Corpus-based Language Universals Analysis using Universal Dependencies

Hee-Soo Choi, Bruno Guillaume, Karën Fort

Corpus-based Language Universals Analysis using Universal Dependencies

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

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

Karën Fort
Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
Sorbonne Université, F-75006 Paris, France
karen.fort@loria.fr

Abstract

This paper presents experiments aiming at verifying Greenberg's universals based on Universal Dependencies (UD) corpora (de Marneffe et al., 2021). We adopt a corpus-based approach that allows us to highlight inconsistencies between corpora of the same language and to explore the causes of these inconsistencies. In addition to intra-language inconsistency, our analysis on 141 corpora, i.e. 74 languages, also shows cross-language inconsistency and questions the adaptability of UD annotations for some linguistic concepts.

1 Introduction and Related Work

Despite the evident diversity of languages, similarities between them have allowed linguists to extract common properties called language universals. The notion itself is difficult to define with the fine line between an absolute universal, a property valid over all languages, and a strong tendency. Linguists generally opposed two approaches: the typological approach and the generative approach (Comrie, 1989; Croft, 2003). While the latter insists on the common genetic character of languages (Chomsky, 1982), the typological approach relies more on empirical data from different languages, represented in Greenberg's work (Greenberg, 1966). Considered as one of the pioneers of modern typology, Greenberg defined 45 universals dealing with basic word order, morphology and syntax, based on 30 languages. In particular, he established a classification of the 30 languages according to three basic factors: the existence of prepositions as against postpositions, the order of subject, object and verb in declarative sentences with nominal subject and object and the order of the qualifying adjective in relation to the noun.

The development of natural language processing (NLP) tools and digital resources, especially treebanks, allowed for the multiplication of multilingual research work aiming at testing typological features on a large number of languages (Dryer, 1992; Liu, 2010; Östling, 2015; Futrell et al., 2015; Levshina, 2019). More recently, UD treebanks have been used in several typological and language universal studies: Sharma et al. (2019) showed that language networks constructed from UD treebanks cluster correctly languages based on Greenberg's word order typology, allowing for a representation of linguistic generalizations, while Gerdes et al. (2019) explored Greenberg's universals in a quantitative way on treebanks annotated in dependency syntax using an annotation scheme derived from UD (SUD) and identified quantitative universals using diagrams, which they call "typometric diagrams" (Gerdes et al., 2021). Moreover, Dönicke et al. (2020) proposed a framework to investigate Greenberg's typological universals on UD by using real-valued logics.

Our study aims at examining Greenberg's observations on the basis of the large amount of data provided by the UD corpora.[1] Our results represent empirical information that can confirm or even com-

[1]We decided not to use SUD for our study for two reasons. First, many occurrences of verbal forms with a subject and an object would require a much more complex set of patterns (a specific pattern is needed when there is one auxiliary which is the governor of the subject in SUD; another one when there are two auxiliaries...). Second, most of SUD corpora are produced automatically from UD data and we prefer to work on original data.

plete existing typological databases. Although our experiments are in line with the work of Gerdes et al. (2021), we decided not to group corpora of the same language together to preserve their specificities and evaluate their influence. Our paper is structured as follows: in Section 2, we present our sample of 141 UD corpora as well as the tool and metrics we used for our experiments; in Section 3, we describe our experiments on determining three word orders and verifying four universals; finally, in Section 4, we discuss issues in UD annotations that our analysis brought to light.

2 Material and Methods

2.1 From UD 2.7 to UD 2.7_{1K}

In our work, we used UD version 2.7, which contains 104 languages and 183 corpora. Since we consider each corpus separately, we decided not to take into account the corpora containing less than 1,000 sentences, which we consider too small to be representative. We thus obtained a set of 74 languages and 141 corpora, which constitutes our experimental data, UD 2.7_{1K}. There is a strong bias in the languages represented in UD in general, therefore in our data: 76% of the sentences and 65% of the corpora in UD 2.7_{1K} belong to the Indo-European family. Of the 30 languages in Greenberg's sample, only 14 are present in UD 2.7_{1K}, including all the European languages used by Greenberg.

2.2 GREW: a tool for fine-grained observation on corpora

In our experiments, we need to count the number of occurrences of specific patterns in each corpus. We use the graph matching mechanism available in the GREW tool[2] (Guillaume, 2021). It allows to write complex patterns, with combination of constraints on dependency relations, on node features; it is also possible to add negative constraints to refine the queries. Figure 1 shows a query with constraints on relations, on words, on word order (<< symbol) and with negative constraints (*without* part).

