New Proposal of Greenberg's Universal 14 from Typometrics

Antoni Brosa-Rodríguez*, **Sylvain Kahane**†
*Universitat Rovira i Virgili
Avda. Catalunya 35 - 43002 (Tarragona, Spain)
antoni.brosa@urv.cat

†Université Paris Nanterre
UFR Philia - 92001 (Nanterre Cedex, France)
skahane@parisnanterre.fr

Abstract

In his Universal 14, Greenberg stated that the normal and dominant order in all world languages was to place the condition before the conclusion in conditional sentences. We take this claim to review it quantitatively and based on occurrences in real texts in more than 50 languages. We can see that Greenberg's proposal is correct but that it needs a reformulation to be true at all. We propose a quantitatively based and updated Universal 14, which gives a better account of the representation of the different languages analyzed and which is fulfilled in 100% of the cases (as opposed to Greenberg's 60% in our sample). In addition, we also analyze adverbial sentences. Once we obtain the occurrence data in their direction (before or after the main verb), we plot a new Universal in a typometrical way: 100% of the languages show a higher proportion of preceding conditional clauses than of adverbial clauses, regardless of their type or the direction preference for adverbial clauses. The relationship between the SOV type and a stricter initial conditional location is also proposed.

Keywords: Conditional, Universal Dependencies, Token-based Typology

1. Introduction

In this study, we propose an approach that combines Linguistic Typology and Computational Linguistics to obtain new data about Language Universals from a quantitative perspective, also known as Typometrics (Gerdes et al., 2021) or Token-based Typology (Levshina, 2019). Specifically, our objectives are:

- To review Greenberg's Universal 14 based on occurrences in real texts from different languages.
- To reformulate the Universal in quantitative terms.
- To propose a new Typometric Universal related to conditional sentences, the subject of Universal 14, as well as adverbial clauses.
- To describe the distribution order of conditional sentences and adverbial sentences in some languages never attested before.

As far as we know, this Universal has not been systematically and typologically reviewed in the literature in Computational and Quantitative Linguistics. Hence the interest in this study. However, given the novelty of the approach and the limitations of space, we will limit ourselves to represent an approximation to its study, without being able to cover all possible details of the phenomenon. Our study is based on Universal Dependencies (UD), with 150 languages and over 200 treebanks in its 2.11 version. This means that there are limitations to the labeling of some linguistic elements, especially in word morphology. This gap has a direct influence on languages of a typology with a tendency to be polysynthetic or similar, as the conditional information is contained within words that have not been detailed in depth. Therefore, this language type cannot be checked in this study, which is preliminary and can be considered a good starting point to point out interesting trends and the methodology we propose to investigate. These results may be complemented in the future with the languages not represented in UD and which usually coincide with the typology described above.

Once we know these limitations, we have conducted the research with the rest of the languages. To do so, we have formalised the conditionals with the Grew-Match syntax, which is the tool we have used to carry out the queries, and we have obtained the lexical information through PanLex.

2. State of the art

The linguist Joseph Greenberg (1963) completely transformed the discipline of Linguistic Typology with a paradigm shift from morphology to syntax (word order). His proposal is based on 45 Universals that he claimed based on 30 languages, either known to him or based on data extracted from grammars, that is, what is known as an armchair Linguistics perspective (Clark and Bangerter, 2004). Universal 14, which we want to address in

LREC-COLING 2024, pages 12217–12226

this paper, states:

> In conditional statements, the conditional clause precedes the conclusion as the normal order in all languages.

In other words, what Greenberg indicates is that when a language wants to express a conditional sentence, it usually places the conditional part before the conclusion part. That is, all languages prefer to express "*If you want, you can come to my house*" instead of "*You can come to my house if you want*".

Even if this Universal has not been analyzed in Quantitative Typology, we find other proposals that treat similar elements, which try to offer a much finer-grained approach to different linguistic phenomena, in line with what we propose here. One of the foundational and most relevant publications is probably that of Liu (2010), trying to understand the functioning of languages quantitatively concerning their predisposition to be head-initial or head-final. Other studies in this regard that can be highlighted are those of Guzmán Naranjo and Becker (2018) or Levshina (2019), although they usually focus more on the potential for explaining the order of the direct object in relation to the verb and even, in the case of Levshina (2019), to understand other metrics such as linguistic entropy.

In strict relation to the study of Greenberg's Universals in Quantitative Typology, we find some studies such as those of Brosa-Rodríguez and Jiménez López (2023), Choi et al. (2021a), Choi et al. (2021b) and Gerdes et al. (2019). In these, the proposals of Universals 1, 3, 4, 17, 19, and 25 are reviewed. That is, as we have announced, there is room for research on conditional structures in the field of Linguistic Typology and specifically about Greenberg's proposal thanks to new methods and data.

Of course, we do not want to assert that conditionals have not been extensively treated in Linguistics, as there are different approaches to the phenomenon, which we do not fully develop here due to lack of space, but which can be expanded in Liu (2019). Surely, the most prolific are the comparative studies between two languages and the different behavior of their conditionals, such as the example of Hammadi (2019) or Hasselgård (2014), regarding English and Arabic and English and Norwegian, respectively. The field where there is probably the greatest interest in the phenomenon is the cognitive one. In these types of studies, they try to analyze in a detailed way the different occurrences and cases that the manifestation of conditional structures entails to try to understand it through processing explanations. However, we must emphasize that these studies tend to be focused on a maximum of 2 or 3 languages and, therefore, it is difficult to extend these results to many other languages.

