Glossa
a journal of general linguistics

Typometrics: From Implicational to Quantitative Universals in Word Order Typology

RESEARCH

KIM GERDES
SYLVAIN KAHANE
XINYING CHEN
Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

This paper develops the concept of word order universals based on a data analysis of the Universal Dependencies project, which proposes treebanks of more than 90 languages encoded with the same annotation scheme. The nature of the data we work on allows us to extract rich details for testing well-known typological implicational universals and, further, explore new kinds of universals that we call quantitative universals. We show how such quantitative universals are in essence different from implicational universals, including statistical universals, by the fact that they no longer lay down any claims on categorical statements, but rather on continuous parameters, opening a new field of research we propose to call *typometrics*.

CORRESPONDING AUTHOR:
Xinying Chen
Xi'an Jiaotong University, School of Foreign Studies, Xianning West Road 28, CN – 710049 Xi'an, Shaanxi, China
xy@yuyanxue.net

KEYWORDS:
Typology; Syntax; treebanks; computational typology; word order; statistical universals

TO CITE THIS ARTICLE:
Gerdes, Kim, Sylvain Kahane and Xinying Chen. 2021. Typometrics: From Implicational to Quantitative Universals in Word Order Typology. *Glossa: a journal of general linguistics* 6(1): 17. 1–31. DOI: *https://doi.org/10.5334/gjgl.764*

1 INTRODUCTION

Modern research in the field of language typology (Croft 2002; Song 2001), mostly based on Greenberg (1963), focuses less on lexical similarity and relies rather on various structural linguistic indices for language classification and generally puts much emphasis on the syntactic word order of some grammatical relations in a sentence (Haspelmath et al. 2005). Considered as the founder of word order typology, Greenberg (1963) proposed 45 linguistic universals and 28 of them refer to the relative position of syntactic units, such as the linear relative order of subject, object, and verb in a sentence. A more empirical way of examining word order typologies, testing correlations between two binary grammatical relations such as OV vs. VO and SV vs. VS, can be found in Dryer (1992) (following Lehmann 1973), in which some detailed word order correlations based on a sample of 625 languages are reported.

Along with the development of Corpus Linguistics and driven by the boost of Natural Language Processing, treebanks with fine-grained syntactic annotations have been developed for various languages. They allow for corpus-based methods applied to the study of diverse syntactic phenomena. Treebanks of different languages based on a similar annotation scheme provide direct access to measures of basic word order phenomena, an essential starting point for any linguist working on comparative language studies. With the appearance of larger sets of treebanks, research has begun to test existing word order typology claims or hypotheses based on treebank data. Investigating treebanks of 20 languages, Liu (2010) tested the 'traditional' typological claims with the subject-verb, object-verb, and adjective-noun data extracted from the treebanks, with coherent results, also showing that these 20 languages can be arranged on a continuum with absolute head-initial and head-final patterns as the two ends. Liu further states that treebank-based methods will be able to provide more complete and fine-grained typological analyses, while previous methods usually had to settle for a focus on basic word order phenomena (Hawkins 1983, Mithun 1987).

It is noteworthy that the field of word order typology has a strong empirical tradition, working with data and trying to describe the data with great precision. From a perspective of data analysis, new language data is emerging every day in this so-called era of 'big data'. It has never been a better moment than today to challenge, test, and corroborate existing ideas based on better and bigger data. The Universal Dependencies project (UD, Nivre et al. 2016), the basis of the present study, has seen a rapid growth into its present ample size with more than 100 treebanks of 72 different languages.

Such newly available syntactic treebanks in a wide range of languages can put typological and comparative studies on a new empirical base. These new resources allow reviewing and verifying well-known typological claims based on annotations of authentic texts (Liu et al. 2009, Liu 2010, Futrell et al. 2015).[1] They also allow developing new types of typological universals that are based on and require numerical empirical data. In a perspective of describing numerical data, methods of quantitative analysis are under constant development, in particular, which concerns us here, visualization techniques. The new typological patterns that this paper presents require a compilation of massive amounts of data into diagrams. Thus, the type of research presented here is highly dependent not only on the data itself but also on data processing and visualization technology, that enables typologists to move from empirical, data-based methods to actual data-driven research.

Following these ideas, this paper has a double objective: Based on the data analysis of a set of uniformly annotated texts in diverse languages, we first test well-known existing word-order universals and, secondly, explore how these universals can be embedded and conceived as a special case of more general empirical universals that we propose to call **quantitative universals** of human languages because of their inherently quantitative character. By means of concrete examples, we show how such quantitative universals differ from the classical implicational universals, including statistical universals, and provide new insights on word order typology thus opening a new field of research we propose to call **typometrics**. Even though the

1 The development of treebanks is cumbersome work. Even 75 languages only cover a modest segment of the world's languages. Another direction investigated in Östling (2015) is the use of parallel texts, the available translations of the New Testament in 986 languages. Such methods are not the subject of our paper, but they are certainly worth considering for future works, bearing in mind that translations contain some bias and are not fully representative of the target language (especially when the source text belongs to a marked genre such as religious texts).

set of languages of UD is currently not well-balanced in terms of language diversity (half of the languages of the database are Indo-European languages and non-Indoeuropean treebanks are often too small to be taken into account for some measures; cf. Bell 1978, Perkins 1989, 2001, Dryer 1989, 1992, Croft 1991, Whaley 1996, Cysow 2003, Dik 2010 on language sampling), and the results will have to be confirmed in the future on an even wider collection of languages, this resource allows us to have a new take on the question of language universals.

In this paper, we will look at one and two-dimensional diagrams such as ***Figures 1*** and ***2***, and we will show how these diagrams allow for a typological interpretation as a new type of syntactic universals: Quantitative Universals.

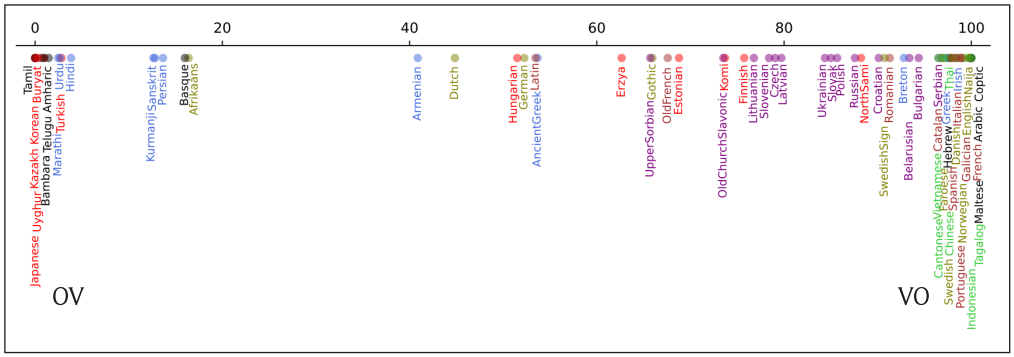

Figure 1 Percentage of VO, that is of nominal object (O) on the right of the verb (V). On the left of the graph, we see OV languages and on the right are the VO languages.

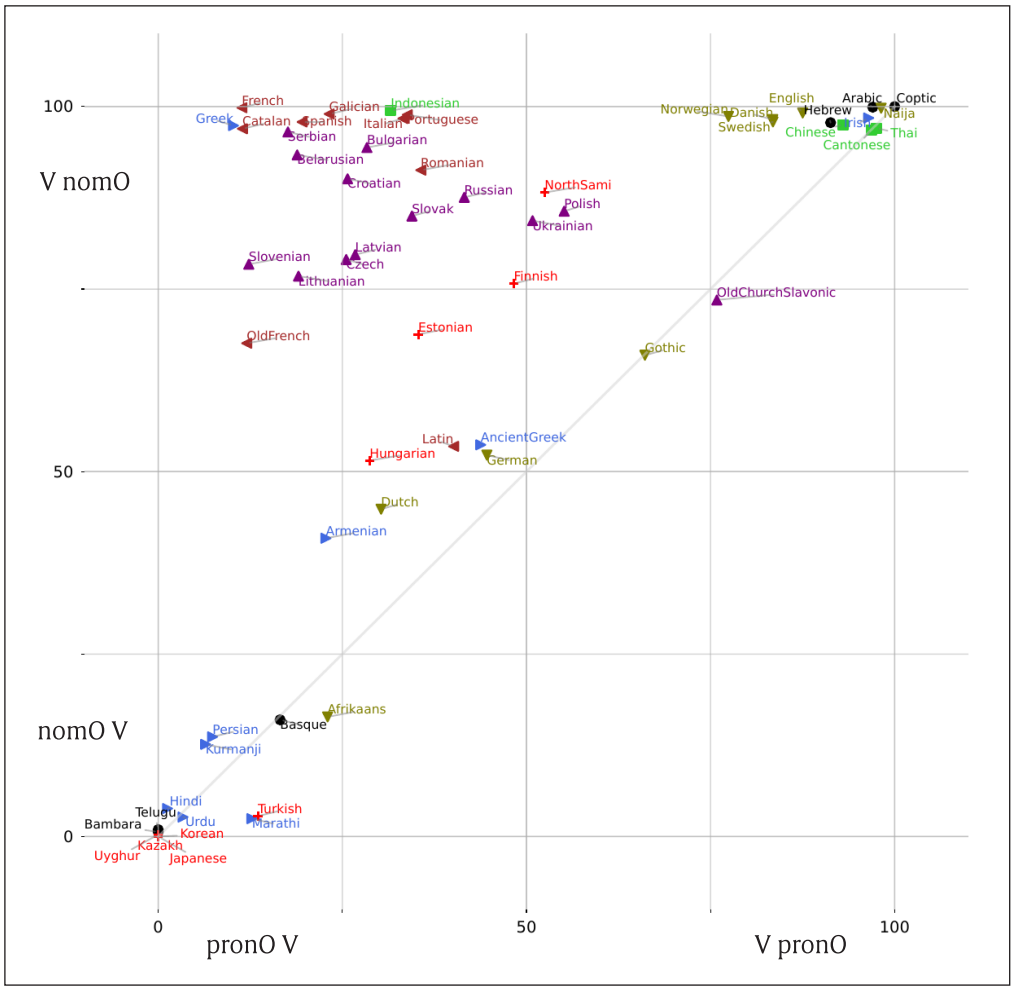

Figure 2 Scatter plot of the percentage of V pronO compared to V nomO.

Figure 1 shows the distributions[2] of languages across the following measurement: for each language, we counted how many times object dependency relations between a verb and a noun go from a left governor to a right dependent, and then compared this to the total number

2 We use the term *distribution* in the sense of a probability distribution. Our graphics show percentages of observed relations, which can be taken as estimates of probabilities.

of object dependency relations between a verb and a noun in a dependency treebank (following the idea developed by Liu et al. 2009).

For instance, the position of Arabic in ***Figure 1*** indicates that Arabic nominal objects are always on the right of their governor.[3] In the German treebank, 53% of all dependency relations between a verb and its nominal object go from left to right. Plotting this information on a scale from 0% to 100%, we obtain the beige dot representing German at 53%[4] in ***Figure 8***, thus comforting the view of German as a mixed word order language.

Traditional studies in typology are based on a categorical classification of languages in OV languages, VO languages, and languages without a dominant order. Such a classification can be problematic as it supposes the introduction of a threshold beyond which we consider an order to be dominant (Mithun 1987), which oversimplifies the data. We think that there are many advantages to working with quantitative data, which we attempt to show in this paper.

In ***Figure 2***, we visualize the average number of pronominal objects (pronO) and nominal objects (nomO) on the right of the verb (V), on a two-dimensional diagram. Our data shows for example that 100% of nominal objects in French are on the right of the governor. Inversely, only 7% of the pronominal objects are on the right, reflecting the low frequency of constructions like *je mange ça* 'I eat that' or *prends-le* 'take it' compared to *je le vois* 'I see it', lit. I it see. In this way, we can plot the French triangle at the point (7%, 100%) i.e. in the top left corner of our graph. By taking the same measures on all treebanks, we obtain the scatter plot.

We also grouped languages by rough language classes:

- Indo-European languages: triangles
- Indo-European-Romance: brown ◀
  - Indo-European-Baltoslavic: purple ▲
  - Indo-European-Germanic, including the English Creole *Naija*: olive ▼
  - Other Indo-European: blue ▶
- Sino-Austronesian: green squares ■
- Agglutinating languages: red plus signs +
- Other languages (Afroasiatic and Dravidian languages as well as Basque): black circles ●
- Some language points are hidden because the available treebank data for the language is not sufficient to provide significant measurements; more specifically, we decided to eliminate every language with less than 50 occurrences of one of the two compared types of relations.

In spite of the typological imbalance of the set of languages, we notice a remarkable shape of the scatter cloud, which is in fact the approximate top left triangle of the plot. Languages on the diagonal have the same percentage of pronominal and nominal objects to the right of their governor, while languages above the diagonal have more nominal objects than pronominal objects on the right of their governor. In other words, the triangle shape of the scatter cloud indicates a strong tendency among the studied languages to have their nominal object more frequently on the right of their governor than their pronominal objects.

In Section 2, we present a first example of a quantitative universal and compare it with absolute or statistical implicational universals.

3 We have chosen to indicate the number of dependencies that go to the right (head-initial relations), which means that the position of a language in the diagram corresponds to the relative position of dependents in this language. We could have chosen to indicate the number of dependencies that go to the left and the position of a language would have been interpreted as the relative position of heads in this language. This latter choice might seem more appropriate, especially for linguists thinking in terms of phrase structure and head positions, because in this case Arabic would appear on the very left of the graph as it is a head-initial language. Nevertheless, our choice to favor the position of the dependent is motivated by our goal to compare the relative position of different dependents of the same head (Figure 2) and our choice will give a more natural interpretation of the two-dimensional diagrams in terms of universals.

4 The transformed treebanks and transformation grammars are available on *https://surfacesyntacticud.github.io/*. Further scripts and precise numerical results are freely available on *https://github.com/typometrics*.

In Section 3, we introduce dependency treebanks, and we discuss how such scatter plots can be obtained from the Universal Dependencies treebanks. We explain amendments of the current annotation scheme that were necessary to obtain typologically relevant data.

In Section 4, we define Typometrics and propose some typometrical studies concerning word order, discussing in particular the traditional language classifications of Tesnière (1959). The section ends with a data-based analysis of the notions of free and mixed word-order.

We then study in Section 5 the interpretation of one-dimensional scatter plots as Greenbergian Universals (Greenberg 1963).

Section 6 is devoted to the classical Subject-Verb-Object versus Subject-Object-Verb classes of languages as seen through the lens of two-dimensional scatter plots.

Further two-dimensional language distribution patterns are examined in Section 7, in particular the interpretation of language distributions with low correlation coefficients.

In the conclusion, we discuss the implications of such methods and tools on studies in syntactic typology as a whole.

2 QUANTITATIVE UNIVERSALS

We will start our discussion based on the scatter plots of ***Figure 2***. In Section 2.1 we will compare this diagram with Greenberg's Universal 25 and see how it can be reformulated in terms of quantitative universals. In Section 2.2, we will see how Universal 25 can be generalized, giving us a new view on quantitative universals. Section 2.3 provides a comparison of qualitative and quantitative universals.

2.1 FROM AN IMPLICATIONAL UNIVERSAL TO QUANTITATIVE UNIVERSALS

The scatter plot of ***Figure 2*** is related with Universal 25 proposed by Greenberg (1963: 91):

> "Universal 25. If the pronominal object follows the verb, so does the nominal object."

Universal 25 is also what we will call a *qualitative* or *categorical universal*: A universal referring to a qualitative absolute property such as the "basic word order" of a language, and not to a numerical threshold. It supposes that we can categorize languages into languages where "the pronominal object follows the verb" and languages where "the pronominal object does not follow the verb", as well as languages where "the nominal object follows the verb" and languages where "the nominal object does not follow the verb".

Universal 25 is an *implicational universal*, because it has the form of an implication between two statements: "the pronominal object follows the verb" (V pronO) and "the nominal object follows the verb" (V nomO). Universal 25 can be abbreviated as V pronO → V nomO.

Universal 25 is an absolute universal: This statement is true for (nearly) all languages.[5] Absolute universals are opposed to statistical universals, i.e. a universal which holds for more languages than would be expected by a random distribution of the considered language property and, more generally, which is true for a significant percentage of languages. Greenberg's Universal 4 is an example of a statistical universal:

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

Note that this statistical universal is still qualitative (or categorical): it supposes that we can categorize languages in "languages with normal SOV order" and in "postpositional languages".