2.3 Quantifying qualitative concepts

We used the same method as in (Choi et al., 2021) and defined the notion of dominant word order quantitatively as follows. We computed the ratio between the two most frequent orders: if it is greater than or equal to 2, the most frequent order is the dominant order, if it is strictly inferior to 2, the corpus is considered to have no dominant order (NDO). In the case where only two orders are possible (for example, adjective-noun / noun-adjective), if the ratio is greater than 2, the frequency of the most frequent order is greater than $\frac{2}{3}$.

In order to compare if two corpora show the same distribution of some observations, we compute the cosine value between the two vectors representing the proportion of each observation. For instance with the orders between Subject, Object and Verb, we have vectors with six dimensions for the proportion of the six possible relative orders. We expect two corpora of the same language to have the same distribution and thus, the cosine to be closed to 1. In our experiments, a lower cosine value indicates a greater inconsistency between the two corpora.

3 Experiments and Results

3.1 Order of Subject, Object and Verb

To be consistent with Greenberg's observations, we determined a dominant order of Subject (S), Object (O) and Verb (V) in declarative sentences with nominal subject and object. Inspired by Choi et al. (2021), we decided to refine the GREW pattern by setting the POS tags to filter the nominal subjects and objects (see Figure 1). Indeed, even if the `nsubj` relation inherently involves nominal dependents, we noticed that this is not the case in all corpora (see Section 4). Furthermore, UD annotations do not allow us to filter declarative sentences precisely, we therefore relied on punctuation by eliminating all sentences whose verb is linked to a question mark or an exclamation mark (which corresponds to the *without* part of the pattern). However, this method remains fragile since corpora without punctuation exist and we

[2] `https://grew.fr`

have no certainty that the punctuation system is the same in all the concerned languages. Moreover, we could not filter out imperative sentences because some corpora do not indicate the mood of the verbs.

```
pattern {
  V [upos=VERB];
  V -[1=nsubj]-> S; S[upos=PROPN|PRON|NOUN];
  V -[1=obj]-> O; O[upos=PROPN|PRON|NOUN];
  S << V; V << O;
}
without {
  V -[punct]-> P; P [lemma="?"|"!"];
}
```

Figure 1: GREW pattern for SVO order.

Using criterion described in Section 2.3 on 141 corpora, we observed that 91 are SVO, 24 are SOV, 4 are VSO and 22 have no dominant order (NDO). Of the 29 multi-corpora languages, all corpora of the same language show the same dominant order, except for six languages: German, Arabic, Ancient Greek, Latin, Dutch and Romanian. Our analyses of the inconsistency in these six languages agree with those of Choi et al. (2021), however we gained in consistency with seven languages with a minimum cosine value of less than 0.95 against ten languages for them.

3.2 Prepositions/Postpositions

In his work, Greenberg uses the terms "prepositional language" or "language with prepositions" but does not provide a precise definition of what this means. Inspired by WALS (Dryer, 2013b), we decided to extract occurrences of adpositions linked to a noun phrase, which corresponds to a noun, a pronoun or a proper noun in UD. Figure 2 shows a GREW pattern for the order adposition - noun phrase (preposition). We fixed a node A labeled as an adposition (ADP), related to a noun phrase N by a `case` relation. Moreover, we decided to exclude sentences where the adposition is part of a multi-words expression, marked with a `fixed` or a `flat` relation in UD.

```
pattern {
  A [upos=ADP];
  N [upos=NOUN|PRON|PROPN];
  N -[1=case]-> A;
  A << N;
}
without {
  A -[1=fixed|flat]-> X
}
```

Figure 2: GREW pattern for adposition - noun phrase order.

The results show a marked tendency for one of the two types of adposition. Of the 141 corpora, 108 have prepositions (Pr), 30 have postpositions (Post), one corpus has no dominant order (`Chinese-PUD`) and two corpora show no occurrence of either type (`Korean-PUD` and `Sanskrit-Vedic`). For the 29 multi-corpora languages, all corpora of the same language present the same type of adposition, except for Chinese and Korean due to the two atypical corpora. To measure the degree of consistency, we computed the cosine values and extracted the minimum cosine between all possible pairs. 27 languages show a high consistency between corpora with a minimum cosine above 0.99. Only two languages have minimum cosine values below 0.99: Chinese (0.8771) and Persian (0.9658)[3].

Chinese The `Chinese-PUD` is the only Chinese corpus without a dominant order. The results on other Chinese corpora in Table 1 show that they generally use prepositions[4]. The minimum cosine value of 0.8771 is observed between the `Chinese-PUD` and the `Chinese-GSD`.

[3]For Korean, we have excluded the corpus `Korean-PUD` which present no occurrence of either adposition.

[4]There is a `Chinese-GSDSimp` which is a simplified Chinese version of the corpus of the `Chinese-GSD`. We consider only the second one here.