In Linguistic Typology, we find some attempts to treat the phenomenon from a cross-linguistic perspective. However, this is not an easy task. As Comrie (1986) points out, the fact that there can be both intralinguistic and cross-linguistic variation in the manifestations of conditionals is problematic. This makes it difficult to identify them completely in all languages and also to define them in a precise but extensible way. Following Comrie (1986) again, abstractly and logically, we can understand conditionals as a relation between two propositions, one which is the protasis (traditionally called *p*, the condition for Greenberg) and another which is the apodosis (traditionally called *q*, the conclusion for Greenberg). These can be both true, both false or *p* can be false and *q* can be true, from a logical point of view.

From a typological point of view, we can present the different manifestations we have to mark a conditional structure in languages in the following possibilities:

- That there is a marking in the protasis and there is no marking in the apodosis, or it is optional.
- That there is a mark in the apodosis and there is no mark in the protasis, or it is optional.
- That there is a mark on the protasis and a mark on the apodosis, both of which are mandatory.
- That there is no mark on either the protasis or the apodosis.

The universal trend is that languages manifest the condition through the first case, as in English, where we have a mandatory *if* element and an optional *then* element. However, we can find some exceptions, known as *rara* in typology in the other 3 cases. In the case of Mandarin, we can find some occurrences where only the conclusion of the condition is marked (Comrie, 1986), although there is also the marking of the condition only in other examples (Liu, 2019). The obligatory marking of both parts has been found only in New Guinea Pidgin, up to our knowledge (Comrie, 1986). On the other hand, Vietnamese and Mandarin, according to Comrie (1986), adhere to the last of the cases, where there is no mark. Olguín Martínez and Lester (2021) add to this last group the Imonda language. It is known that such a construction is conditional thanks to implicatures, context, and common knowledge.

Subsequently, we must review different ways languages manifest to mark this element in the protasis. According to Traugott et al. (1983), conditional

markers are most commonly particles, clitics, or affixes. In addition, they also state that the universal structure is that they appear within the clause already pointed out by Comrie (1986) before the condition.

Again, the most common will be using an independent word or particle. In the case of using clitics or affixes, these can appear either in initial position, as in Swahili (*ki-, ngeli-, ngali-, nge-*) (Nicolle, 2017), or in final position as in Rama (*-kata*) (Olguín Martínez and Lester, 2021). It is also possible to find that the condition mark is the inversion of the order in some cases of German, e.g., however, both German and Swahili and the vast majority of languages present a clear and separate word to mark the conditions of less marked cases, leaving order changes or suffixes for different conditional types (Comrie, 1986; Haiman, 1983).

In addition, we must warn that we are fully aware that there may be alternatives in each language to create conditional sentences using other structures, such as, for example, in English, the sentence: "*You drink one more beer and I think I'm leaving*". In this case, there is no direct equivalent element to 'if', but rather we find a coordinative conjunction that joins two sentences corresponding to slightly different temporal moments. In this case, the conditional structure is inferred. We will not address all these cases for three different reasons.

Firstly, Greenberg was always analyzing the more canonical cases, even though his Universals are formulated in such a general way that these structures could be included. If we want to review what Greenberg considered more faithfully, perhaps we should focus on the canonical structures.

Secondly, it is impossible to formalize the different non-canonical possibilities that each language has for creating conditional sentences. We do not have this information for many of the world's languages. Furthermore, it makes no sense to propose an ad hoc formalization for these cases when it would not be Universal.

Thirdly, according to Weisser (2019), the behavior of all these non-canonical cases can be encompassed and represented by the canonical cases within each language, as demonstrated by these adverbial conditional sentences.

In short, we have seen how the universal behavior is using a specific word located in the condition of the subordinate clause.

As we have mentioned, Universal Dependencies has bet on an analysis of syntactic elements to the detriment of morphological ones, which penalizes the level of detail that we can obtain in these types of languages. There is still a lack of a clear, agreed-upon, and Universal commitment to reflect all the detailed morphological information that allows this type of analysis. We observe a clear under-representation of (poly)synthetic languages in Universal Dependencies, probably due to the lack of consensus even today and the difficulties in its labeling. We believe, furthermore, that offering interglossae could help to homogenize all of this and make the comparability between languages of this type and others with a more analytical or inflectional profile more natural and feasible.

3. Methodology

Once we have observed the lack of computational data regarding conditional adverbial clauses in languages worldwide, we must detail how we propose to approach this phenomenon.

3.1. Sample, size, and source

To perform a quantitative analysis based on data extracted from real texts, it is reasonable to look for corpora that contain grammatical tagging of the object under analysis. Additionally, since we intend to compare linguistic structures in different languages worldwide, we must strive to facilitate comparability by using the same terminology and process of annotation. Therefore, this linguistic information will be extracted from Universal Dependencies 2.11 (Nivre et al., 2023).

The sample with which we work is a convenience sample (Miestamo et al., 2016). Although in typological studies, it may be recommended to create a variety sampling or probability sampling, we do not consider it possible or interesting given the current availability of languages in this resource. However, we believe that in the future, thanks to a greater presence of languages, it will be possible to approach both positions. Additionally, we believe it is not entirely advisable to exclude any language to maximize our goal of offering a typological characterization of conditionals in all possible languages worldwide (as in many of them, we do not have this information).