Let us now examine how Universal 25 is related to the scatter plot in ***Figure 2***. We can observe that Greenberg's statement is not totally clear. What does it actually mean that "the pronominal object follows the verb"? Does it mean that pronominal objects *always* follow the verb or does it mean that *in most cases* they follow the verb? Is there any quantitative statement hidden in Greenberg's statement? Whatever the answer to these questions, we can translate the categorical statements of Universal 25 into quantitative statements and check whether the

5 According to Næss (2006), Äiwoo (Austronesian, Oceanic) could be an exception. Cf. also Universal 502 in Constance University's Universal Archive (*http://typo.uni-konstanz.de/archive/*).

implication is verified on our data.[6] In other words, "the pronominal object follows the verb" (V pronO) can be interpreted as: "the percentage of pronominal object on the right of the verb is greater that *a*", where *a* is some relevant threshold. For instance, for *a* = 75%, we verify what is a first tentative quantitative universal:

> Universal 25': For every language, if the percentage of pronominal objects on the right of the verb is greater than 75%, so is the percentage of nominal objects on the right of the verb.

We abbreviate Universal 25' by: V pronO ≥ 75% → V nomO ≥ 75%.

Universal 25' is illustrated by ***Figure 3***. Let us recall that the negation of a property A → B is A & ¬B. Thus, Universal 25' claims that there are no language with V pronO ≥ 75% and V nomO < 75%, that is, that the corresponding rectangle in ***Figure 3a*** (hatched in gray) is empty of any language.

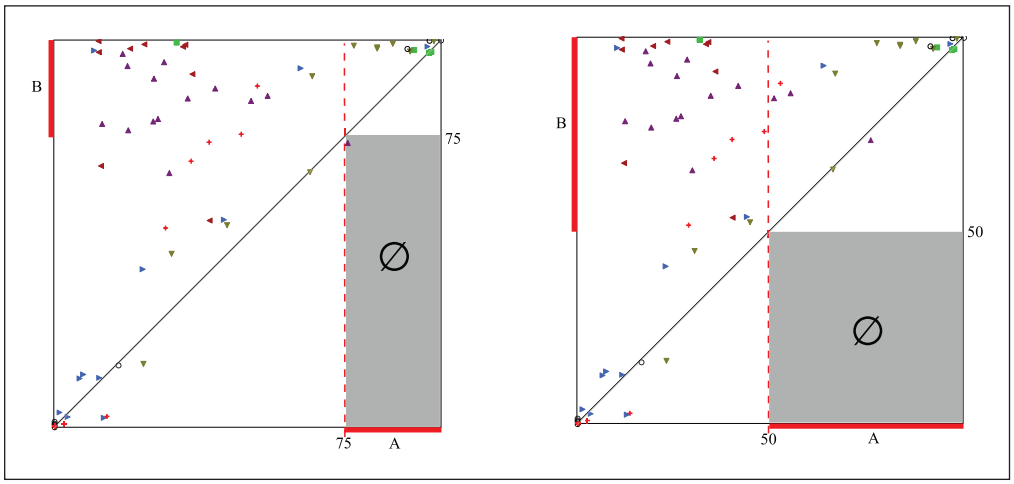

Figure 3 Universal 25' **a.** V pronO ≥ 75% → V nomO ≥ 75% **b.** V pronO ≥ 50% → V nomO ≥ 50%.

Yet, we do not know what the relevant threshold *a* actually is. If *a* = 100%, Greenberg's universal only concerns languages with very strict order where all pronominal objects are on the right of the verb. On the other side, if *a* = 50%, it concerns many more languages, that is, all the languages that place more pronominal objects on the right of the verb than on the left. But if the universal concerns more languages, the statement for each of these languages is also less strong, because it only says that these languages place more nominal objects on the right than on the left. In any case, this last statement is verified on our data as illustrated by ***Figure 3b***.

We conclude this subsection by saying that qualitative universals such as Universal 25 (V pronO → V nomO) could be interpreted by quantitative universals such as "V pronO ≥ *a* → V nomO ≥ *a*" for some *a*, but such an interpretation would suppose that we can define the relevant threshold *a*. We want to note that absolute implicational universals correspond to empty zones in our scatter plots, and statistical implicational universal correspond to almost empty zones. We will now see how the study of such zones in our diagrams can lead us to new formulations of universals.

2.2 A NON-IMPLICATIONAL QUANTITATIVE UNIVERSAL: THE TRIANGULAR PATTERN

Implicational quantitative universals correspond to empty or near empty rectangles in a scatter plots, as we have seen just before. But when we look at a scatter plot such as ***Figure 2*** we see empty zones which are not necessary rectangles. For instance, the scatter plot of ***Figure 2*** show a triangular pattern with almost all languages above the diagonal (see ***Figure 4***).

The diagonal represents the languages for which the percentage of nominal object on the right of the verb (V nomO) equals the percentage of pronominal object on the right of the verb (V pronO). If a language is above the diagonal, the percentage of V nomO is higher than the

6 It is not our purpose to interpret Greenberg's statements and to discuss if qualitative statements are justified or not as this is a difficult and controversial question. We would just like to point out that there is a possible interpretation of qualitative statements in terms of quantitative statements and that we aim to understand how they can be translated.

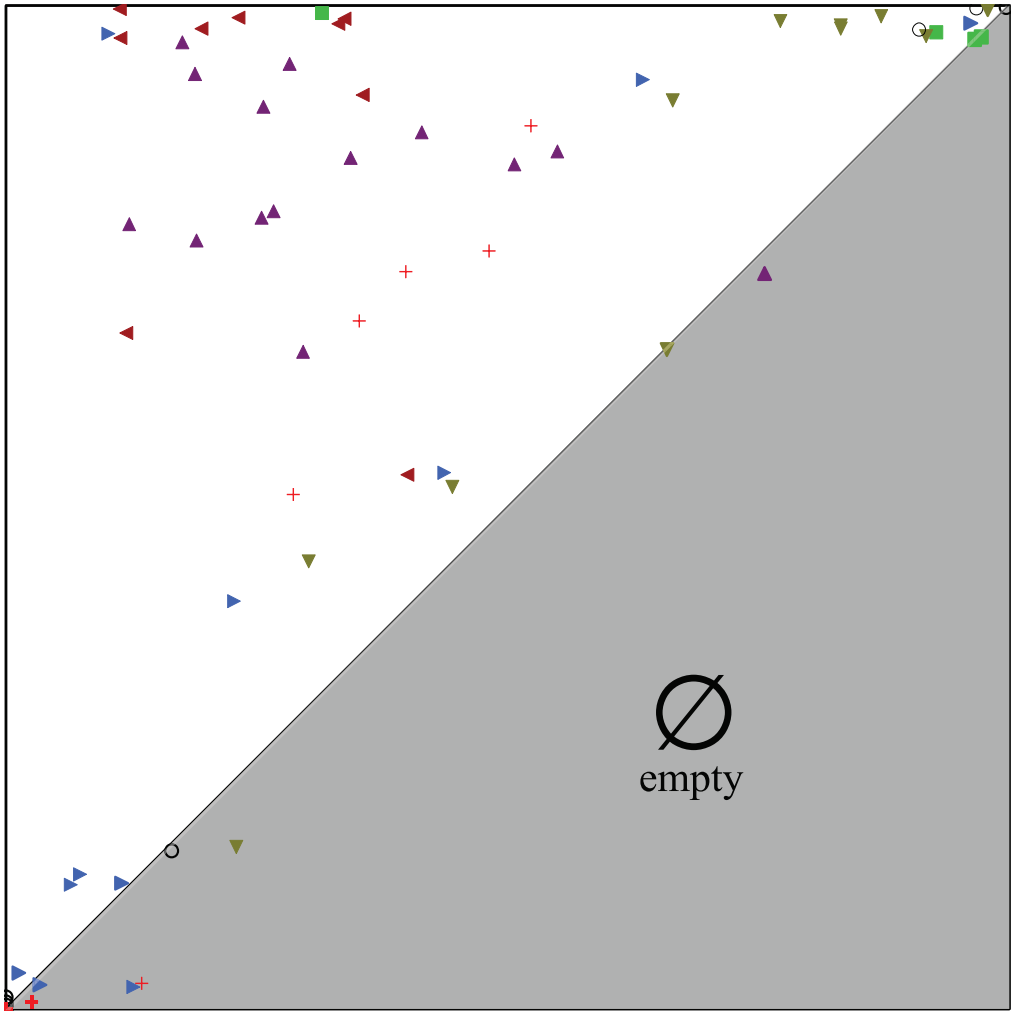

Figure 4 The Triangular pattern of V nomO ≥ V pronO.

percentage of V nomO. If X → Y is a relation between a governor X and a dependent Y, we note XY(L) the percentage of Y on the right of X in the language L.[7]

What our scatter of ***Figure 2*** shows is the following property of our sample:

> For almost every language L, V nomO(L) ≥ V pronO(L).

This is a quantitative language universal that we can abbreviate as:

> V nomO ≥ V pronO.

Such a universal is called a *quantitative universal* because the statement we make about languages is based on quantitative data. But we can remark that the statement itself is quantitative: It puts into relation two numerical values. This was made possible because we work with quantitative data. We can also put it into words as follows:

> Inequality Universal (for the pronominal and nominal objects). Almost every language has a higher proportion of nominal objects than of pronominal objects on the right of the verb.

2.3 COMPARISON OF QUALITATIVE AND QUANTITATIVE UNIVERSALS

When comparing the two universals, our V nomO ≥ V pronO and Greenberg's *V pronO → V nomO*, we can see that our universal is a quite powerful extension of Greenberg's universal.

First, we point out that our universal can also be expressed in an implicational form:[8] V nomO ≥ V pronO means that:

> For all L and all *a*, if V pronO (L) = *a*, then V nomO ≥ *a*.

7 Readers must be aware that the symbol "→" is used both for logical implications between statements and for syntactic dependencies between linguistic units.

8 This implicational form does not mean that the statement is an implicational universal. It does not have the form "for every language (or for almost every language), A → B".

Or equivalently:

> For all L and all *a*, if V pronO ≥ *a*, then V nomO ≥ *a*.

In other words, what seems to be claimed by Greenberg for some particular *a* appears to be true for every *a*, including values of *a* lower than 50%. For such values of *a*, it is better to use the contrapositive version of the implication:[9]

> For all L and all *a*, if V nomO < *a*, then V pronO < *a*.

Figure 5a illustrates the statement for the particular value *a* = 25%. In other words, our inequality universal can be interpreted as a new qualitative universal, that mirrors Universal 25, but seems to be statistical rather than absolute:[10]

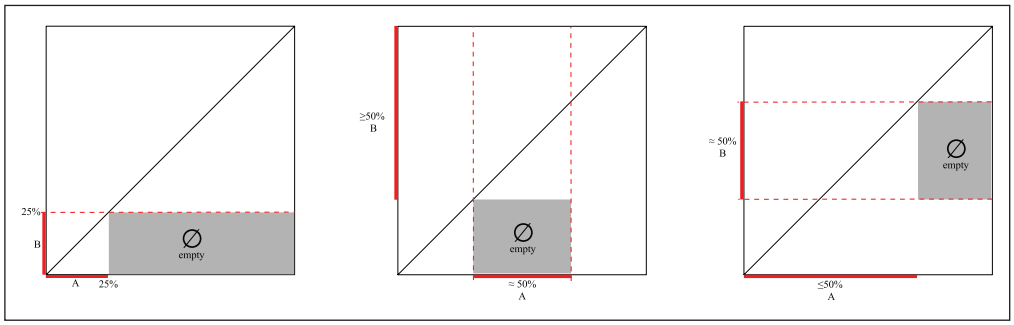

Figure 5 Mirrored Universal 25 and claims for free word order languages: **a.** V nomO < a → *V* pronO < a with a = 25%. **b.** A → B, with A = *"V pronO ≈ 50%"* and B = *"V nomO ≥ 50%"*. **c.** B → A, with A = *"V pronO ≤ 50%"* and B = *"V nomO ≈ 50%"*.

> Mirrored Universal 25. With overwhelmingly greater than chance frequency, if the nominal object precedes the verb in a language, then the pronominal object does so too.

But our universal V nomO ≥ V pronO is even more informative than the two preceding qualitative universals (Universal 25 and Mirrored Universal 25) because our quantitative universal also concerns free word order languages. It contains the two following qualitative universals:

> First Claim: If the pronominal object has a free word order, then the nominal object has a free word order or it follows the verb (***Figure 5b***).

> Second Claim: If the nominal object has a free word order, then the pronominal object has a free word order or it precedes the verb (***Figure 5c***).

In other words, while qualitative universals claim the absence of languages in rather small parts of the diagrams, our qualitative diagrams eliminate half of the diagram, as shown by ***Figure 4***.

Note that both quantitative and qualitative universals may hold for all or only for most languages, i.e. they may be absolute or statistical. Actually, the statement that V nomO ≥ V pronO is not true for four of the UD treebanks, Afrikaans, Turkish, Marathi, and Old Church Slavonic. Note, though, that these four languages are not highly deviant, i.e. far from the diagonal. Thus, we can say that our universal is statistical rather than absolute.

A comment can be made on this last point. We could propose an absolute quantitative universal by relaxing our statement to V nomO ≥ V pronO – c, for some constant c. For instance, this inequality holds for c = 10% for the scatter plot of ***Figure 2***.

Yet, in general, quantitative universal should be seen as distribution of languages in terms of cloud patterns. These cloud patterns can include statistical outliers. Any completely empty (rectangular) zone can be interpreted as an absolute (implicational) universal. Any almost empty zone can be interpreted as a statistical universal. But when we look at the scatter plots, we could also be interested in non-empty zones. We see zones with more or fewer languages, and the whole distribution of languages could be relevant. In other words, quantitative universals are statistical in nature and can be seen as a generalization of statistical qualitative universals.

9 The contraposition law says that: A → B is equivalent to its contrapositive ¬B → ¬A.

10 We must remember that our speech sample is unbalanced, and any universal statements we try to make must be made with great caution and considered as a method of data analysis that is bound to improve with more varied data.

Before enlarging our discussion to the study of other distributions on scatter plots, it is necessary to specify how our diagrams are obtained and some possible biases of a quantitative study such as ours.

3 FROM UD TO SCATTER PLOTS

In this section, we will outline how measures of directional argument and modifier placement of the type we have seen in the preceding section can be taken on dependency treebanks.

Dependency syntax encodes the syntactic structure of a sentence as a Directed Acyclic Graphs (DAG) of relations between words. Each relation is represented as a directed edge that goes from the head word to another word of the phrase (Tesnière 1959 [2015], Mel'čuk 1988, Gerdes & Kahane 2011). The direction of dependencies, which indicates the relative position of a phrase towards its governor, is the base of our measures. In ***Figure 6***, the (linearly) ordered dependency tree has three *head-initial* relations (for example *understand* → *typology*) and three *head-final* relations (for example *treebanks* ← *help*).[11] Dependency Syntax considers syntactic relations between words independently of word order, and dependency trees can be represented as simple dominance relations. No hypothesis on a basic word order has to be stipulated for the representation itself and the notion of *basic word order* is foreign to Dependency Syntax: When studying word order in Dependency Syntax, we assess the different linearizations of an unordered dependency tree. Each dependency has two possible linearizations (governor → dependent or dependent ← governor), one of which may be *dominant* in the sense that it appears more frequently.

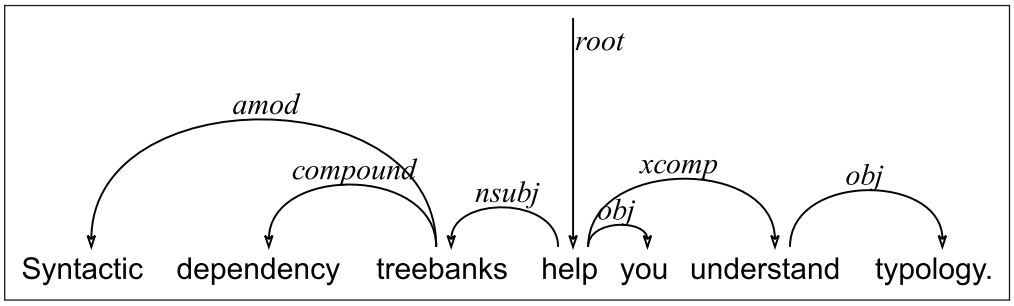

Figure 6 Example of an ordered dependency tree.

