages in common and observe the same dominant order for 38 of these. Table 3 reports what we observe for the 14 other languages. Out of these languages, 9 NDO languages have the same first order as Östling (2015). The 5 remaining ones (Amharic, Ancient Greek, Basque, Latin and Welsh) present different first orders.

7 Influence of Implicit Subjects

As said earlier, we decided to enrich the basic UD annotations, but the choice we made is quite arbitrary. In order to evaluate how this may have impacted our observations, we conducted the same experiment without taking into account the implicit subjects.

Table 4 lists the five languages for which the two experiments predict a different word order for at least one corpus or different first rank in NDO ordering. Adding `isubj` changes the dominant word order in four corpora: CZECH-FICTREE, ESTONIAN-EWT, GERMAN-HDT and TURKISH_GERMAN-SAGT. However, we note that in the four cases, one of the two ratios is close to the threshold. We observe an unexpected change for LATIN-LLCT which remains NDO, but with different top ranks (SOV/SVO with `isubj` and OSV/SVO without `isubj`). Latin is the language where we see the most important changes as can be observed by comparing the heatmaps in Figure 6.

Table 5 reports the corpora where the two experiments show a high difference (more than 5%) in term of first rank word order prediction. Only the WELSH-CCG corpus has a significant lower first rank with `isubj`, other corpora with a large difference present an higher first rank when `isubj` are taken into account. Again, the LATIN-LLCT exhibits a strange behavior with different first rank word order.

8 Conclusions and Perspectives

The main outcome of these experiments is the determination of the dominant word order for 74 languages, based on large amounts of annotated data. This result can be used for NLP applications.

On the linguistic side, our findings could be used to reinforce the results published in WALS and complete them in some cases. However, our results differ from WALS' for 11 languages, and for these,

		Without `isubj`		With `isubj`	
Language	Corpora	Order	Ratio	Order	Ratio
Czech	CAC	SVO	4.27	SVO	5.28
	CLTT	SVO	6.85	SVO	8.18
	FicTree	**NDO (SVO/SOV)**	1.97	**SVO**	2.20
	PDT	SVO	3.36	SVO	3.96
	PUD	SVO	6.58	SVO	6.17
Estonian	EDT	SVO	3.80	SVO	3.19
	EWT	**SVO**	2.05	**NDO (SVO/SOV)**	1.70
German	GSD	NDO (SOV/SVO)	1.03	NDO (SOV/SVO)	1.03
	HDT	**NDO (SOV/SVO)**	1.87	**SOV**	2.01
	LIT	SOV	2.30	SOV	2.53
	PUD	NDO (SOV/SVO)	1.47	NDO (SOV/SVO)	1.62
Latin	ITTB	NDO (SVO/SOV)	1.22	NDO (SVO/SOV)	1.12
	LLCT	**NDO (OSV/SVO)**	1.07	**NDO (SOV/SVO)**	1.40
	PROIEL	NDO (SOV/SVO)	1.21	NDO (SOV/SVO)	1.16
	Perseus	SOV	2.42	SOV	2.17
Turkish-German	**SAGT**	**NDO (SOV/SVO)**	1.95	**SOV**	2.22

Table 4: Corpora for which the word order changes with/without `isubj` and associated ratio (in bold, corpora for which the order changes).

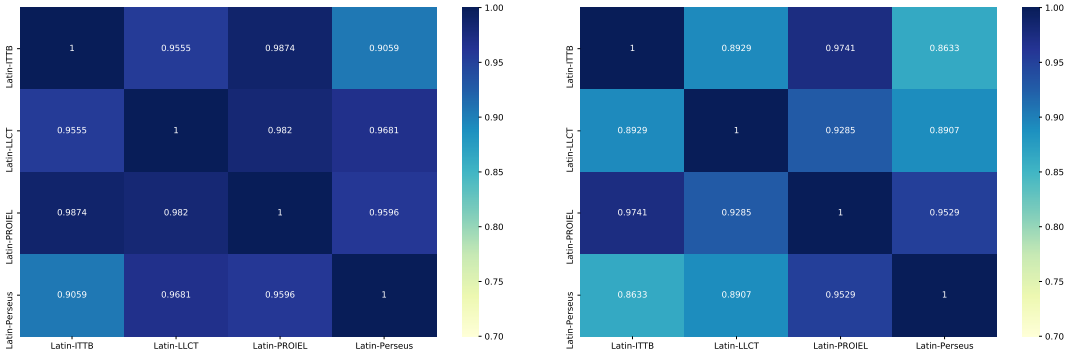

Figure 6: Cosine values between the Latin corpora, with `isubj` on the left, without `isubj` on the right.

a more thorough analysis should be conducted by specialists of said languages. We are planning to experiment using graph rewriting to explore other universals, like Greenberg's (Greenberg, 1963) or other missing features in WALS.