Corpus	Pr	Post
Chinese-GSD	99.92%	0.08%
Chinese-HK	87.40%	12.60%
Chinese-PUD	**64.57%**	**35.43%**

Table 1: Proportions of prepositions and postpositions in the Chinese corpora.

Corpus	acl	appos	case	case:loc	conj	mark
Chinese-GSD	**98.41%**	0.10%	0.20%	0.00%	0.20%	1.09%
Chinese-HK	0.00%	0.00%	9.68%	**90.32%**	0.00%	0.00%
Chinese-PUD	0.00%	0.00%	0.00%	**99.39%**	0.00%	0.61%

Table 2: Syntactic relation types between the noun phrase and the postposition in the Chinese corpora.

Table 2 presents syntactic relation types between postpositions and noun phrases. In the Chinese-GSD, postpositions are not annotated with a case relation but with a acl relation, they are therefore not covered by the pattern we used. In contrast, the Chinese-PUD has 99% of postpositions identified with a case:loc relation, a subtype of the case relation covered by the pattern. The same goes for the Chinese-HK with 90.32% of case:loc relation and 9.68% case relation. One could imagine that the relations case and acl involve postpositions of a different nature, but all Chinese corpora use postpositions that generally correspond to location-indicating postpositions such as: 上 (shàng = above/on), 中 (zhōng = between, in the middle), 下 (xià = below/under). In this case, the annotators simply made a different choice of annotation. Taking into account the postpositions related by the relation acl in Chinese-GSD, we obtain a distribution of 72.87% of prepositions and 27.13% of postpositions, which is consistent with the distributions of the Chinese-PUD and the Chinese-HK.

The UD guidelines describe the acl relation as a clausal modifier of noun. In Chinese, clausal modifiers may precede the noun and may be formed with the particle 的. They may also follow the noun, in which case they will be juxtaposed after the noun[5]. The acl relation must link a noun and the head of clause. It is quite rare to find an adposition depending on a relation acl, which would be contrary to the UD guidelines. We therefore assume that there is an inconsistency in annotations in these corpora[6].

Conflict between ADP and PART tags The distinction between a particle and an adposition is difficult to define. Adpositions, coordinating conjunctions and subordinating conjunctions are considered particles in UD, but the guidelines specify that the most precise label should be used.

Korean is an agglutinating language using postpositions at the end of words to designate their functions. These postpositions are sometimes referred to as particles. In UD, it is stated that the PART tag designates a functional word and should only be used when the word does not fit the definitions of other functional words such as adpositions and conjunctions[7]. However, in the Korean-PUD corpus, the ADP tag is not used and all adpositions are annotated with the PART tag. Moreover, in version 2.7 of the corpus, adpositions are not linked with a relation case but with a relation dep:prt[8].

As in Korean-PUD, our pattern did not allow to find any occurrences of adpositions for the Sanskrit-VEDIC because they are annotated as a particle. With the PART annotation, we obtain 79.76% of postpositions and thus 20.24% of prepositions. Sanskrit being a dead language, it is not listed in WALS and we cannot offer a more in-depth analysis without a specialist of the language.

Finally, in Persian, the cosine between the two corpora is 0.9658. While both corpora present mostly prepositions, the Persian-PerDT presents 21.18% of postpositions and the Persian-Seraji only 0.01% because postpositions are annotated as PART and not ADP in the corpus.

[5] https://universaldependencies.org/zh/dep/acl, August 2021.

[6] This inconsistency was corrected after we entered an issue on the GitHub.

[7] https://universaldependencies.org/u/pos/PART, August 2021.

[8] In version 2.8, the relation dep:prt has been replaced by the relation case.

Comparison with WALS WALS presents the order of adposition and noun phrase under the feature 85A (Dryer, 2013b). On the 74 languages of our study: 50 languages have the same order as WALS, 21 languages are not in WALS (mainly dead languages) or do not present the feature 85A, three languages do not have the same order: Chinese, Cantonese and Amharic which are considered NDO by WALS.

The `Cantonese-HK` has 87.84% of prepositions. Since Cantonese and Chinese have relatively similar syntax, we can assume that Cantonese can indeed have prepositions and postpositions but that this corpus either has few constructions with postpositions or the annotations do not allow us to cover all cases with our pattern. As for the Amharic corpus (`Amharic-ATT`), it has 83.81% of prepositions. Greenberg also considers Amharic to be a prepositional language.

3.3 Order of the Adjective and the Noun

Determining a dominant order of Adjective (Adj) and Noun (N) turns out to be simpler, since the concepts are present in all languages and are precisely annotated. We used the pattern presented in Figure 3, where we define a noun N and an adjective Adj linked by the relation `amod`.

```
pattern {
  N [upos=NOUN];
  Adj [upos=ADJ];
  N -[1=amod]-> Adj;
  Adj << N
}
```

Figure 3: GREW pattern for the Adj-N order.