Recently, many corpora have been enriched with dependency-based syntactic analysis (so-called dependency treebanks). Universal Dependencies (UD) is the largest collection of dependency treebanks, annotated in a common cross-linguistically consistent annotation scheme. UD has been developed with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a perspective of language typology (de Marneffe et al. 2014, Nivre et al. 2016, Croft et al. 2017). The annotation scheme is an attempt to unify previous dependency treebank developments based on an evolution of (universal) Stanford dependencies (de Marneffe et al. 2006), Google universal part-of-speech tags (Petrov et al., 2012), and the Interset interlingua for morphosyntactic tagsets (Zeman, 2008). The general philosophy is to provide a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages, while allowing language-specific extensions when necessary. UD expects the schema, as well as the treebank data, to be "satisfactory on linguistic analysis grounds for individual languages", and at the same time, to be appropriate for linguistic typology, i.e., to provide "a suitable basis for bringing out cross-linguistic parallelism across languages and language families".[12]

One outstanding advantage of using this data set for language typology studies is the sheer size of the data set: UD 2.2 includes 110 treebanks in over 70 languages and is constantly

11 The syntactic analysis of this sentence is subject to debate. The proposed analysis corresponds to what is commonly done in dependency syntax. The annotation choices are based on theoretical considerations, for instance the analysis of *you* as an object of *help* rather than as a subject of *understand*. See Hudson 1998 for a comprehensive overview of the stakes of this particular question in a dependency perspective.

12 UD introduction page *http://universaldependencies.org/introduction.html* consulted in August 2017.

growing.[13] Moreover, most importantly, all UD treebanks use the same annotation scheme.[14] Therefore, UD can, to a certain extent, provide rich informative evidence that can be easily compared and interpreted across authentic texts of various languages.

Our study is based on Surface-Syntactic Universal Dependencies (SUD), a variant of the UD annotation scheme (Gerdes et al. 2018, 2019). SUD is better suited for word order studies as it is based on distributional criteria whereas UD favors relations between content words (Gerdes & Kahane 2016, Osborne & Gerdes 2019). In SUD, contrary to UD, prepositional phrases are headed by prepositions, and auxiliaries and copula are analyzed just like other matrix verbs, taking the embedded verb as a dependent. For a discussion on the criteria that allows deciding whether a construction is clearly headed (endocentric in the terms of Bloomfield 1933), see for instance Criteria B of Mel'čuk (1988).

The choice of the SUD version is particularly important when we consider a comprehensive view of all constructions of one language, for example Japanese is nearly completely head-final in SUD whereas Japanese UD has a number of head-initial relations such as adposition-noun constructions and auxiliary-verb constructions. More generally, the choice of functional words as heads is strongly correlated with other relations as shown by many studies in typology and confirmed by our data (see Section 7.2).

Some noise remains nonetheless as a result of the UD to SUD transformation. For instance, the UD scheme considers an *expl* relation for expletive elements. While this can be justified for languages such as German (which can have an expletive *es* in the preverbal position), it is problematic for languages such as English when the surface subject *it* of impersonal constructions is analyzed as an expletive while the demoted subject remains analyzed as a subject (see the *expl* and *subj* dependencies in ***Figure 7***), thus giving a VS-structured sentence. In original SUD format, it is preferable to analyze "it" as the subject (*subj*) and "to make your contribution" to be the quasi subject but the translation UD-SUD rules do not allow to recover the information about *it* being the subject.

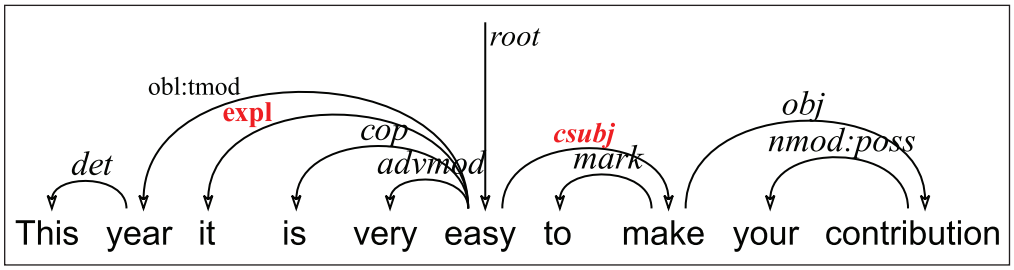

Figure 7 Expletive subject in English (erroneous SUD analysis translated from UD-English v2.0).

Problems also arise with the relation *dislocated*. For instance, in the Cantonese treebank, 100% of the object relations go to the right, and left peripheral objects are all annotated with a *dislocated* relation even if there is no resumptive pronoun (Wong et al. 2017). Put differently, UD obliges the annotator to choose between the annotation of the (rather discursive) information about the word's dislocation status and the (rather syntactic and valency based) information about the word being an object. For our study, dependencies such as *dislocated* are kept only when considering all the head-daughter dependency relations and not taking into account label distinctions.

13 The most recent version at the time of writing is UD 2.5 with 157 treebanks of 90 languages. The present analysis is based on the Surface-Syntactic Universal Dependencies (SUD) version 2.2. The languages added in the most recent versions are mostly very small and would have been filtered out by our threshold values.

14 This point raises a general concern on the possibility of using the same categories for languages that categorize the world differently (Sapir 1985, Croft 1991, 2002). These papers show that even if an agreement on a common set of categories can be reached for some constructions the projection on the presupposed universal scheme will remain problematic and arbitrary. Moreover, some treebanks are a result of multiple transformations of previous phrase-structure and dependency treebanks without manual corrections, therefore often multiplying already existing annotation errors and annotations of the different UD treebanks in the different languages have been done by different teams, which come from different theoretical backgrounds. But we assume that such problems are not specific to our study but rather faced by every study in typology when data collected by different linguists coming from different horizons is compared. It is important to point out that the UD project has a very lively open discussion group where everybody can ask questions about the annotation of a particular phenomenon and a higher inter-language agreement can be expected to be reached over time.

From the set of SUD treebanks, we can compute for any relation the percentage of head-initial links. We can also filter the links of any given relation by the POS of the governor or of the dependent to look into more specific sub-cases (such as the pronominal objects of a verb). For each relevant triple (POS of the governor, relation, POS of the dependent – the POS can remain under-specified) and each of the UD languages (merging all treebanks of the same language),[15] we computed the number of head-initial and head-final dependencies.

The scatter plot of ***Figure 8*** shows the percentage of head-initial head-daughter dependencies, that is, dependencies that link a head with a constituent that is subordinated to it.[16] The set of SUD/UD relations taken into account for our study includes clear-cut cases of headedness such as the relation between the verb and its object (*obj*) as well as more controversial relations between functional words and content words, as discussed above.

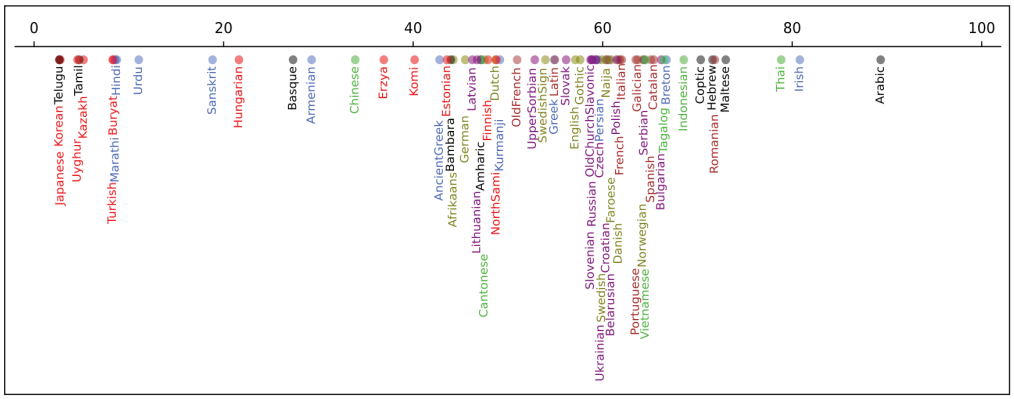

Figure 8 Percentage of head-initial head-daughter dependency relations in the SUD treebanks ranging from 3% for Japanese to 89% for Arabic.

We do not consider SUD/UD relations the direction of which are fixed by the annotation scheme such as *conj, appos, reparandum, fixed, flat, list, parataxis, orphan, goeswith*. Moreover, the *punct* and *root* relations are not syntactically relevant relations and the *dep, clf*, and *compound* relations are not homogeneously annotated on all treebanks. We decided to keep the *det* relation for determiners, even though the relation linking a determiner and a noun does not always provide a clear-cut head (cf. the DP-hypothesis; Hudson 1984, Abney 1987). One of the reasons we keep the *det* relation is that it has been used even in some languages, such as Japanese, which do not have clear determiners, for the closed classes of adjectives which have a similar meaning as English determiners.[17]

4 TYPOMETRICS

This section starts by defining the notion of typometrics and places it in the context of typological studies in Section 4.1. In Section 4.2, we describe one-dimensional distributions, starting with the distribution of languages across the head-final and head-initial spectrum. We compare the resulting distributional diagram with Tesnière's classification. This raises the question of the contrast between free and mixed word order languages, which is developed in Section 4.3.

4.1 DEFINITION AND HISTORICAL CONTEXT

We propose to call *typometrics* the open field of the study of the distribution of languages in a distributional scatter diagram based on empirical measures on corpora. The term *typometrics* is directly inspired by the term *textometrics*, sometimes used for *textual data analysis.* Contrary to textometrics that generally compares texts, authors, or genres from the same language, typometrics attempts to compare languages.

15 We are aware that treebank properties not only reflect the language but also show genre differences as well as annotation choices. As shown in Chen & Gerdes (2017), the global measures for different treebanks of the same language remain nevertheless quite homogeneous.

16 Our approach does not consider movement; we only take measures on observed "surface" data. Thus, the discussions inside Generative Grammar about head-initial and head-final deep structures is not of our concern.

17 We consider that a language has clear determiners when the noun cannot be used alone in some argument positions.

This paper focuses on word order typometrics, but typometrical statements can be made on all levels of the linguistic analysis, from phonetics to semantics, on grammar as well as on the lexicon. Greenberg (1954[1960]) is to our knowledge the first to propose a typometrical study. He introduces numerous interesting typometrical parameters such as the degree of synthesis (the number of morphemes per words), the degree of agglutination (the percentage of concatenative combination of morphemes) that are evaluated on a 100-word text in 8 languages. Among the ten parameters he considers, three parameters characterize the marking of syntactic dependencies: the percentage of dependencies that are marked by an agreement morpheme ("concordial index"), a case morpheme ("pure inflectional index"), or only by word order ("isolational index"). Greenberg presents the quantitative measures in the form of tables.

Similarly, Krámský (1959, 1972) proposes a typometrical study on the distribution of types of consonants. Referring to Isačenko (1939), who classified Slavic languages into three categories (radically vocalic, radically consonantal, and mixed type) according to the relative frequency of consonants and vowels, Krámský (1959) has expanded this previous work "by comparing the occurrence of vowels and consonants in the phonemic inventory with the occurrence of vowels and consonants in coherent texts".

It seems that these pioneering works was not followed by many similar works over the next three decades (perhaps due to the dominance of Generative grammars). Givón (1983: 21) contains a quantitative study of the distance between coreferent elements in eight languages, showing that this distance is correlated with the realization of the second element (zero anaphora or left vs. right dislocation). According to Cysouw (2005), Myhill (1992) discusses a variety of indices based on text counts, for example, languages that are mostly SV use VS to mark non-temporal sequencing. Fenk-Oczlon & Fenk (1999) study correlations between a number of phonemes per syllables and syllables per words or clauses in order to confirm Menzerath's law. Their work is based on counts of a small set of non-attested data (22 sentences translated in every language) in 34 languages (with the similar bias as our study, since half of the languages are Indo-European). The study mixes these typometrical parameters with categorical statements on word order (OV vs. VO) and on morphological type (isolating, agglutinative, fusional). As already mentioned in our introduction, the availability of syntactic treebanks in several languages in the 2000s opened the door for typometrical studies in syntax, such as Liu (2010) on word order. In the same way, Futrell et al. (2015) test on a sample of 34 dependency treebanks that languages with more word order freedom have more case marking.

Typometrics is a subfield of quantitative typology. Some research in the field of quantitative typology aims to introduce general models of language, such as probabilistic and information-theoretic models (Perfors et al 2010, Ferrer-i-Cancho 2015 & 2018), among others. Yet most work in the field of quantitative typology is based on categorical statements. Typology has been fueled by the growing number of languages for which traditional categories of typology have been instantiated. On the basis of this significant amount of data, it became possible to carry out statistical studies and to check whether the qualitative universals were verified (instead of basing studies only on a balanced sample of a few dozens of languages). Many remarkable studies have been conducted in this direction (in particular Nichols 1992, Dryer 1992 or for more recent works, Daumé III & Campbell 2009, Ferrer-i-Cancho 2008, 2016) without being typometrical studies. Typometrics thus becomes a branch of quantitative typology, which is characterized by the study of quantitative values obtained by empirical measurements on corpora and does not categorize them before studying their distribution among languages.

It is worth noting the work of Justeson & Stephens (1984) on the relationship between the numbers of vowels and consonants in phonological systems. Although this work is not pure typometrics (they do not count data in texts but in grammars), they are the only ones, to our knowledge, to propose a scatter plots of languages (***Figure 9***). They argue that the number of vowels and the number of consonants have a log-normal distribution across all languages in the world, and using the Pearson correlation, they conclude that these two variables are uncorrelated.

Typometrics is also not primarily concerned with sampling and the choice of a statistically relevant sample of languages, that is part of quantitative typology (Bell 1978, Perkins 1989, 2001). For a survey and a classification of works in quantitative typology, see Cysouw (2005).

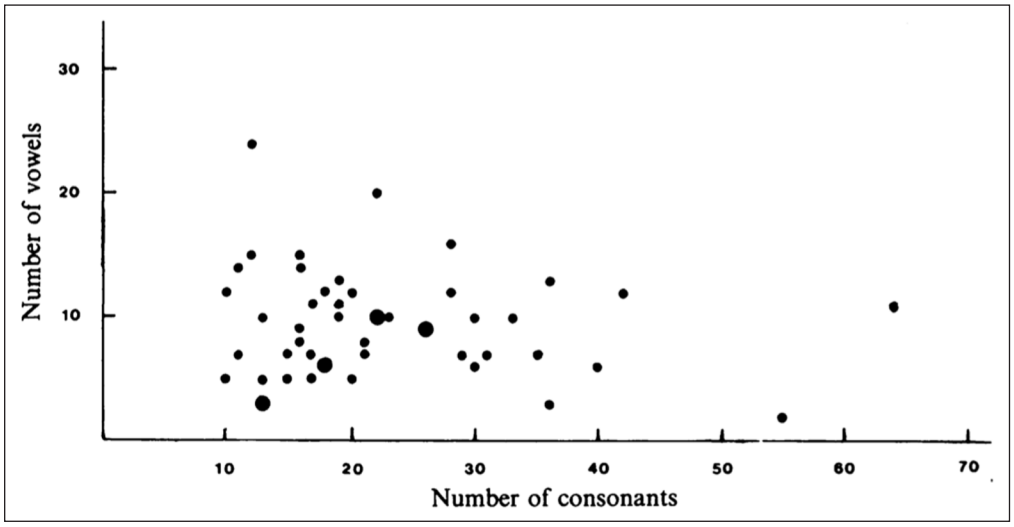

Figure 9 Scatter plots by Judeston & Stephens (1984: 534). Original title: Scattergram of languages according to the numbers of vowels (V) and consonants (C) in their inventories; heavier dots indicate that two languages have the same values of V and C.

Some works in typology which are not typometrical can be based on typometrical studies when the types are defined quantitatively, but one of the specificity of typometrical works such as ours is to not introduce types and to proceed to the typological studies with continuous parameters. In some sense, typometrics is typology without types.

4.2 TESNIÈRE'S CLASSIFICATION IN HEAD-FINAL AND HEAD-INITIAL LANGUAGES

In this section, we will present a language classification of Tesnière (1959), which, without being a typometrical studies, can be compared to our quantitative data.

Some years before the seminal work of Greenberg on word order universals, Tesnière (1959) proposed a classification of languages based on the dependency direction referring to Steinthal (1850) and Schmidt (1926).[18] He introduced the terms *centrifugal* for head-initial languages and *centripetal* for head-final languages.[19] Moreover, he opposes *strict* word order, when head-daughter relations mostly go in one direction, to *mitigated* when the head is amidst its dependents going out in both directions.