Graph rewriting can be used to enrich the UD annotations but it can also be used to reorganise more deeply the tree dependency graph. In Gerdes et al. (2019b), the observations were done on such a deeper reorganisation of the dependency tree structure, proposed in Surface Syntactic Universal Dependency (Gerdes et al., 2019a) which was already produced using GREW-based graph rewriting.

Our experiments can be replicated and extended: all the tools and resources are freely available and we also provide the patterns and scripts to be used[12].

Acknowledgements

We thank the reviewers for their useful remarks. We also wish to thank the colleagues who kindly took the time to answer our questions concerning some of the results we obtained in languages we do not speak: Sashi Narayan and Lydie Lemoine for Hindi, Hilda Mock for Arabic, Kim Gerdes for German and Vincent Vandeghinste for Dutch. The internship of the first author, during which this work has been done, was funded by the Lorraine Université d'Excellence OLKI research project[13].

[12] https://gitlab.inria.fr/ud-greenberg/ranlp-2021

[13] https://olki.loria.fr/

Corpora	first rank without `isubj`	first rank with `isubj`	diff
WELSH-CCG	VSO (80.6%)	VSO (71.1%)	-9.5%
LATIN-LLCT	**OSV (32.0%)**	**SOV (42.7%)**	+10.7%
AKKADIAN-RIAO	SOV (56.6%)	SOV (66.3%)	+9.7%
ICELANDIC-ICEPAHC	SVO (57.0%)	SVO (63.4%)	+6.4%
WOLOF-WTB	SVO (59.3%)	SVO (64.8%)	+5.5%
OLD_CHURCH_SLAVONIC-PROIEL	SVO (52.0%)	SVO (57.4%)	+5.5%
CZECH-FICTREE	SVO (47.6%)	SVO (52.7%)	+5.1%

Table 5: Corpora for which the first rank difference with or without `isubj` is greater than 5% (in bold, the corpus for which the first rank changes).

References

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2440–2452, Minneapolis, Minnesota. Association for Computational Linguistics.

Chiara Alzetta, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2018. Universal dependencies and quantitative typological trends. a case study on word order. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.

Emily M. Bender. 2016. Linguistic typology in natural language processing. *Linguistic Typology*, 20:645–660.

Aleksandrs Berdicevskis and A. Piperski. 2020. Corpus evidence for word order freezing in russian and german. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 26–33, Barcelona, Spain (Online).

Yevgeni Berzak, Roi Reichart, and Boris Katz. 2014. Reconstructing native language typology from foreign language usage. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 21–29, Ann Arbor, Michigan. Association for Computational Linguistics.

Kira Droganova and Daniel Zeman. 2019. Towards deep Universal Dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 144–152, Paris, France.

Matthew S. Dryer. 2013. Determining dominant word order. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

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

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Quantifying word order freedom in dependency corpora. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 91–100, Uppsala, Sweden. Uppsala University, Uppsala, Sweden.

Ryan Georgi, Fei Xia, and William Lewis. 2010. Comparing language similarity across genetic and typologically-based groupings. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 385–393, Beijing, China. Coling 2010 Organizing Committee.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2019a. Improving Surface-syntactic Universal Dependencies (SUD): surface-syntactic relations and deep syntactic features. In *TLT 2019 - 18th International Workshop on Treebanks and Linguistic Theories*, Paris, France.

Kim Gerdes, Sylvain Kahane, and Xinying Chen. 2019b. Rediscovering greenberg's word order universals in UD. In *Proceedings of the Third Workshop on Universal Dependencies (UD Workshop, SyntaxFest 2019)*, pages 124–131, Paris, France. Association for Computational Linguistics.

Joseph H. Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Human Language*, pages 73–113. MIT Press, Cambridge, Mass.

Haitao Liu. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 120(6):1567–1578. Contrast as an information-structural notion in grammar.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency

parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 629–637, Jeju Island, Korea. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in natural language processing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1297–1308, Osaka, Japan.

Robert Östling. 2015. Word order typology through multilingual word alignment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 205–211, Beijing, China. Association for Computational Linguistics.

Sebastian Schuster and Christopher D. Manning. 2016. Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2371–2378, Portorož, Slovenia.