Of the 141 corpora, 84 present the Adj-N order, 43 present the N-Adj order and 14 do not have a dominant order: one French corpus, three Italian corpora, two Polish corpora, the two Ancient Greek corpora, the four Latin corpora, the Old Russian corpus and the Gothic corpus.

All corpora of the same language have the same order, except for French, Italian and Polish. 24 multi-corpora languages have a high consistency in their corpora with minimum cosine values above 0.99. Five languages are below 0.99: Italian (0.9146), French (0.9171), Romanian (0.9771), Polish (0.9779) and Latin (0.9836). In French, Italian and Polish, the adjective can be placed either before the noun or after the noun. Generally, the place of the adjective does not change the meaning of the sentence, except in special cases[9]. However, there are rules such as placing short adjectives before the noun or placing adjectives of color after the noun in French and Italian. It is also possible to change the place of the adjective to emphasize the quality carried by the adjective.

French The `French-Spoken` is the only corpus of French without a dominant order. This corpus is the only spoken French corpus among the seven French corpora. As shown in Table 3, the written French corpora have very similar proportions with about 70% N-Adj, while the `French-Spoken` has both orders relatively homogeneously. The minimun cosine value is 0.9171 between the `French-Spoken` and the `French-Sequoia`.

Corpus	Adj-N	N-Adj
French-FQB	29.31%	70.69%
French-FTB	30.86%	69.14%
French-GSD	29.95%	70.05%
French-ParTUT	27.10%	72.90%
French-PUD	31.13%	68.87%
French-Sequoia	24.24%	75.76%
French-Spoken	**46.71%**	**53.29%**

Table 3: Proportions of Adj-N and N-Adj orders in the French corpora.

[9]For example, in French "une ancienne maison" (what used to be a house) has a different meaning than "une maison ancienne" (an old house).

The written corpora are mostly from newspaper articles and Wikipedia, therefore the sentences present adjectives from more scientific domains which tend to be placed after the noun. On the other hand, the oral French corpus shows an over-representation of common adjectives such as "petit" (small) and "grand" (big) which are placed before the noun.

Italian In Italian, three corpora have a dominant N-Adj order and three have no dominant order. Of the three corpora without dominant order, two are tweets corpora, the `Italian-PoSTWITA` and the `Italian-TWITTIRO`. The third corpus is the `Italian-ParTUT`, but this result is certainly due to a threshold effect, the distribution being 33.99% and 66.01%.

As in French, the three N-Adj dominant order corpora as well as the `Italian-ParTUT` have texts extracted from newspaper articles and Wikipedia. The distribution of the orders is similar in these corpora with a proportion around 70% for the N-Adj order. The results show that the genre of the corpora has an influence on the type of adjectives used and thus on the place of the adjective in relation to the noun.

Corpus	Adj-N	N-Adj
Italian-ISDT	29.89%	70.11%
Italian-ParTUT	**33.99%**	**66.01%**
Italian-PoSTWITA	**41.31%**	**58.69%**
Italian-PUD	31.04%	68.96%
Italian-TWITTIRO	**51.69%**	**48.31%**
Italian-VIT	31.76%	68.24%

Table 4: Proportions of Adj-N and N-Adj orders in the Italian corpora.

Polish In Polish, adjectives can be placed before or after the noun, except for short adjectives which are always placed before. Placing an adjective after the noun is possible to emphasize it. Unlike in French and Italian, we cannot explain the differences in distribution by the genre of the texts. The corpora are all composed of various genres: newspaper articles, fiction, Wikipedia.

Corpus	Adj-N	N-Adj
Polish-LFG	71.30%	28.69%
Polish-PDB	**64.48%**	**35.52%**
Polish-PUD	**59.73%**	**40.27%**

Table 5: Proportions of Adj-N and N-Adj orders in the Polish corpora.

Dead languages Latin, Ancient Greek, Old Russian and Gothic corpora do not have a dominant order of Adjective and Noun. The proportions between the two orders are homogeneous, with ratios very close to 1 for these languages. We therefore can assume that these languages allow for both orders, as they are considered free word order languages (Levshina, 2019).

Comparison with WALS WALS presents the order of Adjective and Noun in the feature 87A (Dryer, 2013a). Of the 74 languages: 54 languages show the same order as WALS, 17 languages are not in WALS or do not have the feature 87A, three languages have some corpora with no dominant order in our results: Italian, French and Polish. However, the corpora where order is defined for these languages are all consistent with WALS' results.