Tesnière worked on a sample of languages that is quite close to what we find in SUD. ***Figure 10*** shows that his classification agrees well with our measures. On the left is his distribution diagram.[20] On the right is the one-dimensional diagram of ***Figure 8***. The order proposed by Tesnière is reasonably well respected:

Semitic < Celtic < Romance < Germanic < Slavic < Chinese < Ural-Altaic

Of course, Tesnière's classification is sometimes much too coarse when he puts all American languages, all Papuan languages, and all "Black-African" languages (except Bantu and South-African languages) in one position, but these languages are not represented in the UD database except for a very small Bambara treebank.[21] Tesnière did not consider the Indo-Aryan branch of Indo-European languages, which are quite well represented in our database (Persian, Sanskrit, Urdu, Hindi, Kurmanji, and Marathi). Basque is the only language that was not well classified and appears to be much more head-final than foreseen by Tesnière. Slavic languages appear

18 Tesnière's book was published five years after his death. He never published his typological work in his lifetime, which is condensed in a few pages of his book (Chapter 14). Tesnière Funds of the French National Library contains four big tables with a total number of 150 languages (BNF, Fonds Tesnière, Box 39, Folder 3). For each language, Tesnière has indicated the direction of 5 relations (Noun-Genitive, Noun-Possessive, Verb-Subject, Verb-Object, Noun-Adjective), as well as whether the language has prefixes, infixes, suffixes, prepositions, or postpositions. Note that Tesnière did not consider the Adposition-Noun relation as a head-daughter relation.

19 In head-initial languages, the dependents "escape from" the head, which gives *centrifugal*, and, in head-final languages, the dependents "seek" the head, which gives *centripetal*. Henri Weil had already introduced the notion of ascending (head-final) and descending (head-initial) in his thesis on word order in 1844 (p. 73).

20 From Tesnière 1959, ch. 14. The present version is extracted from the English translation, which is consistent with the original.

21 Bambara is a Mande and not a Bantu language, but both groups belong to the Niger-Congo family.

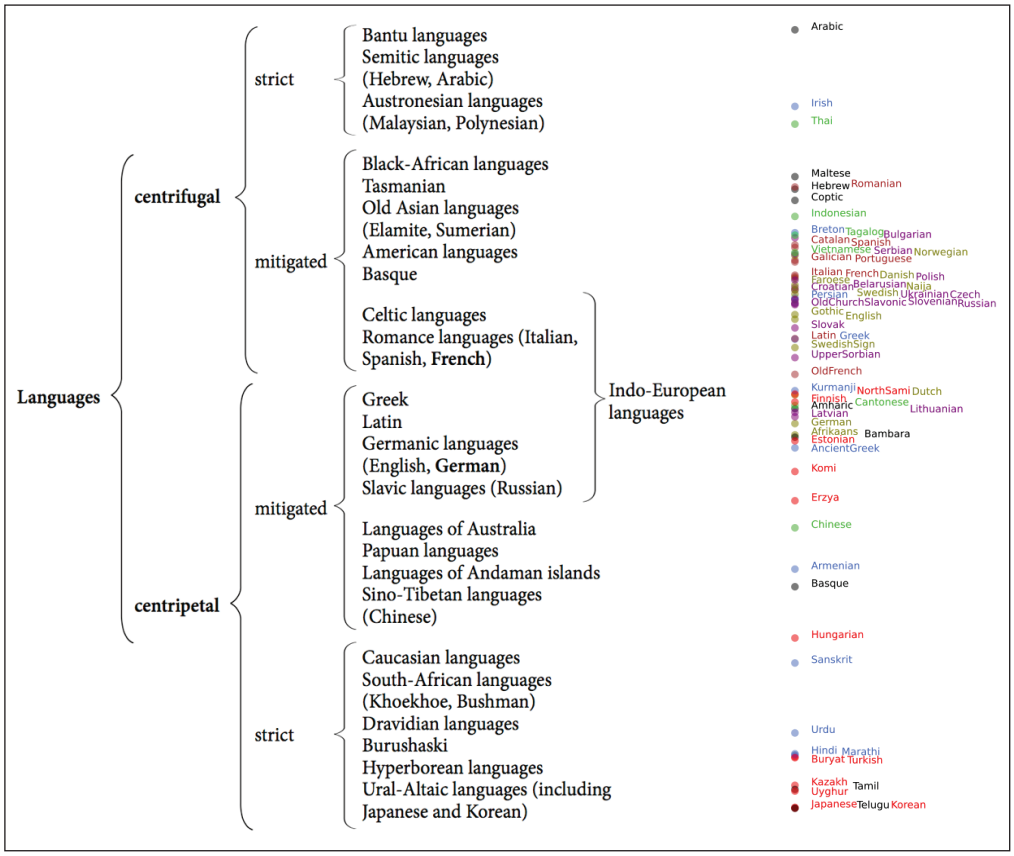

Figure 10 Tesnière's "typological classification of languages according to the nature of linearization" and a vertical display of Figure 8.

to be more head-initial than expected,[22] and the same holds for Finno-Ugric languages (Sami, Estonian, and Finnish), which belong to the Ural-Altaic family.

Liu (2010) was the first to order a set of languages according to the percentage of left and right branching dependencies. His results, based on 20 dependency treebanks available at that time, with different annotation schemes, were also compared to Tesnière (1959)'s classification, by means of histograms, which did not make the comparison as explicit as here.

4.3 FREE VS. MIXED WORD ORDER

Tesnière's term *mitigated*, for the languages in the middle of ***Figure 10*** that do not have a clear tendency towards either head-initiality or head-finality, conflates two notions: free and mixed word order languages. We usually mean that a language has a *free word order* if the **same** syntactic (dependency) tree can frequently be ordered in different ways, i.e. in many topological paraphrases. Two topological paraphrases differ only in their communicative structure (information packaging). *Mixed word order* languages have head-initial and head-final dependencies, but a given syntactic dependency usually accepts only one specific direction for the dependent.

In French, for example, pronominal objects are commonly on the left whereas nominal objects are on the right of the verb. We must not conclude from the balanced left/right distribution of objects that French is a free word order language. On the contrary, French hardly allows any exception in this matter and few dependency trees have variations in the object placement. Equally, Persian is not a free word order language just because nominal objects are placed to the left of their verbal governor while verbal dependents that fill the object slot go to the right of the governing verb. The distinction between free word order and mixed word order is possible if we introduce the right features. For instance, for French, we notice that the object word order is not free as soon as we distinguish pronominal and nominal objects.

The 33.9% of Chinese head-initial head-daughter dependency relations comfort the view of Chinese as a mixed word order language (Li & Thompson 1989). Note that the values depend again on the theoretical view taken on a series of phenomena. For instance, the analysis of 把

22 Nevertheless, Tesnière is known to have been a very good slavist. His PhD thesis was on the dual in Slovenian, and he wrote a grammar of Russian (Tesnière 1945).

ba (the marker for direct-objects in so-called 'ba' sentences) and 被 bei (a passive marker) as prepositions instead of coverbs, directly creates a higher number of head-final object relations (Sun & Givon 1985). This analysis has been chosen for the Mandarin UD treebank, contrarily to the UD analysis of Cantonese (Wong et al. 2017) which accounts for the 14% gap to the 47.5% of Cantonese head-initial relations.

Futrell et al. (2015) proposes a way to estimate to which extent the order possibilities of a given relation are syntactically determined, that is, determined by syntactic features such as the POS of the governor or the dependent and the relations and POS of codependents. Put more simply, a narrow selection of syntactic relations, for example, the direction of the nominal object, allows for measures that can be expected to be correlated with what is commonly called "free word order": In a free word order language, it should be hard to make out the predominant order of the direction of nominal objects (Hawkins 1983, Mithun 1987). In Slavic languages, for example, there is no good way to identify on a syntactic basis a verbal dependent relation that is strict concerning order.

We see that languages in the middle of the scatter graph of ***Figure 11***, where around 50% of the nominal objects go to the right and the other half to the left of their governor, are commonly classified as free word order languages. Note that Czech, just as other Slavic languages, does not fall at 50% but rather at an 80% average of head-initial nominal object relations because Slavic languages have the predominant word order VO, the OV expressing special communicative configurations (theme/rheme structures) that appear less frequently, in particular in written texts. German and Dutch appearing closer to the 50% mark of head-initial nominal object relations does not necessarily indicate that these languages have a freer word order than Slavic languages in the sense that a given dependency tree has a wider choice of possible linearizations but rather that both orders appear with nearly equal frequency. If we separated finite and infinite verbal governors for German and Dutch, we would have predominantly VO in the first case and OV in the second. This not only shows the difficulty of defining a basic word order for German and Dutch but also the interdependence of our results on the treebanks' part-of-speech and morpho-syntactic distinctions.

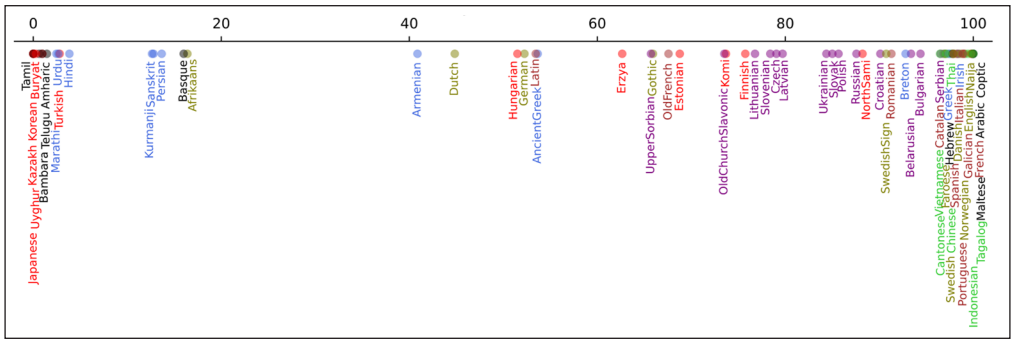

Figure 11 Percentage of head-initial nominal object relations with a verbal governor. On the left of the graph, we see OV languages and on the right are the VO languages.

5 A ONE-DIMENSIONAL DISTRIBUTION RELATED TO A ONE-DIMENSIONAL UNIVERSAL

In this section, we relate one of Greenberg's universals to a one-dimensional distribution. In Section 5.1, we show how other distributions can be interpreted as quantitative and qualitative universals.

For each syntactic relation we consider, it is possible to order our set of languages and to produce a one-dimensional diagram. Most Greenbergian universals (Greenberg 1963) relate several types of relations, and we will look at these universals in Section 6. Just one Greenbergian word order universal involves only one relation and can be tested on a one-dimensional diagram – we will call such universals *one-dimensional universals*.

> "Universal 19. When the general rule is that the descriptive adjective follows, there may be a minority of adjectives which usually precede, but when the general rule is that descriptive adjectives precede, there are no exceptions."

This universal means that languages with dominant *AN* order, that is, with a head-final NOUN → ADJ relation, must necessarily have a very low percentage of head-initial occurrences. In other words, a gap in the area of moderately head-final languages is expected for this relation.

If we look at the distribution of languages for the NOUN → ADJ relation in ***Figure 12***, we see that Universal 19 is more or less confirmed. On one hand, there is no real gap in the distribution of dominant head-final languages, due to the presence of Polish and Old French between 20% and 50%.[23] On the other hand, we observe that the distribution of head-initial languages is much more uniform than the distribution of head-final languages, whose languages are highly concentrated between 0% and 5%. More precisely, the average percentage of head-initial languages is 83.4% with a standard deviation (SD) of 14.2. On the left side of the graph, we obtain an average of 3.8% and an SD of 9.1, which confirms the universal statistically.[24]

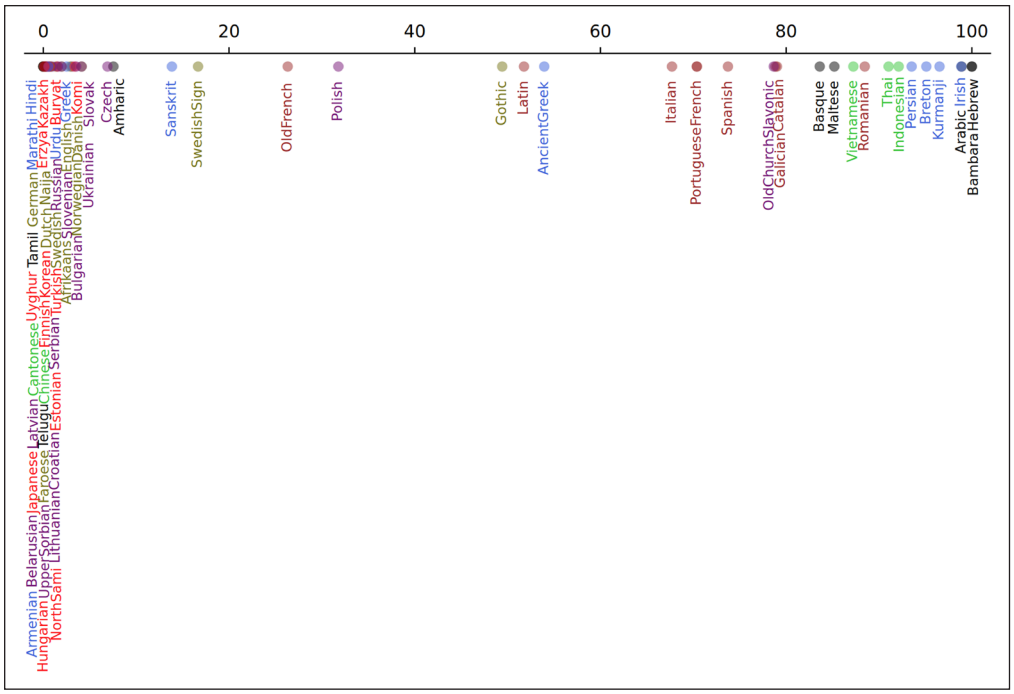

Figure 12 Language distribution for the direction of the NOUN → ADJ relation. On the left of the graph, we see *AN* languages, and on the right are the *NA* languages.

When analyzing further the Greenbergian Universal 19, we note that the interpretation of the condition "when the general rule is that the descriptive adjective follows" is difficult to apply empirically. If we take this rule to hold for all languages with predominant NA order (i.e. with a NA score of more than 50%), we include the classical languages Latin, Gothic, and Ancient Greek in this group although their position is just above 50%. A universal such as Universal 19 tries to describe the (quantitative) distribution of languages considering a special feature (the distribution of ADJs towards the NOUN) in qualitative terms, which is not straightforward. We believe that a diagram such as ***Figure 12*** can be a more satisfying alternative to such descriptions.

5.1 OTHER ONE-DIMENSIONAL DISTRIBUTIONS

The distribution for other relations can be examined in order to produce claims similar to Universal 19. For instance, the ADP → NOUN relation (***Figure 13***, ADP = adposition) lends itself even better to universal claims since the distribution is even less homogeneous than the NOUN → ADJ relation: We have no languages between 25% and 65 % and only few languages between 5% and 95% (***Figure 13***). In other words, languages tend to be strongly prepositional or strongly postpositional, and few languages allow both prepositions and postpositions in a

23 A possible explanation for the presence of Old French is that the Old French UD treebank covers a wide period (842 to 1225, see Stein & Prévost 2013), where Latin, positioned at around 50% in our diagram, was influenced by Germanic tribes. We have no explanation why Polish is an outlier among the modern Slavic languages.

24 Recall that the standard deviation measures the average deviation of language positions from the mean. In other words, these measures confirm what can be observed in the diagram: The languages on the left side of the diagram are more concentrated and very much left-leaning, while the languages on the right side are more central and more balanced.

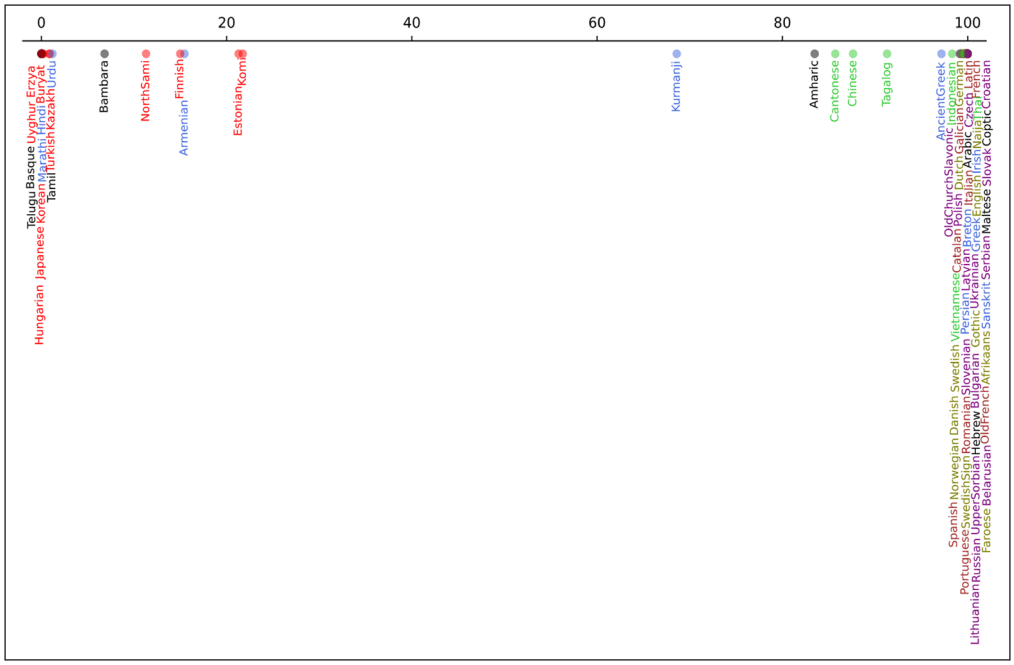

Figure 13 Language distribution for the direction of the complement of an adposition.On the left of the graph, we see postpositional languages and on the right are the prepositional languages.

significant number (see the remarks on Chinese and Cantonese in Section 4.3 that explain their astonishing position in this graph). For postpositional languages (below 50% in the graph), we obtain an average 4.8% with an SD of 7.7, the prepositional languages have an average of 98.1% with an SD of 5.5. In words, this could be spelled out as follows:

> Universal about adpositions. When the general rule is that adpositions precede the noun, there are almost no exceptions. Similarly, when the general rule is that adpositions follow the noun, there are almost no exceptions either.