In any case, we believe that our results can be a good starting point (as there are no other similar studies with this object of study computationally, up to our knowledge) to understand these universal trends. In the future, the exact quantitative result may be modified, although we believe it will be only slightly, and, therefore, the claims we may propose will not change in general.

The initial selection of languages that it is possible to analyze consists of a fairly varied representation of the different basic linguistic types, as can be seen in the comparison that we establish between our percentages and those that exist in the WALS in table 1.

As can be seen, the representation of linguistic types is quite similar to the comparison element

Table 1: Sample Type Comparison with WALS

Type	Our Sample %	WALS' Sample %
SVO	58	41
SOV	21	35
VSO	7	7
VOS	0	2
OVS	1	0
OSV	0	1
NDO	12	14

(WALS), although there is a slight preference for SVO languages. These data are interesting because they allow us to point out a possible correlation between SOV-type languages (therefore, with a head-final tendency) and a greater rigidity in the initial placement of IF-clauses.

Regarding the genetic aspect, more than half of the languages (58%) are of the Indo-European family, one of the best-known biases in UD. Regarding the area, the great majority of languages are from Eurasia (88%), followed by Africa (6%), Papunesia (4%), North America (1%), and South America (1%). Only Australia is not represented (0%). These data, besides demonstrating the obvious bias of the UD database, show us that the data we can show here, although interesting, will have to be checked in the future with a better sample to be confirmed.

As is known, the corpus typology in Universal Dependencies is varied. In this case, the possible bias of some languages (especially those with few resources and, therefore, only one treebank) is questionable if the information comes only from literary or sacred texts, for example. However, we still believe that if we have no additional information, we must consider this data as the most suitable to understand the language, as in traditional Linguistic Typology when considering dead languages whose only preserved texts are of a very specific type (usually sacred). Therefore, we do not apply any type of textual typology restriction to obtain the maximum number of languages to analyze. For expanding this information, we recommend Levshina (2021).

In short, we rely on corpora, usually drawn mostly from wikis, news, fiction, legal texts, and blogs or reviews. As can be seen, the typology of texts is quite varied and interesting, although written content prevails over oral content, something that should be considered. We have carried out a check on whether the corpus typology influences the results obtained (in the case of languages with variety). We have not observed any clear and strong patterns which would allow us to indicate that there is a conditioning, neither of the frequency of occurrence of conditional structures in general nor of a greater occurrence of a certain order of occurrence. However, we would like to point out that corpora containing news or oral inputs tend to show more conditional structures than corpora based on wikis. The size of the treebanks is very different between languages, so we will show the results as a percentage, which allows for homogenization and comparison of results.

However, the significance test (the p-value) we have carried out is the most interesting measure to validate the reliability of the data we present and filter possible conclusions. In this, we check the size N necessary to accept as valid a given distribution of occurrences in a binomial law X. If we postulate a null hypothesis (H0), we can check if the individual results in each language are less than 0.05. This will mean that the sample of conditionals available to us is large enough to state that the results would have the same trend with any other sample and larger size. Once automatically calculated in Python, we can draw much more reliable conclusions. Of the 57 languages for which such a result can be analyzed in conditional constructions, 39 pass this test with a large margin, and 18 are slightly above the limit. The full data can be seen later in figure 7, where the case of languages we should isolate for this lack of confirmation is marked with a special color. However, all languages, whether they show that level below the formulated threshold or not, behave in the same way and contribute in the same way to the Universal. Therefore, there is no change in whether these languages are filtered or not.

3.2. Tools

This study, given the complexity of the linguistic structure we want to analyze, requires the use of two different tools. First, it is necessary to obtain the equivalent of 'if' in the different languages we want to analyze. Ideally, this information would be easily retrievable through interglossae in Universal Dependencies. However, this is not possible, as most Treebanks have not chosen to provide such information, leading us to search for this information externally. Therefore, we have compiled a list of all the languages we have Treebanks for and have used the PanLex tool (Kamholz et al., 2014), following a methodology already proposed in Brosa-Rodríguez and Jiménez-López (2023); Brosa-Rodríguez and Jiménez-López (2024) or McCarthy et al. (2020), all three studies on Universals, but in the lexicon, about the implicational hierarchy of the Berlin and Kay colors.

With this resource, we can obtain an equivalent for the meaning of 'if' in various languages. According to the creators, there is lexical documentation for 5,700 languages (although, obviously, in some cases, the variety of the present lexicon is limited). The special interest of PanLex in represent-

Table 2: Little Sample of Languages and their IF-Value

Language	IF-Value
Croatian	*ako*
Czech	*jestliže* | *jestli* | *pokud*
Danish	*hvis*
Dutch	*als* | *indien*

ing languages with few resources and traditionally marginalized is an important aspect when it comes to typologically representing a broad range of languages, and it is also interesting for the future, as if data availability scales, we will be able to continue working with this tool. Additionally, thanks to the design of this tool, we have been able to avoid two classical problems in lexical searches, such as homonymy and synonymy. Through a double translation process, we could verify if the form obtained in that language also corresponds to other elements that must be considered. However, we are not overly concerned about homonymy, as we require this 'if' to have a specific part of speech and be anchored in a very specific structure, which rules out other uses. In other words, the formalizations shown below allow us to manage this possible homonymy and filter out only the cases we are interested in.