3.4 Greenberg's Universal 1

Universal 1 *In declarative sentences with nominal subject and object, the dominant order is almost always one in which the subject precedes the object.*

The three orders where the subject precedes the object are: SVO, SOV and VSO. The results obtained show that on 141 corpora 91 are SVO, 24 are SOV, four are VSO and 22 are NDO. Our results therefore

confirm Greenberg's Universal 1 for 119 corpora and 59 languages. For NDO corpora, without taking into account the ratio, the most frequent order is either SVO, SOV or VSO, except for two corpora: the `Amharic-ATT` and the `Latin-LLCT` (see Table 6).

Corpus	SVO	SOV	VSO	VOS	OSV	OVS
Amharic-ATT	4.70%	**28.86%**	8.95%	0.45%	12.97%	**44.07%**
Latin-LLCT	**30.18%**	**29.00%**	4.31%	0.93%	**32.65%**	2.91%

Table 6: Distribution of Subject, Object and Verb orders in `Amharic-ATT` and `Latin-LLCT`.

For Amharic, the most frequent order is OVS at 44.07%, followed by SOV at 28.86%. The OVS order remains a rather rare case of dominant order. According to WALS and Greenberg, Amharic is a SOV language, which is the second most frequent order in our results. Without a speaker of Amharic, we can only assume that there are either: i) annotation errors in the corpus, ii) the possibility of using the OVS order in some sentences, iii) an influence of the genre of the texts of the corpus, which is very heterogeneous since the sentences can be examples of grammars, extracted from texts of fiction, from the bible and from newspapers among others.

The `Latin-LLCT` has a relatively homogeneous distribution between three orders SVO, SOV and OSV which are frequent at about 30%. Latin being a free word order language, this may explain the absence of a dominant order. Moreover, the texts come from different centuries, which has some influence on the way sentences are constructed.

3.5 Greenberg's Universals 3 and 17

Universals 3 and 17 concern VSO languages, so we treat them together in this section.

Universal 3 *Languages with dominant VSO order are always prepositional.*

Universal 17 *With overwhelmingly more than chance frequency, languages with dominant order VSO have the adjective after the noun.*

Table 7 shows that our results confirm Greenberg's Universals 3 and 17, but only on four corpora. For Arabic, two other corpora are available but do not present a dominant order, with a conflict between VSO (31.20% and 49.45%) and SVO orders. These corpora are largely prepositional and have almost 100% N-Adj order. It is interesting to note that `Arabic-PADT` exhibits these characteristics while having a higher frequency of SVO than VSO (48.15%). This result is consistent with Greenberg's observation that SVO languages are more correlated with prepositionnaly and N-Adj order than postpositionnaly and Adj-N order.

Corpus	VSO	Pr	N-Adj
Arabic-NYUAD	54.56%	99.97%	99.69%
Irish-IDT	99.14%	99.78%	98.91%
Scottish_Gaelic-ARCOSG	97.49%	100%	84.82%
Welsh-CCG	78.57%	100%	82.54%

Table 7: Proportions of prepositions (Pr) and N-Adj order in the VSO corpora.

3.6 Greenberg's Universal 4

Universal 4 *With overwhelmingly greater than chance frequency, languages with normal SOV order are postpositional.*

According to our results, 24 corpora have a dominant SOV order, which corresponds to 15 languages. To visualize this universal, we used the typometric graph of Gerdes et al. (2021) in Figure 4.

Corpora at the top right of the figure correspond to very strongly SOV and postpositional corpora. The languages represented are: Bambara, Hindi, Japanese, Kazakh, Korean, Telugu, Turkish,