One of the most unbalanced relations concerns pronominal subjects.[25] Except for Tagalog and Irish,[26] all languages in our sample have preverbal pronominal subject (that is head-final VERB-subj→PRON relations with less than 25% of pronouns on the right of the verb, ***Figure 14***). The total average is 9.5% with an SD of 16.8.

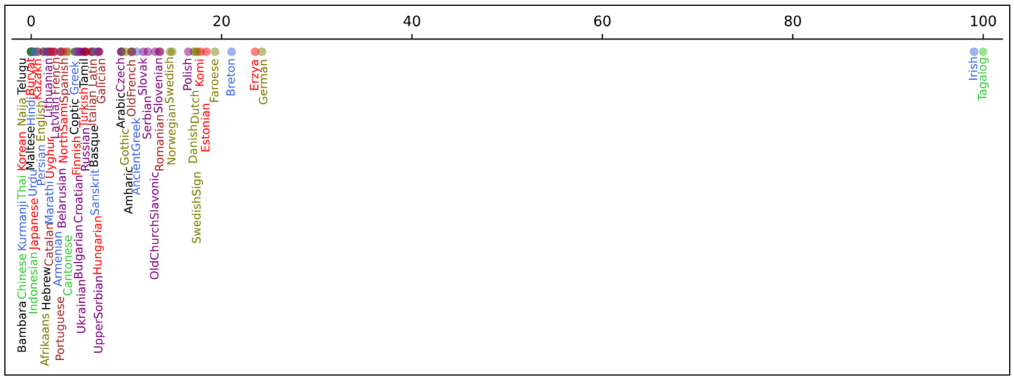

Figure 14 Language distribution for the direction of the pronominal subjects.

On the other hand, the language distribution for pronominal objects is well-balanced, although we observe a trend of pronominal objects on the left of the verb (***Figure 15***): The total average is 39.3% with an SD of 33.1. This remarkably balanced distribution cannot be described as a qualitative universal à la Greenberg, precisely because no configuration is excluded. This particular distribution is nevertheless a striking property of the set of languages that deserves to be recorded among the typological facts of Language.

Nominal objects (***Figure 16***) tend to be more on the right than pronominal objects (***Figure 15***). And verbal (= clausal) objects even further (***Figure 17***). We have already seen in Section 2.2 that the correlation between the rightness of nominal and pronominal objects shows a noteworthy

25 Note that we only consider VERB-subj→PRON relations. All subject relations depending on other POS such as auxiliaries, adjectives, and nouns are not taken into account in order to simplify the discussion.

26 Irish has clitic-like conjunctive forms of subject pronouns that may be in a grammaticalization process.

DOI: 10.5334/gjgl.764

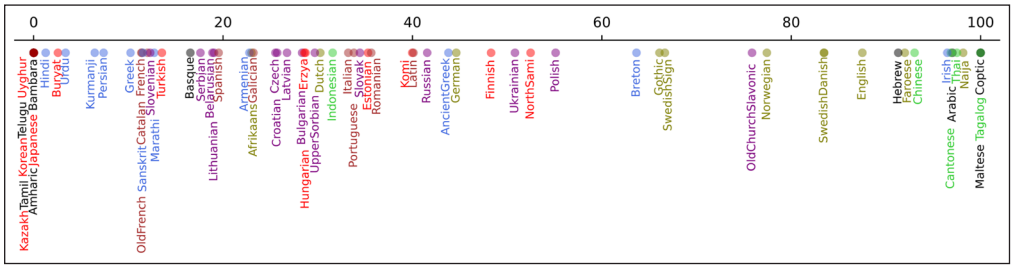

Figure 15 Language distribution for the direction of the pronominal objects.

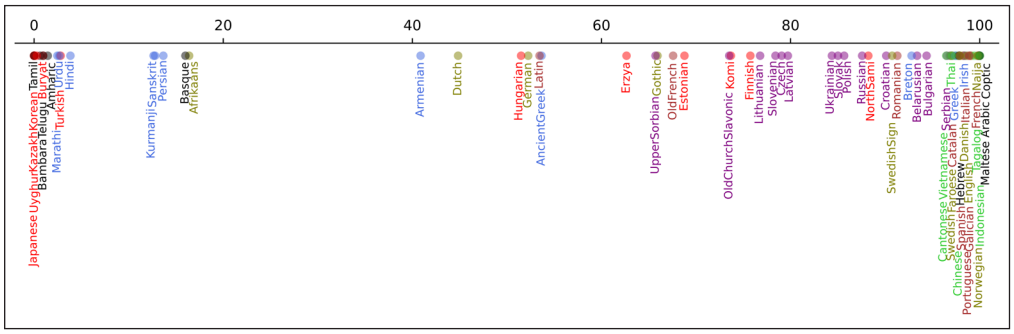

Figure 16 Language distribution for the direction of the VERB-object-NOUN relation.

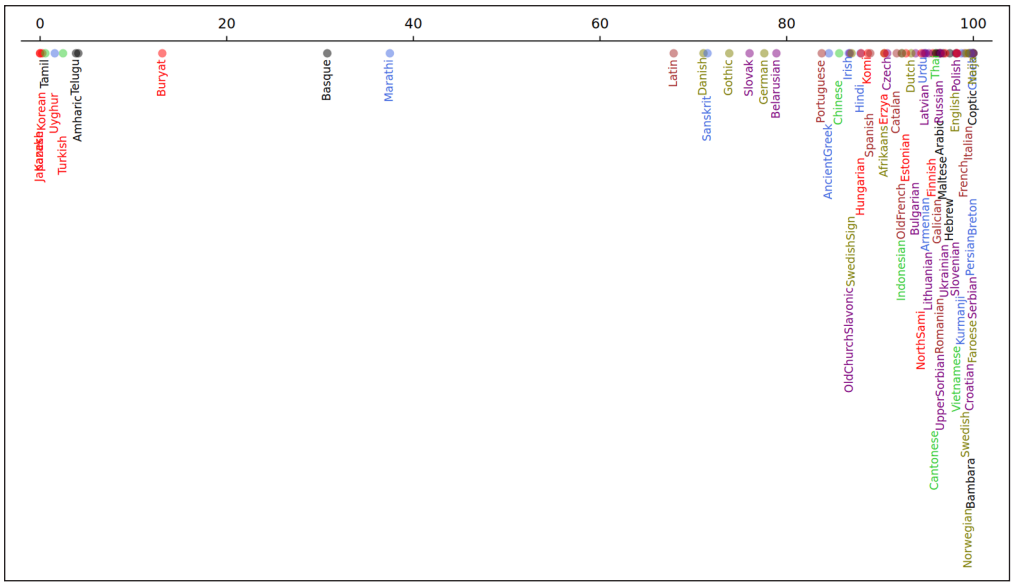

Figure 17 Language distribution for the direction of the verbal (= clausal) complements of a verb.

pattern. The total averages for pronominal, nominal, and clausal objects are respectively 39.3%, 65.2%, and 79.3%.

If we compare language distributions for adverbial modifiers and adverbial clauses, we also observe that clauses tends to be more on the right than other adverbials (***Figures 18*** and ***19***). The average of non-clausal adverbials is 25.1% vs. 51.5% for clausal adverbials.

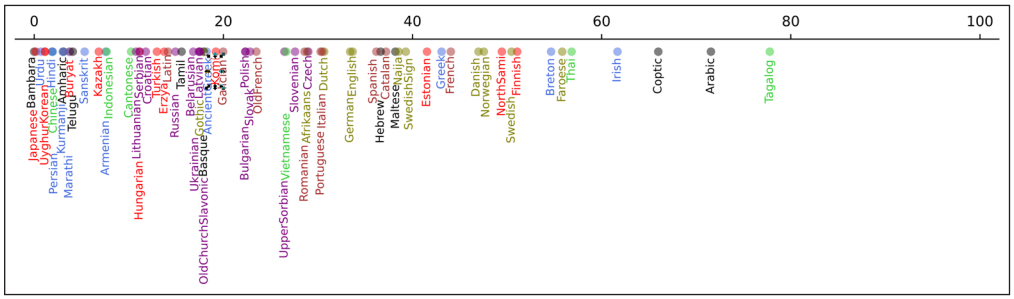

Figure 18 Language distribution for the direction of the adverbial modifiers.

More generally it appears that light dependents tend to be more on the left than heavier dependents if we consider pronouns to be lighter than noun phrases, noun phrases to be lighter than clauses, and adverbial modifiers to be lighter than adverbial clauses.

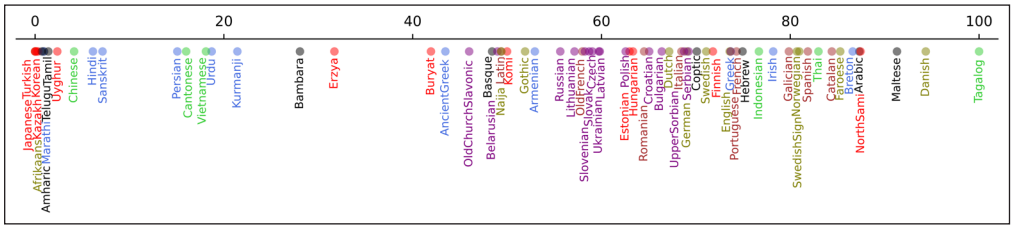

Figure 19 Language distribution for the direction of the adverbial clauses.

Universals à la Greenberg can only point to empty spaces in the distribution but fail to give a more global view of the directional tendencies. For the study of typological features, we see that a graphical representation of directional tendencies provides us directly with a clear overview of the distribution of the studied languages along with their characteristics. It not only allows comparing languages for a given feature but also comparing one feature with another through the analysis of different scatter plots of languages. This advantage of diagrams will become even more explicit when looking at two-dimensional scatter plots.

6 TWO-DIMENSIONAL LANGUAGE CLASSIFICATION

We have seen how to order our set of languages and to produce a one-dimensional diagram for each syntactic relation we consider. This can be generalized to two-dimensional diagrams and further to multi-dimensional diagrams if we consider more than two quantitative features.

As we have seen in the preceding section, it is possible to infer universals from distributional diagrams. A universal is *n*-dimensional if it involves *n* quantitative features, producing an *n*-dimensional diagram.

Dryer (1992) studies a range of two-dimensional universals by comparing the correlation between nominal objects (VO in his notation) and other relations. An example of a multi-dimensional universal can be found in Greenberg (1963): Universal 5 is at least three-dimensional; three-dimensional if we count the SOV order as only one dimension and four-dimensional if with count it as two dimensions, SV and OV:

> "Universal 5. If a language has dominant order SOV and the genitive follows the governing noun, then the adjective likewise follows the noun."

Putting aside the absence of genitive encoding in the present UD scheme, it is difficult to resolve the problem of operational visualizations for 3-dimensional typometrical scatter plots. Nevertheless, there is no theoretical difficulty to numerically analyze multi-dimensional typometrical configurations. We leave it to further studies to characterize relevant multi-dimensional patterns, and we focus on two-dimensional diagrams in this paper, starting in Section 6.1 with the use of such diagrams for language classification.

6.1 LANGUAGE CLASSIFICATION

Since Greenberg (1963), the most emblematic word order classification concerns the respective order of S, O, and V, where S and O stand for the nominal subjects and objects of a verb V. In dependencies, there is a direct relation between verb and subject as well as between verb and object, and we can easily read these two relations, V→S and V→O, from the treebank's dependency relations. In our measures of the V→S relation, we did not separate subjects in transitive and in intransitive constructions.[27]

Figure 20 proposes a two-dimensional diagram, with the percentage of head-initial nominal subject dependencies (VS) on the X-axis and the percentage of head-initial nominal object dependencies (VO) on the Y-axis.

Figure 20 gives us two kinds of information about word order: The first information corresponds to Greenberg's classification. As shown in ***Figure 21***, each quadrant of the diagram corresponds to a dominant word order. As expected, we do not have any OVS language (such languages are very rare). Most of our languages are SVO due to the predominance of Western Indo-European

27 Our computation simply counts overall frequencies and directions of relations. We did not filter configuration consisting of two or more dependency relations such as the transitive construction. Thus, we did not compute the relative order between S and O. According to Greenberg's Universal 1, we suppose that SO order is dominant. This could be verified in future work on the same data.

DOI: 10.5334/gjgl.764

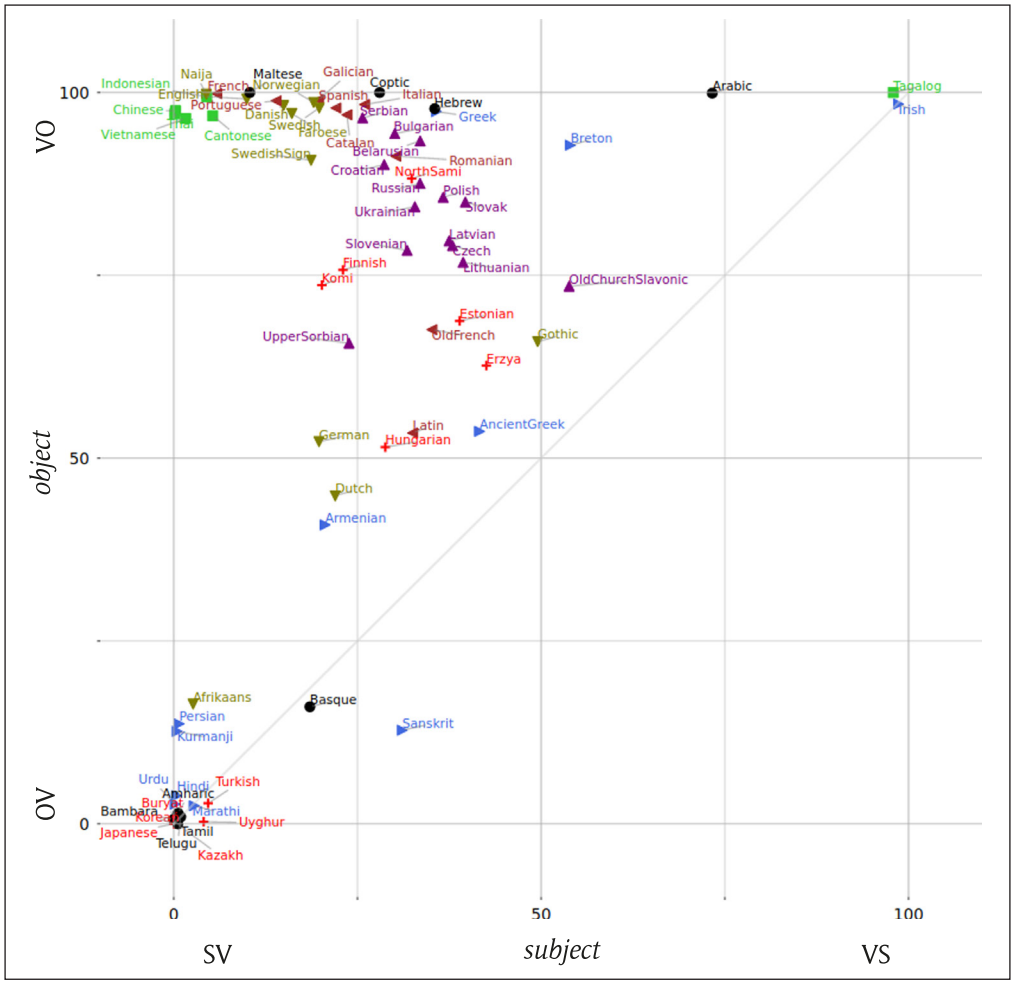

Figure 20 Nominal subject vs. nominal object diagram.