On the other hand, we can also obtain percentages of similarity with other forms that may be equivalent to 'if,' that is, synonyms. Following the precepts of Gries (2013), if there is a similarity of more than 66.66% between two forms, we consider them to be synonymous, as in Polish with "jeśli" and "jeżeli". In summary, we can compile a list of all the words, a reduced version shown as an example in table 2.

Once we have all the usable forms, we must find a tool to analyze this annotated data in Universal Dependencies. The selected tool, due to its completeness and visual and metric facilities, is Grew-Match (Guillaume, 2021). To use it, we just need to formalize the structures we want to analyze using the tool's own syntax.

3.3. Formalization

To capture conditional adverbial structures with different words, we propose the following formalizations[1] 1 and 2:

```
%14-IF-before
pattern { IF [upos=SCONJ, lemma=if] ;
VIF [upos=VERB|AUX] ;
VMAIN [upos=VERB|AUX] ;
VIF -[mark]-> IF ; VMAIN -[advcl]-> VIF ;
IF<<VMAIN }
```
(1)

```
%14-IF-after
pattern { IF [upos=SCONJ, lemma=if] ;
VIF [upos=VERB|AUX] ;
VMAIN [upos=VERB|AUX] ;
VIF -[mark]-> IF; VMAIN-[advcl]-> VIF;
IF>>VMAIN }
```
(2)

In them, we have created a pattern where there is a word that corresponds to the equivalent of 'if' in each language and where we must manually replace the "if" part (corresponding to the English example) with the specific element of each language. This element depends on a verb or auxiliary that is the central element of that structure. The central element of this structure is dependent on another verb or auxiliary, functioning as an equivalent to an adverbial clause. In the first pattern, we require the condition to precede the conclusion, and in the second, the conclusion to precede the condition. To calculate the distribution of adverbial clauses in general in the languages, it is sufficient to remove from the presented formalization "$, lemma = if$".[2] This means that the formalizations we have analyzed above ("if") are also included in this more general query.

However, the formalization presented above only works for languages that mark the conditional by an isolated word or particle (the majority in the languages of the world and especially in UD). We also believe it is necessary to offer a solution to capture occurrences corresponding to clitics, prefixes or suffixes computationally; that is, elements joined in a word. In the case of Basque, for example, we would offer the formalization 3, where we look at a word the form of which begins by *ba*:

```
%14-IF-basque
pattern { GOV -[advcl]-> DEP ;
DEP [upos=VERB, form=re"ba.*"] }
```
(3)

The current limitations with UD, of different origins, do not allow a search like this, as will be detailed in the limitations section. This causes this preliminary study to focus on languages expressing conditions separately.

[1]The use of "whether" is discarded as it shows practically no occurrences in the analyzed corpora. This is because we have proposed a formalization to capture the structure of the conditions. This allows us to filter out cases such as "whether", where there is no condition behind it, as it usually appears in subordinate noun clauses. In the case that there is a shared word for both structures, if and whether (in Spanish or French, for example), the syntactic constraints in the rule we propose allow us to capture only the conditions and discard other options.

[2]However, another more convenient option is to use the clustering function by whether in Grew, as we show in the link https://Universal.grew.fr/?custom=64c66c5ee3027.

3.4. Calculation

As we have mentioned, there are different ways to calculate a Universal. Fundamentally, the calculations we propose respond to two different approaches. On the one hand, we will try to offer a more traditional review to analyze the degree of acceptability of Greenberg's original proposal based on categories based on quantitative data and real texts. On the other hand, we intend to reformulate the Universal in the style of Typometrics, that is, purely quantitative.

To work with Greenberg's Universal, we must translate the concept of "normal order" into quantitative terms, which is what we will obtain. Based on the proposal of Gries (2013), which relates Statistics to Linguistics, we understand basic or dominant order as an element with double or greater occurrence than its next competitor. In the case of a dichotomy, such as Universal 14, this translates into designating one of the two orders if its occurrence is above 66.66%. If, on the other hand, the occurrence is between 33.33% and 66.66%, we must understand that there is no tendency for a language to adopt either of the two orders as its normal order.

The review of this same Universal in quantitative terms should only consider which of the two orders has the highest occurrence. However, since we want to go further and formulate a two-dimensional Typometric Universal too, we must cross these quantitative data from the previous Universal with those of the calculation of adverbial clauses to see the internal proportion of each language and see if we can highlight invariances in the behavior of all of them and their distribution considering both variables at once.

4. Results

After conducting our study, we obtained interesting results with a clear trend, only broken by Norwegian. Therefore, we decided to discuss with a Norwegian linguist and concluded that there was an error in the annotation for this language. Such problems in corpus annotation can be expected in some cases, as is demonstrated in (Wisniewski and Yvon, 2019). Norwegian has two forms equivalent to 'if': *hvis* and *om*. There is no problem with the first one, but the second one has a case of homonymy, as it is also used as an adposition, equivalent to 'about' in English. With our methodology, we cannot discriminate this opposition due to an error in the treebank itself, as an automatic conversion to the UD annotation scheme was made, and all *om* forms were assigned the value of SCONJ, without distinguishing them with the ADP tag. Therefore, we removed the