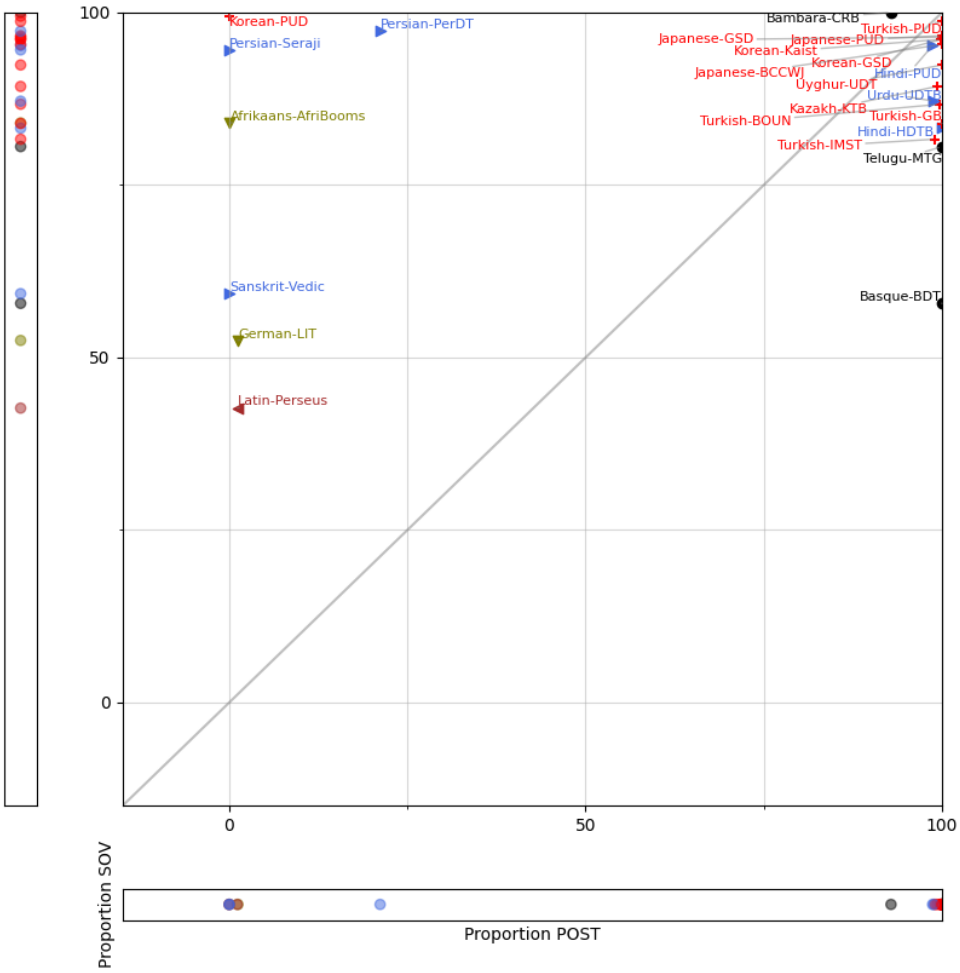

Figure 4: SOV corpora according to their proportions of postpositions.

Urdu and Uyghur. At the top left, some corpora are strongly SOV but with very few postpositions: `Afrikaans-AfriBooms`, `Persian-Seraji`, `Persian-PADT` and `Korean-PUD`. Afrikaans and Persian are both SOV and prepositional languages. The `Korean-PUD` does not present postpositions due to annotation issues detailed in section 3.2. We can also note that the two other corpora of Korean largely present postpositions.

Moreover, three corpora stand out with a percentage of SOV around 50%: the `German-LIT`, the `Latin-Perseus` and the `Sanskrit-Vedic`. The German and Latin corpora have a dominant SOV order but their other corpora do not have a dominant order. These two languages are also generally considered to have no dominant order. We assume that Greenberg's formulation "normal SOV order" allows us not to take into account these languages in the universal. For Sanskrit, we are in the same situation as the `Korean-PUD`: our pattern did not allow us to find any occurrences of pre or postpositions.

Finally, the `Basque` is isolated because of a relatively low percentage of SOV at 57%, but sufficient to be considered the dominant order, the second order being SVO at 19%. WALS also considers Basque to be an SOV and postpositional language.

4 Discussing UD annotations

4.1 Nominal dependent relations

To determine the order of Subject, Object and Verb, we had to consider only nominal subjects and objects. In UD, these functions are annotated with the relations `nsubj` and `obj` respectively. We ran a first experiment without fixing the POS tags of the subject and object as noun, pronoun or proper noun, to avoid redundancy in the patterns. We obtained heterogeneous results, especially between corpora of the same language. On closer inspection of some corpora, we observed several times and in corpora of different languages that the dependent `nsubj` and `obj` relations were not nominal, which is inconsistent with the definitions of these relations.

With GREW, we computed the proportion of non-nominal dependents of the relations `nsubj` and `obj` linked to a verb. Out of the 141 corpora, six corpora present more than 20% of non-nominal `nsubj` dependents and four present more than 20% of non-nominal `obj` dependents. Tables 8 and 9 detail the relative proportions for each corpus as well as the most represented POS tags for these dependents. The non-nominal `nsubj` dependents of the three `PROIEL` corpora are mostly adjectives between 72% and 78%. The `Thai-PUD` and `Slovenian-SST` have a high proportion of determiners and out of the 25% of non-nominal `nsubj` dependents of the `Arabic-PADT`, 50% are annotated X, a label used when the other POS tags are not adapted.

As for the four corpora with more than 20% non-nominal `obj` dependents, the most frequent POS is verb for three corpora and adjective for the last one.