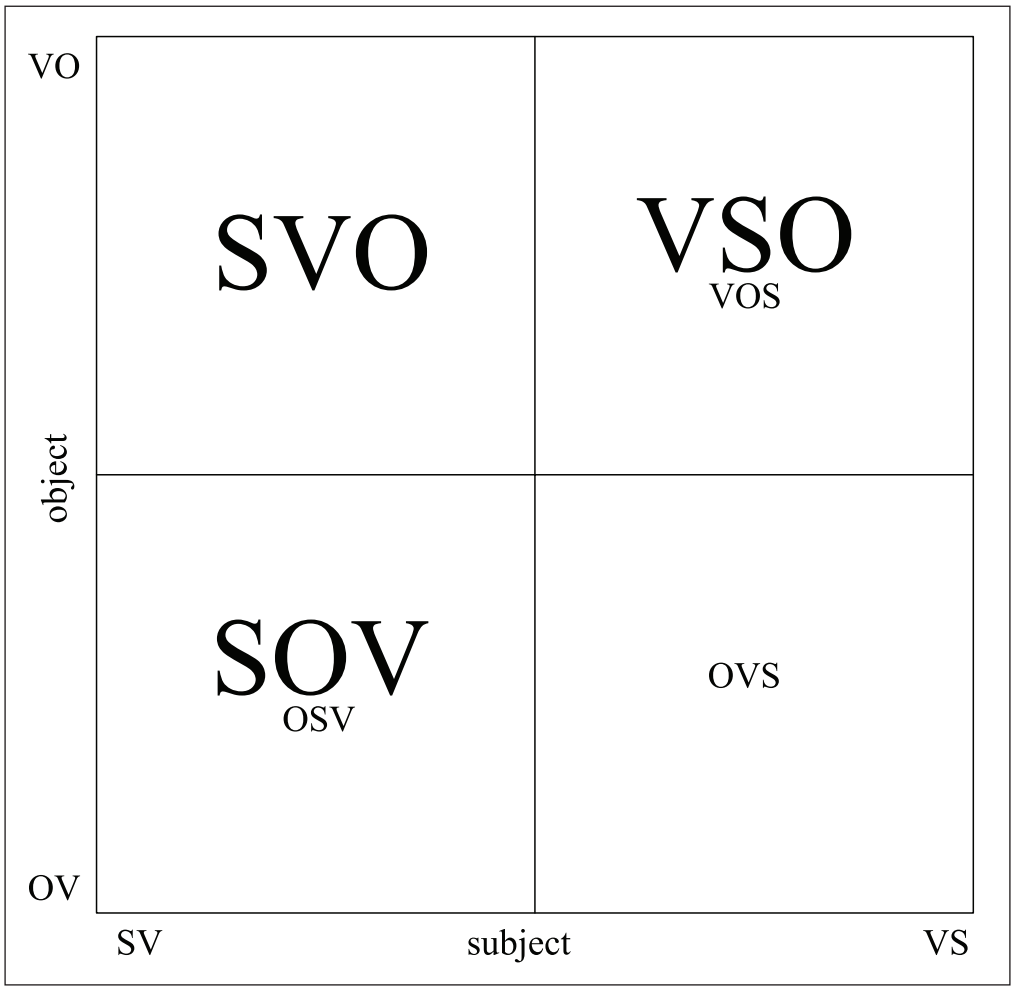

Figure 21 Greenberg's classification on our bi-dimensional scatter.

and Sinitic languages in our sampling. Next are SOV languages, and we have three clear VSO languages (Irish, Arabic and Tagalog).[28]

Some languages are not clearly in one of the quadrants. This conducts us to the second information contained in ***Figure 20***: As shown in ***Figure 22***, the further a language is from a corner of the diagram, the less it can clearly be classified in one of Greenberg's categories and the more it has free or mixed word order.

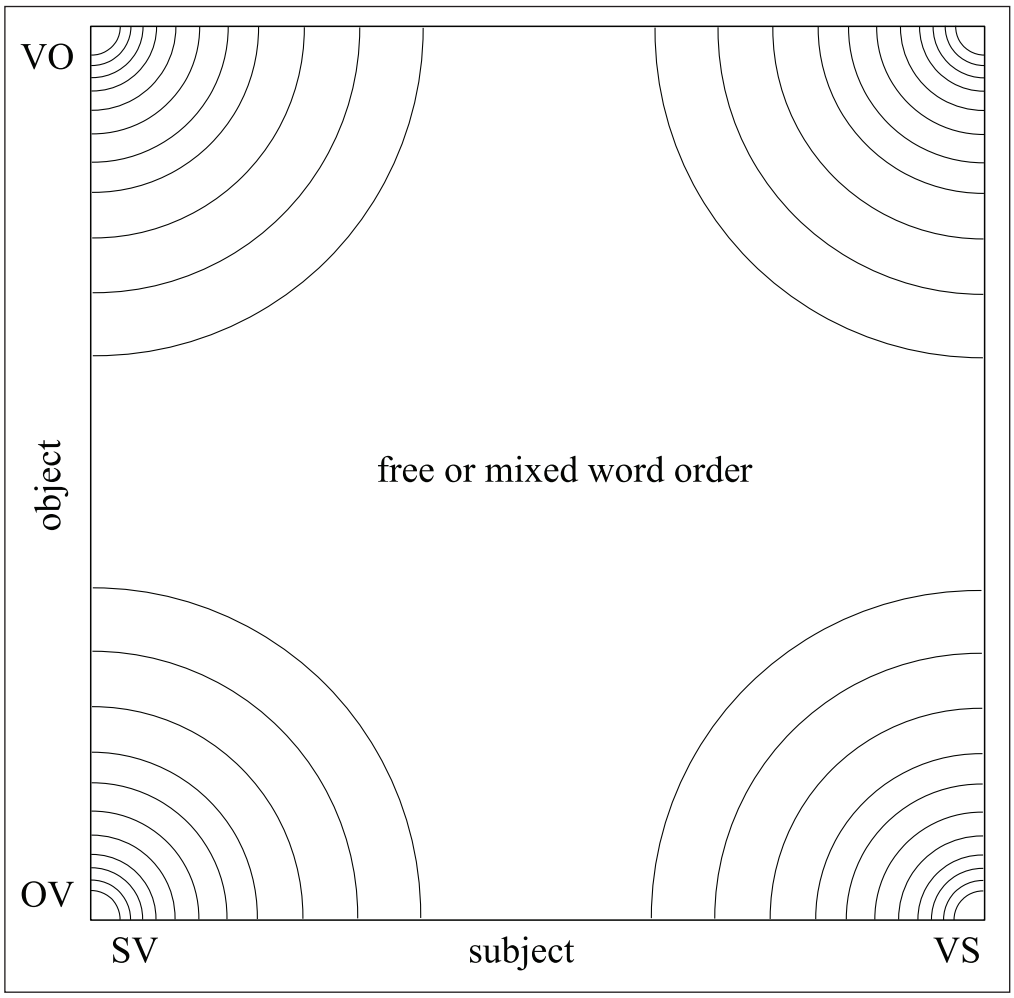

Figure 22 Classification in terms of free/mixed word order.

7 OTHER EXAMPLES OF QUANTITATIVE UNIVERSALS

Section 2.2 introduced a first quantitative universal associated with a particular configuration, the triangular pattern. We then showed the relation between quantitative universals and qualitative universals in Section 2.3. Yet, the triangular pattern is not the only configuration that allows us to state some quantitative universals. We will present the crescent pattern in Section 7.1, the Z-pattern in Section 7.2 and patterns of correlation and non-correlation in Section 7.3.

7.1 THE CRESCENT PATTERN

Looking back on the subject vs. object diagram in ***Figure 20***, we see that the triangular distribution is not completely homogeneous, as the languages are distributed in a crescent shape where the left middle area is empty. This boils down to a new universal:

28 Our diagram is indeed unbalanced with many languages in the bottom-left corner and only two languages in the top-right corner (Irish and Tagalog). Nevertheless, even this sample of languages can be the reflection of another property of languages: The fact that VSO languages tend to have a less rigid word order than SOV languages, which is claimed in Greenberg's Universals 6 and 7:

> "Universal 6. All languages with dominant VSO order have SVO as an alternative or as the only alternative basic order."
>
> "Universal 7. If in a language with dominant SOV order there is no alternative basic order, or only OSV as the alternative, then all adverbial modifiers of the verb likewise precede the verb. (This is the "rigid" subtype of III.)"

Subject-Object Strictness Universal. Languages that strictly place the subject before the verb (VS < 15%), also place the object strictly, either to the left or the right of the verb (VO < 15% or > 85%).

The Subject-Object Strictness Universal can be illustrated by ***Figure 23***.

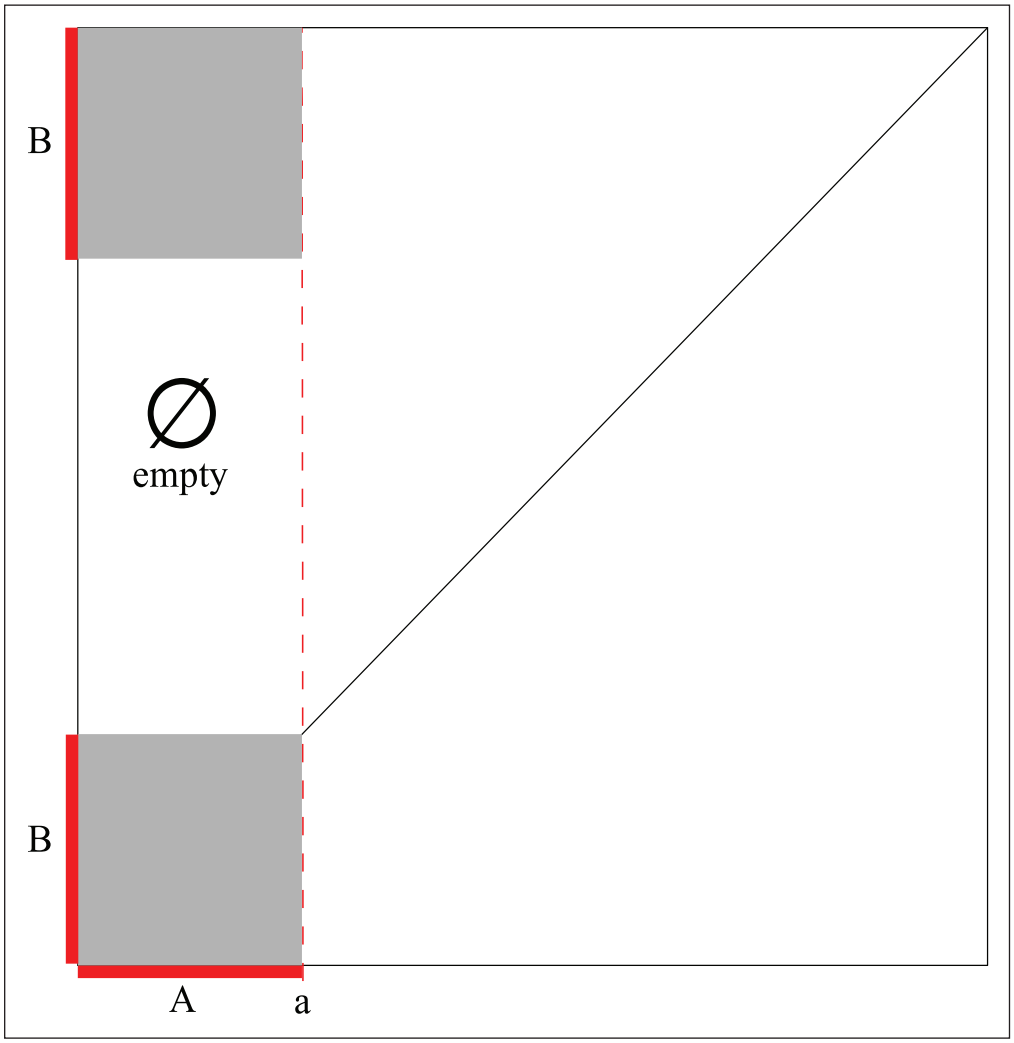

Figure 23 A → B, with A = "*VS < a*" and B = "*VO < a or > 1-a*".

Note that together with the triangular pattern, we actually have a stronger claim

Subject-Object Strictness Universal (generalization). Languages that place the subject strictly (<15% or >85% inversion), also place the object strictly (<15% or > 85%).

The Subject-Object Strictness Universal is a quantitative universal because we can quantify the degrees of strictness, i.e. the zones where languages cannot appear. If, however, we consider that word-order strictness is a sufficiently clearly defined notion, we can interpret the quantitative values as qualitative notions and obtain a qualitative universal:

Qualitative Subject-Object Strictness Universal. Languages that place the subject strictly, also place the object strictly.

7.2 THE Z-PATTERN

Figure 24 gives an example of another remarkable pattern, from the typological point of view, that we call the Z-pattern (cf. ***Figure 25***).[29]

Figure 24 can be viewed as a reformulation of Greenberg's Universals 3 and 4:

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

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

29 An alternative denomination could be "Sigmoid-pattern" or "S-pattern". Although the distribution resembles an inverted letter Z, we prefer the term "Z-pattern" because it emphasizes the three groups of languages that appear in the distribution, as the three straight lines that make the letter Z.

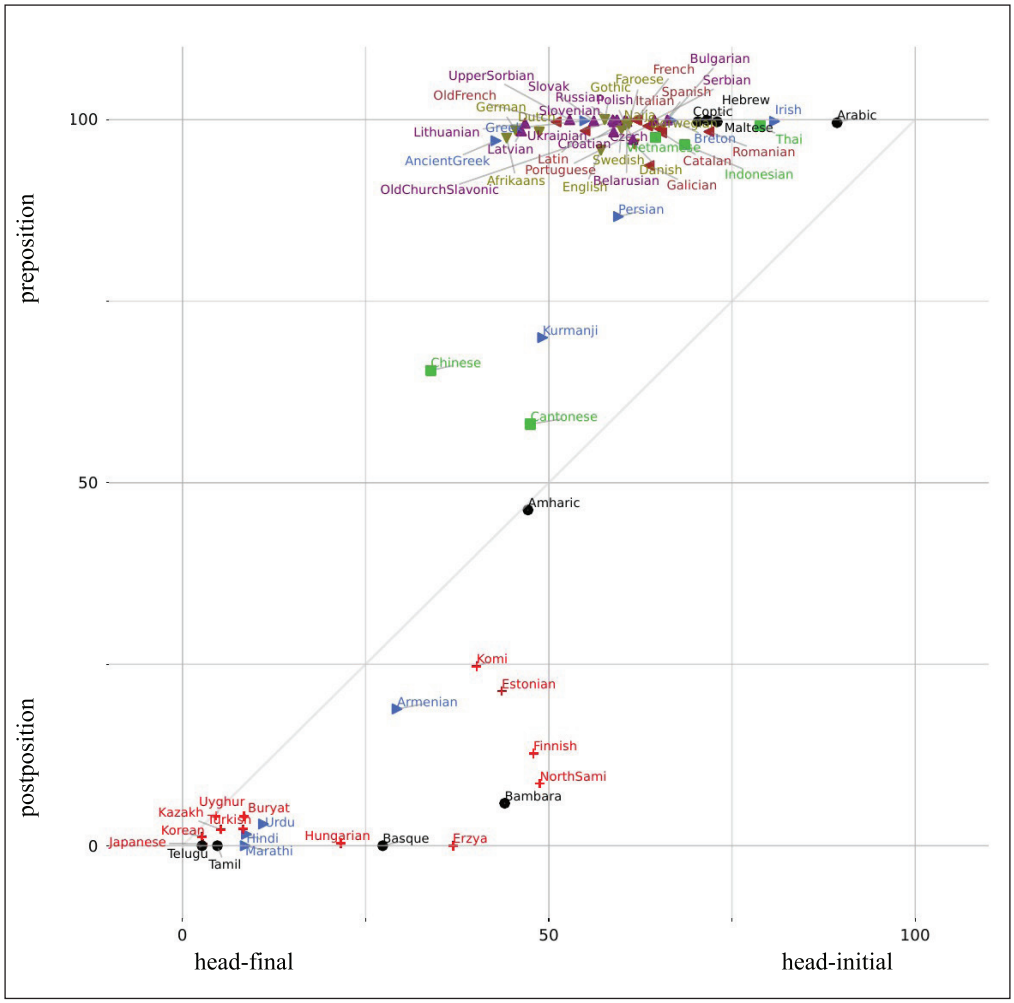

Figure 24 All dependent vs. complements of an adposition diagram.