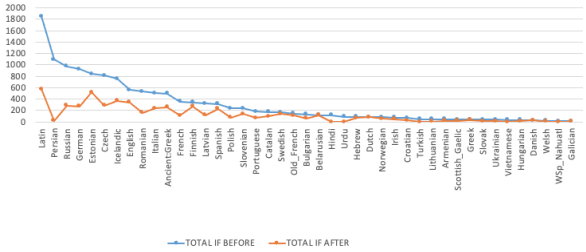

Figure 1: Occurrences of IF-Clauses before or after

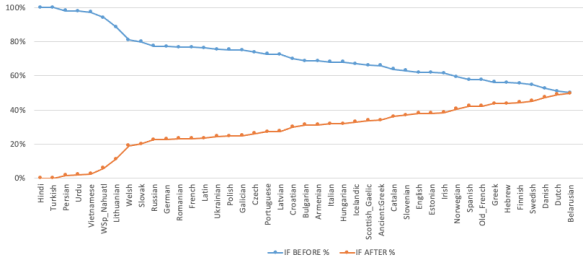

Figure 2: Percentage of IF-Clauses before or after

om lemma until this disparity is fixed and only use *hvis*.

Once this error was clarified, we can offer a graph showing the results in occurrences and percentages of the behavior of the languages regarding the location of the conditional clause in different real texts, as seen in figures 1 and 2. On the other hand, we can offer these two graphs but apply them in general to any adverbial construction, not to a specific conditional construction, as shown in figures 3 and 4. Later, we can cross-reference the data obtained in both percentage graphs to formulate a pure typometrics-style graph, with a 2D plot and the creation of two triangles that correspond to two clusters concerning a greater proportion of one of the two elements over the other in all languages, as seen in figure 5. In addition, we present the detailed data by languages in figure 6. Finally, we would like to provide a list of the obtained results, as they may be relevant for the typological characterization of some languages for which we may not have had this information before. Therefore, we present figures 7 and 8, where salmon color

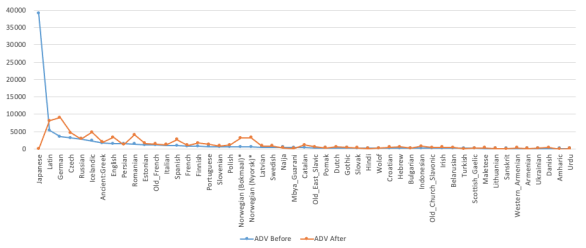

Figure 3: Occurrences of ADV-Clauses before or after

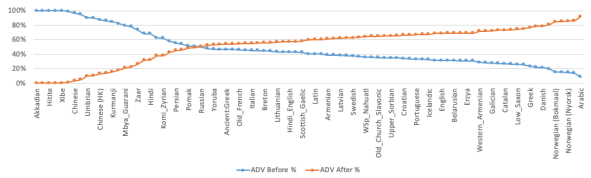

Figure 4: Percentage of ADV-Clauses before or after

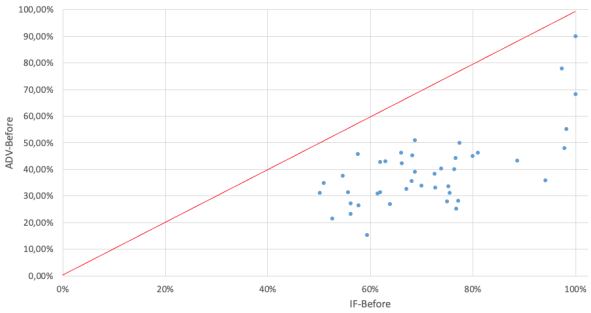

Figure 5: Typometrics 2D-Plot with Adverbial-Clauses and IF-Clauses

means that it has not passed the liability test.

5. Discussion

Firstly, if we want to verify the validity of Greenberg's original Universal, we must say that it does not hold true according to our data. Taking the threshold of 66.66% to determine the normal order of the antecedent 'if' element, around 60% of the languages fulfill the proposed Universal, while 40% do not, which does not seem strong enough to be considered a representative trend.

However, through a label-free approach (more quantitative), we can find a Universal with greater applicability. That is, if we reformulate Greenberg's Universal claim that "In all languages, in conditional statements, there are always more conditional clauses preceding conclusions than vice versa," we find a validity of 100%. In this case, we are not assuming that the fact of placing the if element first is what always occurs in languages (leaving the postposition as something marginal and marked), but we recognize the trend (probably justified by processing preferences) that exists to place the condition before the conclusion in all