Corpus	Non-nominal `nsubj` dependents	VERB	ADJ	DET	X
Old_Church_Slavonic-PROIEL	27.06%	20.94%	**73.89%**	0.00%	0.00%
Thai-PUD	25.86%	14.40%	0.28%	**83.10%**	0.00%
Arabic-PADT	25.15%	0.34%	10.72%	31.88%	**50.34%**
Ancient_Greek-PROIEL	23.67%	17.55%	**78.55%**	0.00%	0.00%
Gothic-PROIEL	21.82%	22.42%	**72.17%**	0.00%	0.00%
Slovenian-SST	20.24%	0.00%	9.46%	**79.28%**	1.35%

Table 8: Corpora with the most non-nominal `nsubj` dependents and proportions of POS tags for these.

Corpus	Non-nominal `obj` dependents	VERB	ADJ
Turkish-IMST	32.80%	**65.72%**	26.03%
Hindi-HDTB	26.43%	**89.00%**	9.55%
Urdu-UDTB	24.87%	**88.83%**	8.14%
Ancient_Greek-PROIEL	20.96%	19.12%	**76.95%**

Table 9: Corpora with the most non-nominal `obj` dependents and proportions of POS tags for these.

4.2 The `case` relation

According to UD, the relation `case` allows to annotate any case-marking element which is treated as a separate syntactic word (including prepositions, postpositions, and clitic case markers). These elements are dependents on the nouns to which they are attached[10]. The relation `case` thus involves nominal governors. In the same way as for the relations `nsubj` and `obj`, we calculated the proportion of `case` relations with non-nominal governors in the 141 corpora. Table 10 shows the seven corpora which have more than 20% non-nominal governors involved in the `case` relation.

We explored the possible reasons for these results in Turkish, Chinese and Korean as we can compare the results between different corpora. We could not add the two mono-corpus languages, Basque and

[10] `https://universaldependencies.org/u/dep/case`, September 2021.

Corpus	Non-nominal `case` governors	VERB	NUM	ADJ	DET	PART
Turkish-BOUN	45.95%	**54.73%**	0.96%	17.46%	1.33%	0.00%
Turkish-IMST	37.49%	**55.40%**	6.89%	29.46%	3.51%	0.00%
Chinese-GSD	28.77%	**60.76%**	1.02%	6.31%	0.09%	27.27%
Basque-BDT	28.27%	0.00%	19.85%	24.95%	**38.94%**	0.00%
Amharic-ATT	22.47%	**75.41%**	1.64%	7.38%	5.74%	0.00%
Korean-Kaist	20.59%	19.51%	**56.50%**	5.69%	0.00%	0.00%

Table 10: Proportions and POS tags of non-nominal `case` governors linked to an adposition.

Amharic in our analysis, as there was no other corpus to compare them to.

Turkish The percentages of non-nominal governors are high for two Turkish corpora (45.95% and 37.49%) and more than half of them are verbs. In comparison, the other two Turkish corpora have low proportions of non-nominal governors: the `Turkish-GB` is at 6.23% and the `Turkish-PUD` is at 4.97%. Syntactic relations of the `Turkish-BOUN` were manually annotated in the UD scheme, while the `Turkish-IMST` is the result of a semi-automatic conversion of the IMST Treebank (Sulubacak et al., 2016). As the `Turkish-BOUN` is manually annotated by native speakers, we can assume that they made an annotation choice that is contradictory to the UD instructions. For the `Turkish-IMST`, this may be due to the semi-automatic conversion. In version 2.8 of UD, four new corpora have been added and it is indicated that updates have been made to gain consistency across all corpora in Turkish[11]. However, our results remain the same for the `Turkish-BOUN` and the `Turkish-IMST` in UD 2.8.

Chinese The `Chinese-GSD` have 28.77% of non-nominal governors in the relation `case`, mostly verbs (more then 60% of cases). Due to the Chinese language structure, the `case` relation definition is slightly different from the universal definition. `case` is used on particles marking relations such as genitive, prepositions including coverbs and valence markers[12].

The coverbs correspond to particles or adpositions linked to verbs such as: 在+ 落在 (fall on), 向+ 奔向 (run to). In `Chinese-GSD`, they are annotated as adpositions and are thus linked to verbs with the relation `case`. In the other two Chinese corpora, the `Chinese-PUD` and the `Chinese-HK`, the percentages of non-nominal governors are 2.29% and 5.22% respectively. In these corpora, adpositions and verbs are linked with the `mark` relation. In Chinese, it is used on a functional word marking a clause as subordinate to another clause[13]. In the universal definition, this relation involves a subordinating conjunction SCONJ rather than an adposition but Chinese corpora use both tags. For Chinese, we face two difficulties: i) the use of `case` and `mark` relations to annotate coverbs, ii) the conflict between the POS tags ADP, PART and SCONJ. As previously observed in Section 3.2, the `Chinese-GSD` differs from the other two corpora because of some inconsistent annotations with the UD guidelines.