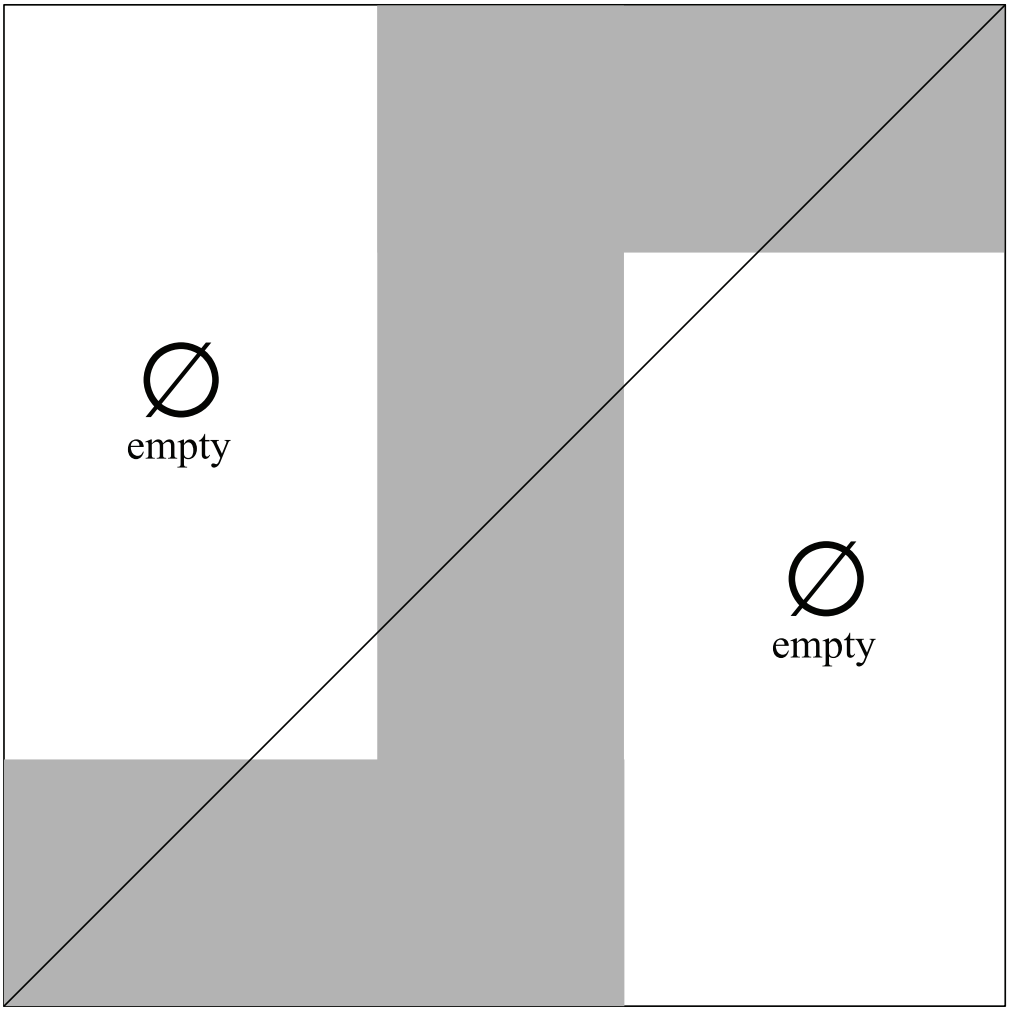

Figure 25 The Z-pattern.

But ***Figure 24*** tells us more:

> Qualitative Adpositional Tendency Universal. Languages that have an (even weak) tendency to be head-initial or head-final have a *strong* tendency to have respectively mainly prepositions or mainly postpositions.

Let us explain this in more details. The Z-pattern of ***Figure 24*** contains three zones (justifying the Z denomination): a central zone, with free or mixed languages (that is languages with 40% to 60% of head-initial dependencies) which can have prepositions or postpositions or both, a zone with languages with less than 40% of head-initial dependencies and more than 90% of postpositions, and a zone with languages with more than 60% of head-initial dependencies and more than 90% of prepositions. Two zones remain empty: we neither have languages with less than 40% of head-initial dependencies and less than 90% of postpositions, nor languages with more than 60% of head-initial dependencies and less than 90% of prepositions.

It comes as no surprise that, if a language has no clear word order tendency, we cannot predict whether it has prepositions or postpositions. But what is remarkable is that as soon as a language has a word order tendency, even a weak one (less than 40% or more than 60% of head-initial dependencies), the tendency is amplified for adpositional objects. Having 40% (or equally 60%) of head-initial dependencies means that the ratio between the dominant direction and the dominated direction is only 1.5, while 90% of prepositions (among all the adpositions) means that the language has 9 times more prepositions than postpositions. In other words, as soon as a language has a ratio higher than 1.5 between right and head-final dependencies, it has a ratio of more than 9 between prepositions and postpositions. We can call this a *strictness reinforcement factor* of 6 (= 9/1.5) between general governor-dependent order and adposition-noun order.

The diagram can also be read in reverse direction, starting from the Y-axis:

> Qualitative Adpositional Tendency Universal (contraposition). If a language has mainly prepositions, it cannot be head-final; if a language has mainly postpositions, it cannot be head-initial; and if a language has neither mainly prepositions nor mainly postpositions, it is a free or mixed word order language.[30]

Note that the distribution of ***Figure 24*** fully justifies to consider that the head of the adposition-noun relation is the adposition because the adposition-noun relation is now a reinforcement factor in the overall dependency direction tendency (cf. the discussion of Section 3 concerning UD vs. SUD).

We observe a similar Z-pattern for the object of an auxiliary in relation to all governor-dependent relations (***Figure 26***), which justifies to consider that the head of the auxiliary-lexical verb relation is the auxiliary. Our pattern can be viewed as a reformulation of Greenberg's Universal 16:

> "Universal 16. In languages with dominant order VSO, an inflected auxiliary always precedes the main verb. In languages with dominant order SOV, an inflected auxiliary always follows the verb."

The Z-pattern of this distribution is more diluted resulting in a "fatter Z". This means that the reinforcement factor is weaker: The central zone is roughly contained between 33% and 66% corresponding to a factor less than 2 between head-final and head-initial dependencies. Languages outside the central zone have less than 20% exceptions in their auxiliary-verb standard direction, corresponding to a factor of 5 between the standard directions and the exceptional directions. Consequently, the strictness reinforcement factor is equal to 2.5 (= 5/2).

7.3 PATTERNS OF CORRELATION AND NON-CORRELATION

One of the patterns that we encounter has no equivalent as an implicational universal precisely because it corresponds to the absence of an implication. As an example consider the relation between the direction of adverbial clauses and dependents of nouns (see ***Figure 27***), where the languages are distributed all over the diagram without a clearly discernible tendency.

30 Note that the third statement is quite astonishing: As soon as the ratio between prepositions and postpositions is less than 9, the language has free or mixed word order.

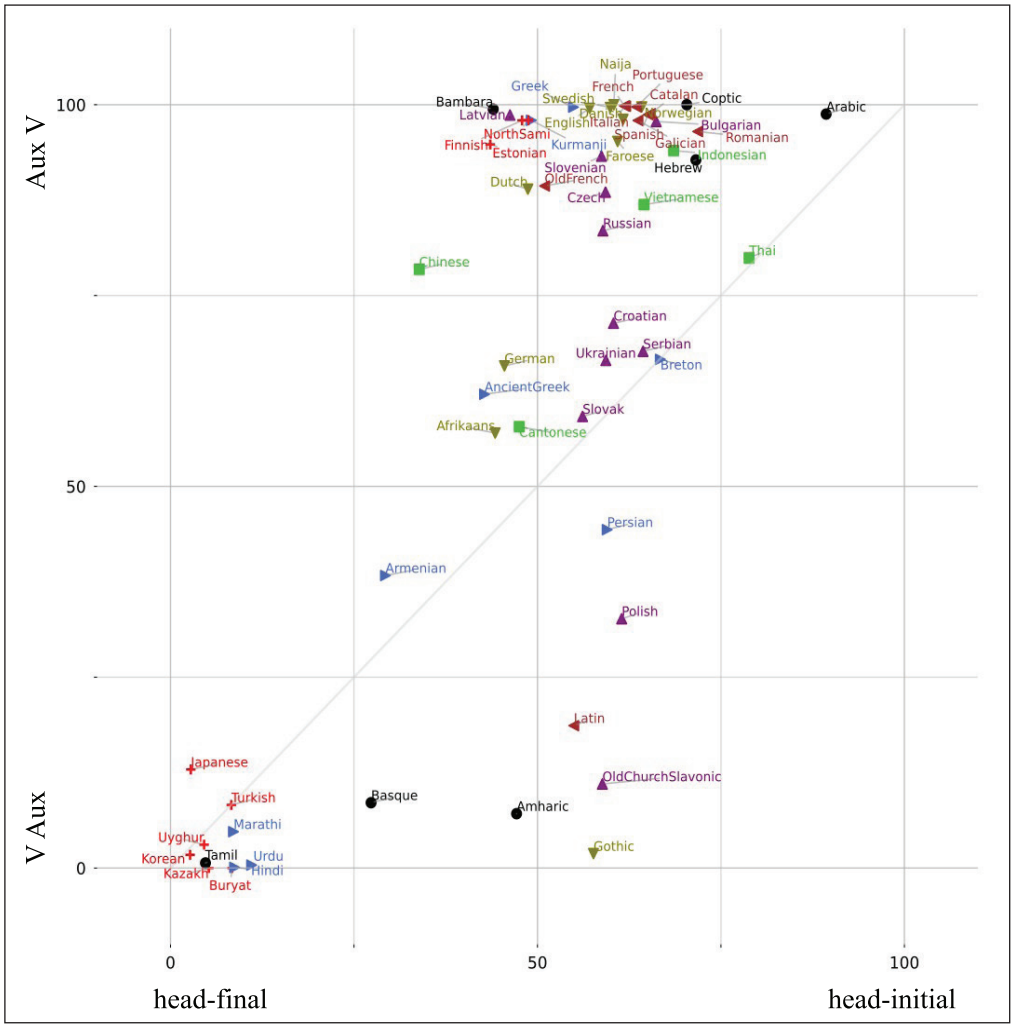

Figure 26 All dependent vs. complement of an auxiliary diagram.

This configuration indicates that knowing how a language places adverbial clauses does not give us enough information about how it places noun dependents, and vice versa.[31]

This can be formulated as the following claim

> Weak Correlation Universal. Even if we know the tendency of a language to place adverbial clauses towards the verb, we cannot deduce from this information any strong tendency concerning the noun-dependent placement and vice versa.

Non-correlation can be gauged by the correlation coefficients, which measure to which extent the data itself (for Pearson) or the rank of the data (Spearman[32]) can be placed on a line.[33] However, both correlation coefficients do not measure uniformity of the data; they measure to which extent the values or rank values of one dimension can predict the values or rank values of the other dimension. In the example of ***Figure 27***, the Spearman correlation coefficient ρ is .40, lower than the .71 for the relation of V pronO compared to V nomO in ***Figure 2*** but higher than the ρ of .33 for the crescent shape of the nominal subject vs. nominal object diagram of ***Figure 20***. Note that even if the correlation is weak, there is nevertheless a significant correlation (for example p-value 0.0005 for the above ***Figure 27***). As our choice of languages is also far from being random (let us recall that half of them are Indo-European), it is difficult to lay down stronger claims, but further investigations might show that many random couples of relations yield a significant correlation. At the very least, it can be noted that we can identify plenty

31 We nevertheless observe one highly specific dependency between adverbial clauses and noun-dependents' directions, which could even be described as an implicational universal: If a language has noun-dependents strictly on the right (beyond 80%), adverbial clauses will be strictly on the right, too, which corresponds to the three nearly empty cases on the right of the diagram of Figure 27 (only Old Church Slavonic is in this region with 77% of dependents of nouns to the right of the noun).

32 The use of the Spearman correlation coefficient rather than Pearson was suggested by an anonymous reviewer because the latter is just a measure of linear association (the true relationship may not be linear)."

33 The absolute value of both correlation coefficients goes from 0 to 1. 1 means that y is a strictly monotonically increasing function of x. If the data points are distributed uniformly on the surface, the value is 0.

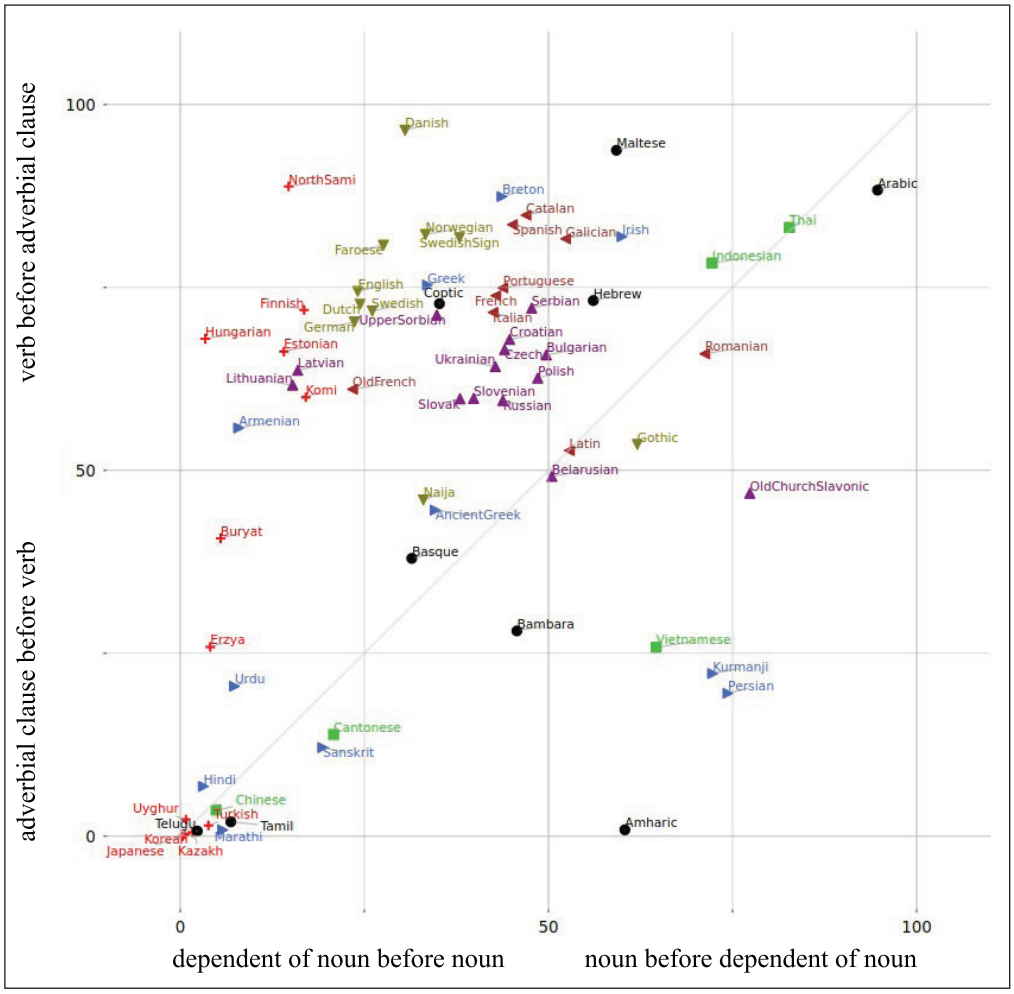

Figure 27 Dependent of a noun vs. adverbial clause diagram. Example of a near-uniform pattern.

of languages that are strictly head-final or strictly head-initial and which therefore cause dependency direction correlations on most couples of relations.

We can compare the configuration of ***Figure 27*** with the one of ***Figure 29***, where the Spearman correlation coefficient is much higher (***Figure 27*** has $\rho = .40$, ***Figure 29*** has $\rho \fallingdotseq .64$). Here we see an almond-shaped distribution of the languages around the diagonal as shown in ***Figure 28***. Head-initial and head-final languages will always be at the extremities of the diagonal. What is interesting is that so-called mixed order languages also have the tendency to give the same left-right distribution to some pairs of relations. This is the case for the direction of dependents of adverbs vs. the direction of dependents of nouns.

> Uncertainty Correlation Universal. The direction of dependents of nouns and of adverbs are correlated in a sense that one value predicts the interval in which the other value can be found, and the less one factor is strict (i.e. approaching 50% from below or above) the more the other factor is uncertain, too (i.e. the greater is the interval restricting the other).