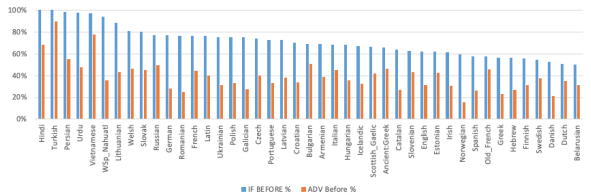

Figure 6: Typometrics Data by Language

Languages	% IF Before	% IF After	p-value	Languages	% IF Before	% IF After	p-value
Bengali	100.00	0.00	0.5	Manx	71.43	28.57	0.226
Kazakh	100.00	0.00	0.5	Croatian	70.00	30.00	0
Uyghur	100.00	0.00	0.5	Bulgarian	68.72	31.28	0
Chinese	100.00	0.00	0.25	Armenian	68.66	31.34	0.001
Sanskrit	100.00	0.00	0.25	Italian	68.16	31.84	0
Kurmanji	100.00	0.00	0.125	Hungarian	68.09	31.91	0.009
Gothic	100.00	0.00	0.031	Icelandic	67.11	32.89	0
Hindi	100.00	0.00	0	Erzya	66.67	33.33	0.253
Turkish	100.00	0.00	0	Scottish_Gaelic	66.15	33.85	0.006
Persian	98.21	1.79	0	Ancient_Greek	66.01	33.99	0
Urdu	97.83	2.17	0	Catalan	63.87	36.13	0
Vietnamese	97.37	2.63	0	Slovenian	62.96	37.04	0
WSp_Nahuatl	94.12	5.88	0	Upper_Sorbian	62.50	37.50	0.362
Lithuanian	88.68	11.32	0	English	61.96	38.04	0
Welsh	80.95	19.05	0.003	Estonian	61.92	38.08	0
Slovak	80.00	20.00	0	Irish	61.48	38.52	0.007
Russian	77.39	22.61	0	Indonesian	60.00	40.00	0.376
German	77.17	22.83	0	Norwegian*	59.42	40.58	0.016
Romanian	76.76	23.24	0	Spanish	57.70	42.30	0
French	76.66	23.34	0	Old_French	57.65	42.35	0.008
Latin	76.42	23.58	0	Greek	56.16	43.84	0.174
Ukrainian	75.47	24.53	0	Hebrew	56.13	43.87	0.073
Polish	75.24	24.76	0	Finnish	55.65	44.35	0.002
Galician	75.00	25.00	0.038	Swedish	54.67	45.33	0.059
Czech	73.77	26.23	0	Danish	52.63	47.37	0.395
Portuguese	72.65	27.35	0	Dutch	50.89	49.11	0.438
Latvian	72.60	27.40	0	Belarusian	50.21	49.79	0.5
Breton	71.43	28.57	0.226	North_Sami	50.00	50.00	0.687

Figure 7: Quantitative Value of IF Position in Each Language

Languages	ADV Before	ADV After	Languages	ADV Before	ADV After
Akkadian	100%	0.00%	Latin	40.01%	59.99%
Guajajara	100.00%	0.00%	Nheengatu	40.00%	60.00%
Hitite	100.00%	0.00%	Armenian	38.98%	61.02%
Japanese	100.00%	0.00%	Gothic	38.63%	61.37%
Xibe	100.00%	0.00%	Latvian	38.15%	61.85%
Amharic	99.09%	0.91%	Old_East_Slavic	37.66%	62.34%
Chinese	96.67%	3.33%	Swedish	37.61%	62.39%
Beja	95.00%	5.00%	Maletese	36.67%	63.33%
Umbrian	90.00%	10.00%	WSp_Nahuatl	35.71%	64.29%
Turkish	89.82%	10.18%	Hungarian	35.54%	64.46%
Chinese (HK)	87.23%	12.77%	Old_Church_Slavonic	35.35%	64.65%
Cantonese (HK)	86.08%	13.92%	Dutch	34.88%	65.12%
Kurmanji	84.62%	15.38%	Upper_Sorbian	34.78%	65.22%
Sanskrit	82.07%	17.93%	Faroese	34.62%	65.38%
Mbya_Guarani	79.17%	20.83%	Croatian	33.69%	66.31%
Vietnamese	77.78%	22.22%	Polish	33.51%	66.49%
Zaar	73.53%	26.47%	Portuguese	33.01%	66.99%
Bambara	68.21%	31.79%	Turkish_German	32.88%	67.12%
Hindi	68.04%	31.96%	Icelandic	32.56%	67.44%
Naija	62.17%	37.83%	Finnish	31.38%	68.62%
Komi_Zyrian	61.90%	38.10%	English	31.34%	68.66%
Wolof	57.82%	42.18%	Ukrainian	31.15%	68.85%
Persian	55.06%	44.94%	Belarusian	31.06%	68.94%
Kiche	54.00%	46.00%	Serbian	30.84%	69.16%
Pomak	51.21%	48.79%	Erzya	30.77%	69.23%
Bulgarian	50.77%	49.23%	Irish	30.75%	69.25%
Russian	49.78%	50.22%	Western_Armenian	28.33%	71.67%
Urdu	47.79%	52.21%	German	28.10%	71.90%
Yoruba	46.81%	53.19%	Galician	27.78%	72.22%
Manx	46.51%	53.49%	Hebrew	27.06%	72.94%
Ancient:Greek	46.22%	53.78%	Catalan	26.94%	73.06%
Welsh	46.15%	53.85%	Spanish	26.30%	73.70%
Old_French	45.70%	54.30%	Low_Saxon	25.81%	74.19%
Ligurian	45.24%	54.76%	Romanian	25.19%	74.81%
Italian	45.15%	54.85%	Greek	23.17%	76.83%
Slovak	44.89%	55.11%	Indonesian	21.70%	78.30%
Breton	44.44%	55.56%	Danish	21.47%	78.53%
French	44.20%	55.80%	Ancient_Hebrew	19.40%	80.60%
Lithuanian	43.31%	56.69%	Norwegian (Bokmaal)*	15.35%	84.65%
Slovenian	42.99%	57.01%	North_Sami	15.09%	84.91%
Hindi_English	42.79%	57.21%	Norwegian (Nyorsk)*	14.65%	85.35%
Estonian	42.78%	57.22%	Coptic	14.03%	85.97%
Scottish_Gaelic	42.23%	57.77%	Arabic	8.37%	91.63%
Czech	40.30%	59.70%			

Figure 8: Quantitative Value of ADV Position in Each Language

languages of the world, without hiding the fact that in many languages, the opposite mechanism is equally habitual, feasible, or natural.