Korean The `Korean-Kaist` has a percentage of non-nominal governors in the relation `case` at 20.59% and 56.50% of these governors are numbers. To annotate numbers followed by symbols, there is no real consensus between languages on the entity that should carry the syntactic relation. For example, in French corpora, the numeral and the symbol are linked by a relation `nummod` and the symbol is linked to the noun to which it is attached by a relation `nmod`. Although the UD guidelines state that the `case` relation involves nominal governors, other languages have `case` relations between an adposition and a numeral. In French, for example, `French-Sequoia` includes this construction in the form “en 1980”.

5 Conclusion

Our results are mostly consistent with Greenberg’s observations, but also with the information from WALS concerning the three orders. Our results constitute new typological information based on large

[11] `https://github.com/UniversalDependencies/UD_Turkish-BOUN`, September 2021.
[12] `https://universaldependencies.org/zh/dep/case`, September 2021.
[13] `https://universaldependencies.org/zh/dep/mark`, September 2021.

amounts of data that can fill in gaps in the existing databases. In particular, we treated seven languages for which values of the three orders are not provided in WALS: Afrikaans, Faroese, Galician, Kazakh, Maltese, Naija and Slovak. The corpus-based approach allows us to evaluate the consistency between corpora of the same language and show a great variation according to the corpus types: oral language, written language in newspapers, tweets, poetry, novels, grammars, etc.

Moreover, our study raises several issues related to UD, especially to the universality of its annotation scheme. UD being initially based on an annotation scheme created for English (de Marneffe et al., 2014), for languages with a different structure than English, adapting the annotations leads the creators of the corpora to make annotation choices that they have to justify and that are not necessarily consistent with other corpora in the language. We therefore end up with some inconsistencies between languages but also between corpora of the same language, inconsistencies for which we provided analyses either by examining the corpus documentation or by asking native speakers. The collaborative aspect of UD project allowed us to share our observations and thus contribute to UD corpora improvement.

It is worth noting that our experiments can be replicated and extended given that the tool GREW is available online[14] along with the UD corpora[15]. Similarly, the scripts and patterns can be found on a Gitlab repository[16].

References

Hee-Soo Choi, Bruno Guillaume, Karën Fort, and Guy Perrier. 2021. Investigating Dominant Word Order on Universal Dependencies with Graph Rewriting. In *Recent Advances in Natural Language Processing (RANLP2021)*, en ligne, Bulgarie, September.

Noam. Chomsky. 1982. *Some concepts and consequences of the theory of government and binding / Noam Chomsky*. MIT Press Cambridge, Mass.

Bernard Comrie. 1989. *Language universals and Typology: Syntax and Morphology*. University of Chicago Press.

William Croft. 2003. *Typology and Universals*. Cambridge University Press, New York.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592, Reykjavik, Islande, May. European Language Resources Association (ELRA).

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308, 07.

Tillmann Dönicke, Xiang Yu, and Jonas Kuhn. 2020. Real-valued logics for typological universals: Framework and application. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3990–4003, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

M. Dryer. 1992. The greenbergian word order correlations. *Language*, 68:138 – 81.

Matthew S. Dryer. 2013a. Order of adjective and noun. 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. 2013b. Order of adposition and noun phrase. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures 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, Suède, August. Uppsala University, Uppsala, Suède.

[14] https://grew.fr/

[15] https://universaldependencies.org/

[16] https://gitlab.inria.fr/ud-greenberg/udworkshop-2021

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

Kim Gerdes, Sylvain Kahane, and Xinying Chen. 2021. Typometrics from implicational to quantitative universals in word order typology. *Glossa: a journal of general linguistics*, 6, 02.

Joseph H. Greenberg. 1966. 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.

Bruno Guillaume. 2021. Graph matching and graph rewriting: GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–175, Online, April. Association for Computational Linguistics.

Natalia Levshina. 2019. Token-based typology and word order entropy: A study based on universal dependencies. *Linguistic Typology*, 23(3):533–572.

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.

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, Pékin, Chine, July. Association for Computational Linguistics.

Kartik Sharma, Kaivalya Swami, Aditya Shete, and Samar Husain. 2019. Can Greenbergian universals be induced from language networks? In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 25–37, Paris, France, August. Association for Computational Linguistics.

Umut Sulubacak, Memduh Gokirmak, Francis Tyers, Çağrı Çöltekin, Joakim Nivre, and Gülşen Eryiğit. 2016. Universal Dependencies for Turkish. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3444–3454, Osaka, Japon, December. The COLING 2016 Organizing Committee.