In the left part of figure 2, where the languages with a more extreme percentage of "if" before the head appear, there is a tendency for SOV lan-

guages to appear. Therefore, although we cannot develop a stronger theory because it is a pioneering approach and due to lack of data, we would like to point out this tendency as a possible correlation between strict antecedent conditional element placement and the SOV linguistic type, while there seems to be greater flexibility in SVO languages. This fact would make sense insofar as this linguistic type always shows rather inflexible behaviors, such as, for example, a low tolerance for prepositions, while some occurrences of postpositions can be observed in SVO languages.

It is also interesting to highlight the group of languages whose results are closest to 50%, as in many cases, they are varieties close to Norwegian either through area or genus. This leads us to think that perhaps in these languages, there is also an alternative use of the elements used for 'if' for different constructions (such as concessive, for example). However, this use would be minor, in any case, and does not affect the Universal. It remains to be seen when more data is available if this is a possible explanation.

In the case of adverbial clauses, it can be observed that, unlike what happens with conditional clauses, no pattern can be established. As seen in figure 4, there is a group of languages clearly with the first position adverbial structures and another group where the adverbial structures are after the head. It should be noted that in this case, the trend is for more languages with a postposition of this type of structure, i.e., the opposite location to what happens with conditional clauses, which is quite striking. We should also highlight that, in most languages, although the percentage of postpositions of adverbial clauses is higher, there is a balanced number between both orders (close to 50%). This behavior makes sense if we understand that many of these structures are not arguments; therefore, there is not such a clear prefixed order as there might be with arguments.

In figure 5, on the other hand, we see what we believe is the most interesting, complete, and alternative prediction for approaching the phenomenon of analyzed conditional clauses. Firstly, we can see how all languages are in the right triangle. That is, no matter how many initial or final conditional structures or adverbial structures, in general, each language has, they will always exhibit a higher proportion of conditionals preceding than adverbials doing so. This seems to indicate a push in the world's languages to enjoy relative flexibility in locating adverbial clauses but greater restriction when locating (preceding) conditional clauses, being this proportionate. It is also possible to highlight the cluster formed with many languages sharing space (indicating that this behavior is Universal and typical), with some languages outside this cluster and at much higher values of "if" preceding. Apart from the fact that the left half of the graph is empty (no occurrence of preceding 'if' less than 50%), we can also highlight the gap in the lower right corner. It seems that when a language shows a very high proportion of conditional clauses preceding, these tend to go towards a higher proportion of adverbial clauses in the first position (as would be expected). In other words, we can witness extreme behaviors of conditional or adverbial clauses before, while there are no examples of the opposite in conditional or adverbial clauses.

6. Conclusion

In conclusion, we can point out that, on the one hand, according to our data, Greenberg did not formulate his Universal 14 quite accurately, leading to inadequacies in its application (60% of accuracy). Therefore, we recommend abandoning the concept of "normal order" as controversial and confusing in this Universal and propose moving towards a purely quantitative proposal: all languages tend to place conditions before conclusions more frequently in conditional clauses. Without mentioning further restrictions, this indicates more preceding conditionals than after the head. Furthermore, thanks to the comparison with the results obtained from adverbial clauses in general, we can point out that, despite the possible imprecision of Greenberg due to the use of the mentioned labels, he was right to propose a preference of languages for placing the condition in the first position. Although this order is not frequent in many languages, and there are cases of almost a tie in many of them, we see a more restricted behavior of languages in general than in adverbial clauses, which seems indicative of this tendency. In other words, all languages exhibit more conditional clauses preceding than adverbial clauses preceding, without exceptions. We should recall that we recommend further study of the links between the SOV type and constrictions with IF or adverbial sentences. We observed a more extreme behavior than with other linguistic types, and it seems coherent and consistent with the precedence of both structures. This may be due to functional processing explanations emphasized in these languages, which we want to investigate further.

Finally, we must remember that we also offer specific data for all the analyzed languages (for conditional constructions and adverbial clauses). This information we believe may be interesting for scholars focused on individual languages, families or types.

Therefore, we hope that in the future, thanks to greater availability of data (more treebanks and more languages) and closer collaboration between

Linguistic Typology and Computational Linguistics, we can better represent and integrate in detail this phenomenon that we already know in a general way but extended to more languages. Moreover, we believe that the methodology we propose and, at the same time, its limitations open up new avenues of research. On the one hand, we can extend the methodology developed in this paper to other universals and even propose new typometric universals in this way (or with new metrics like checking dependency length instead of dependency direction). On the other hand, we consider it necessary to continue working and deepening the aspect of conditionals. We must work with other databases or models to be able to characterize languages that are poorly represented in UD.

7. Acknowledgements

We thank Eskil Blaaflat Mundal for his willingness to discuss the controversial cases/errors in Norwegian UD Treebanks. We would like to thank Félix Kahane for his willingness to help when computing the values of the significance of our hypothesis (p-value).

8. Ethical considerations and limitations

Our study does not raise any ethical considerations that should be mentioned.

However, this pioneering study has a few intrinsic limitations. Beyond the presupposed limits inherent in typological studies, and especially the availability of data that marks limitations in sampling and textual typology previously mentioned (written text, too), we must highlight the limitation to adverbial conditional structures.

The starting bias is in the languages available in Universal Dependencies. There is a clear underrepresentation of agglutinative or polysynthetic languages and the like in this resource. This may be due precisely to the labeling proposal carried out in the different languages, always starting from the word as the basic unit. In the case of languages with affixes, it is difficult to disambiguate when it corresponds to "condition" and to isolate it from other occurrences that coincide in form. We need the information via interglossae, and that is not currently possible.

9. Bibliographical References

References

Antoni Brosa-Rodríguez and M. Dolores Jiménez-López. 2023. La jerarquía implicacional de Berlin y Kay a partir de frecuencias. *E-AESLA*, 8:1–17.

Antoni Brosa-Rodríguez and M. Dolores Jiménez López. 2023. A Typometrical Study of Greenberg's Linguistic Universal 1. In *Distributed Computing and Artificial Intelligence. Lecture Notes in Networks and Systems*, pages 186–196. Springer.

Antoni Brosa-Rodríguez and M. Dolores Jiménez-López. 2024. Quantifying basic colors' salience fromcross-linguistic corpora. *Color Research and Application*, 1(49):34–50.

Hee Soo Choi, Bruno Guillaume, and Karën Fort. 2021a. Corpus-based Language Universals Analysis using Universal Dependencies. In *ACL Anthology*, pages 1–15.

Hee Soo Choi, Bruno Guillaume, Karën Fort, and Guy Perrier. 2021b. Investigating Dominant Word Order on Universal Dependencies with Graph Rewriting. In *International Conference Recent Advances in Natural Language Processing, RANLP*, pages 281–290. Incoma Ltd.

Herbert Clark and Adrian Bangerter. 2004. Changing Ideas about Reference. In I. A. Noveck and D. Sperber, editors, *Experimental Pragmatics*, pages 25–49. Palgrave Macmillan, Houndmills.

Bernard Comrie. 1986. Conditionals: a typology. In Elizabeth Traugott, Alice ter Meulen, Judy Reilly, and Ferguson Charles, editors, *On Conditionals*, pages 77–99. Cambridge University Press, Cambridge.

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.

Kim Gerdes, Sylvain Kahane, and Xinying Chen. 2021. Typometrics: From implicational to quantitative universals in word order typology. *Glossa*, 6(1).

Joseph H. Greenberg. 1963. *Universals of Language*. The M.I.T. Press, Cambridge, Massachusetts.

Stefan Thomas Gries. 2013. *Statistics for Linguistics with R : a Practical Introduction*. De Gruyer Mouton, Berlin/Boston.

Bruno Guillaume. 2021. Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion. In *EACL 2021 - 16th conference of the European Chapter of the Association for Computational Linguistics*, pages 1–9.

Matías Guzmán Naranjo and Laura Becker. 2018. Quantitative word order typology with UD. In *17th International Workshop on Treebanks and Linguistic*, pages 91–104.

John Haiman. 1983. Paratactic IF-Clauses. *Journal of Pragmatics*, (7):263–281.

Samar Sami Hammadi. 2019. Arabic and English Conditional Clauses: A Comparative Study. In *The Eurasia Proceedings of Educational & Social Sciences*, volume 13, pages 109–114.

Hilde Hasselgård. 2014. Conditional clauses in English and Norwegian. In H. P. Helland and C. M. Meklenborg Salvessen, editors, *Affaire(s) de grammaire. Mélanges offterts à Marianne Hobæk Haff à l'occasion de ses soixiante-cinq ans.*, pages 185–204. Novus, Italy.

David Kamholz, Jonathan Pool, and Susan M. Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, pages 3145–3150.

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

Natalia Levshina. 2021. Corpus-based typology: Applications, challenges and some solutions. *Linguistic Typology*, 26:129–160.

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

Mingya Liu. 2019. Current issues in conditionals. *Linguistics Vanguard*, 5(3):263–281.

Arya D. McCarthy, Winston Wu, Aaron Mueller, Bill Watson, and David Yarowsky. 2020. Modeling color terminology across thousands of languages. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 2241–2250.

Matti Miestamo, Dik Bakker, and Antti Arppe. 2016. Sampling for variety. *Linguistic Typology*, 20(2):233–296.

Steve Nicolle. 2017. Conditional constructions in african languages. *Studies in African Linguistics*, (46):1–15.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2023. Universal Dependencies.

Jesús Olguín Martínez and Nicholas Lester. 2021. A quantitative analysis of counterfactual conditionals in cross-linguistic perspective. *Italian Journal of Linguistics*, 2(33):147–182.

Elizabeth Traugott, Alice ter Meulen, Judy Reilly, and Ferguson Charles. 1983. *On Conditionals*. Cambridge University Press, Cambridge.

Philipp Weisser. 2019. Equal rights for all conditionals. *Linguistics Vanguard*, 5(3):1–10.

Guillaume Wisniewski and François Yvon. 2019. How Bad are PoS Tagger in Cross-Corpora Settings? Evaluating Annotation Divergence in the UD Project. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 218–227, Minneapolis, Minnesota. Association for Computational Linguistics.