Figure 28 shows the interval I_a in which the average direction of dependents of adverbials can go (Y-axis), for the value a of the average direction of dependents of nouns (X-axis). In the almond pattern, the size of I_a increases and decreases when a goes from 0 to 100%. The X-axis and the Y-axis can be interchanged.

8 CONCLUSION

Commonly, typological universals declare or can be interpreted as the impossibility (or statistical rareness) of languages with certain properties. For example Universal 6 ("All languages with dominant VSO order have SVO as an alternative or as the only alternative basic order", Greenberg 1963) rules out the existence of a language that only allows the VSO order.

As we have shown in the previous subsections, qualitative universals about word order have this type of configurational interpretation. Quantitative universals generalize qualitative universals,

DOI: 10.5334/gjgl.764

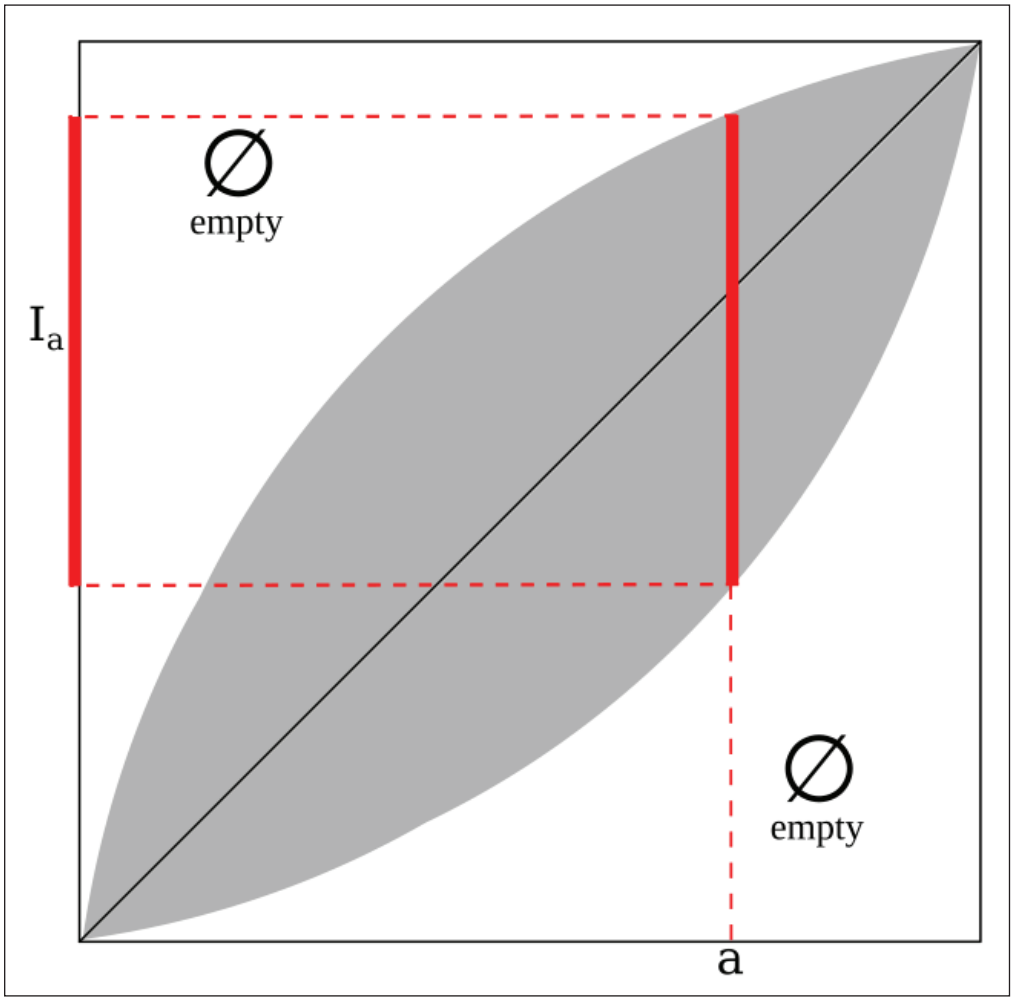

Figure 28 The almond pattern For a being the average direction of a first relation, Ia is the range where the average direction of a second relation can be placed.

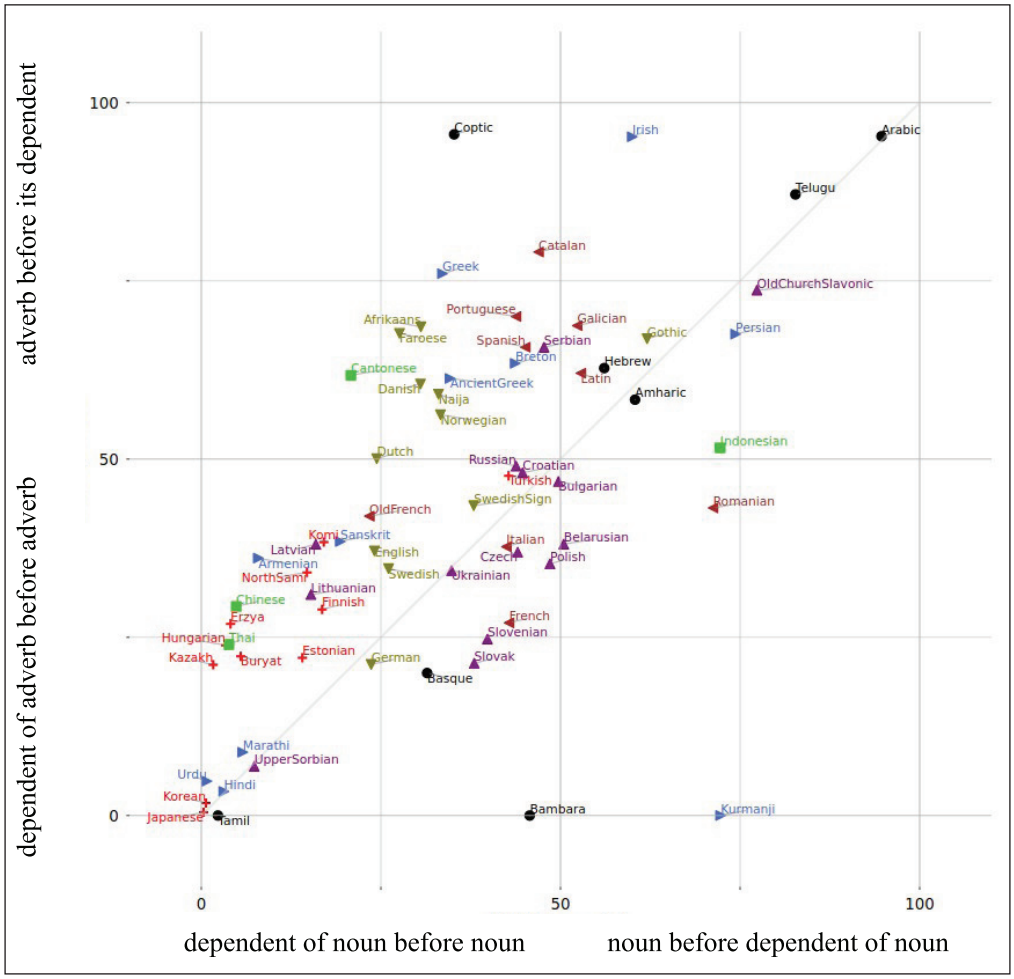

Figure 29 Dependent of a noun vs. dependent of an adverb diagram. Example of an almond pattern.

which only claim that some sub-part of the diagram is empty. More precisely, implicational qualitative universals exclude some configurations of rectangular shape in the corresponding diagram. As we have shown, it is interesting to analyze other shapes of empty zones such as the triangular or the almond pattern, which cannot be expressed easily as a qualitative (implicational) universal.

Beyond confirming, testing, and refining Greenberg's proposed universals, typometrical diagrams allow for a visualization of quantitative language data and can be applied to finding "new" universal patterns.

In this way, we can stake out numerous other kinds of claims about the distribution of languages based on such diagrams. In a way, every diagram tells us something about language properties and universals, including "negative" results – negative in the sense of the absence of a qualitative universal. Notably, we are not only interested in empty zones (cf. absolute qualitative universals). We are also interested in almost empty zones (cf. statistical qualitative universals), more concentrated zones, less concentrated zones, the shape of all these zones, possible attractors in the distribution, etc. In other words, we are interested in the whole distribution of languages in a scatter plot, without limits in the precision of the claims that can be established about language typology. We have identified some patterns – triangle, Z, or almond – for which we can propose an interpretation (cf. the strictness reinforcement factor for the Z-pattern), but a "typology" of relevant patterns remains to be done, as well as a more precise characterization of these patterns: When can we consider that we face a Z-pattern and the strictness reinforcement factor is significant?

Only very few of the possibilities that offer this kind of data analysis has been explored in this article, most questions remain open. We propose to call *typometrics* this open field of the study of the distribution of languages in a distributional scatter diagram based on empirical measures on corpora. Typometrics thus becomes a branch of Quantitative typology, which exists at least since the end of the 1950s (Krámský 1959, 1972, Greenberg 1960, Givón 1983, Fenk-Oczlon & Fenk 1999, Cysouw 2005, Daumé III & Campbell 2007, Liu 2010, Futrell et al. 2015).

The present work only uses simple word-order tendencies taken from treebanks, taking into consideration only measures on simple dependency links between two words. Quantitative universals can be a matter of research based on more complex configurations of trees in the treebank and even use multi-layer analysis of a single sentence. For example, we did not compute the direction between subject and object (SO) because in dependency treebanks this link is not a direct link but would require a more complex search in the trees. We expect similar universals concerning the SO relative order since it is also a gradual tendency rather than an absolute binary feature. More generally, even taking into account complex configurations, the study of word order is not limited to the direction of dependency relations. For example the V2 structure of some languages (in particular Germanic languages) cannot easily be measured in terms of dependency directions.

It is important to point out that all types of syntactic typology have greatly benefited from the community effort to develop and freely distribute homogeneous treebanks for many languages, notably under the impetus of the Universal Dependencies project. At the time of writing, the UD project is only five years old. We can expect a rapid deployment of other forms of syntactic data across languages that might allow for more fine-grained analyses with a more balanced set of languages. Yet, we expect that a typometrical methodology will prove useful for the typological analysis of future treebank data.

ACKNOWLEDGEMENTS

The one- and two-dimensional figures were produced with Matplotlib (Hunter 2007). The labels were adjusted using AdjustText (*https://github.com/Phlya/adjustText*), written by Ilya Flyamer who was kind enough to adapt his code to our needs. We would also like to thank our two reviewers for valuable remarks that conducted us to make several modifications in the presentation of our results.

This work is supported by the National Social Science Fund of China (2018CYY031).

COMPETING INTERESTS

The authors declare that there have been no involvements that might raise the question of bias in the work reported or in the conclusions, implications, or opinions stated.

AUTHOR AFFILIATIONS

Kim Gerdes *orcid.org/0000-0002-9905-0117*
Université Paris Saclay, Lisn (CNRS), Campus universitaire bât 507, F – 91405 Orsay cedex, FR

Sylvain Kahane *orcid.org/0000-0003-3949-8128*
Université Paris Nanterre, Modyco (CNRS), UFR Phillia, F – 92001 Nanterre cedex, FR

Xinying Chen *orcid.org/0000-0002-5052-4991*
Xi'an Jiaotong University, School of Foreign Studies, Xianning West Road 28, CN – 710049 Xi'an, Shaanxi, CN

REFERENCES

Abney, S. P. 1987. *The English noun phrase in its sentential aspect*. Doctoral dissertation, Cambridge: MIT.

Bell, A. 1978. Language samples. In J. H. Greenberg, (ed.), *Universals of Human Language. Vol. 1: Method and Theory*. Stanford: Stanford University Press, 123–156.

Bloomfield, L. 1933. *Language*. New York: Henry Holt.

Chen, X. and K. Gerdes. 2017. Classifying Languages by Dependency Structure. Typologies of Delexicalized Universal Dependency Treebanks. *Proceedings of the conference on Dependency Linguistics (DepLing)*, 54–63.

Croft, W. 1991. *Syntactic categories and grammatical relations: The cognitive organization of information*. University of Chicago Press.

Croft, W. 2002. *Typology and universals*. Cambridge University Press. DOI: *https://doi.org/10.1017/CBO9780511840579*

Croft, W., D. Nordquist, K. Looney and M. Regan. 2017. Linguistic Typology meets Universal Dependencies. *Proceedings of the conference on Treebanks and Linguistic Theories (TLT)*, 63–75.

Cysouw, M. A. 2003. Against implicational universals. *Linguistic Typology* 7(1). 89–101. DOI: *https://doi.org/10.1515/lity.2003.013*

Cysouw, M. A. 2005. Quantitative methods in typology. In *Quantitative Linguistik: ein internationales Handbuch*. De Gruyter, 554–578.

Daumé III, H. and L. Campbell. 2007. A Bayesian model for discovering typological implications. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, 65–72.

De Marneffe, M.-C., B. MacCartney and C. D. Manning. 2006. Generating typed dependency parses from phrase structure parses. *Proceedings of the conference on Language Resources and Evaluation (LREC)*, 449–454.

De Marneffe, M.-C., et al. 2014. Universal Stanford dependencies: A cross-linguistic typology. *Proceedings of the conference on Language Resources and Evaluation (LREC)*, 4585–4592. DOI: *https://doi.org/10.1075/sl.13.2.03dry*

Dik, B. 2010. Language Sampling. In J. J. Song (ed.), *The Oxford Handbook of Linguistic Typology*. Oxford Handbooks.

Dryer, M. S. 1989. Large linguistic areas and language sampling. *Studies in language* 13(2). 257–292.

Dryer, M. S. 1992. The Greenbergian word order correlations. *Language* 68. 81–138. DOI: *https://doi.org/10.1353/lan.1992.0028*

Fenk-Oczlon, G. and A. Fenk. 1999. Cognition, quantitative linguistics, and systemic typology. *Linguistic Typology* 3. 151–178. DOI: *https://doi.org/10.1515/lity.1999.3.2.151*

Ferrer-i-Cancho, R. 2008. Some word order biases from limited brain resources: A mathematical approach. *Advances in Complex Systems* 11(3). 393–414. DOI: *https://doi.org/10.1142/S0219525908001702*

Ferrer-i-Cancho, R. 2015. The placement of the head that minimizes online memory. A complex systems approach. *Language Dynamics and Change* 5(1). 114–137. DOI: *https://doi.org/10.1163/22105832-00501007*

Ferrer-i-Cancho, R. 2016. Kauffman's adjacent possible in word order evolution. In S. G. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Feher and T. Verhoef (eds.), *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11)*.

Ferrer-i-Cancho, R. 2018. Optimization models of natural communication. *Journal of Quantitative Linguistics* 25(3). 207–237. DOI: *https://doi.org/10.1080/09296174.2017.1366095*

Futrell, R., K. Mahowald and E. Gibson. 2015. Quantifying Word Order Freedom in Dependency Corpora. *Proceedings of the conference on Dependency Linguistics (DepLing)*, 91–100.

Gerdes, Kim, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. "Improving Surface-syntactic Universal Dependencies (SUD): surface-syntactic relations and deep syntactic features." In *TLT 2019-18th International Workshop on Treebanks and Linguistic Theories*. 2019. DOI: *https://doi.org/10.18653/v1/*

W19-7814

Gerdes, K. and S. Kahane. 2011. Defining dependency (and constituency). *Proceedings of the conference on Dependency Linguistics (DepLing)*, 17–27.

Gerdes, K. and S. Kahane. 2016. Dependency annotation choices: Assessing theoretical and practical issues of Universal Dependencies. *Proceedings of the Linguistic Annotation Workshop* (*LAW*), 131–140. DOI: *https://doi.org/10.18653/v1/W16-1715*

Gerdes, K., B. Guillaume, S. Kahane and G. Perrier. 2018. SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. *Proceedings of the Universal Dependencies Workshop (UDW)*, 66–74. DOI: *https://doi.org/10.18653/v1/W18-6008*

Givón, T. 1983. *Topic continuity in discourse: A quantitative cross-language study*, vol. 3, Benjamins. DOI: *https://doi.org/10.1075/tsl.3*

Greenberg, J. H. 1954 [I960]. A quantitative approach to the morphological typology of language. In R. F. Spencer (ed.), *Method and Perspective in Anthropology: Papers in Honor of Wilson D. Wallis*, 192–220. Minneapolis: University of Minnesota Press. Reprinted in *International Journal of American Linguistics* 26(3). 178–194.

Greenberg, J. H. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg (ed.), *Universals of grammar*, 73–113. Cambridge: MIT.

Haspelmath, M., M. S. Dryer, D. Gil and B. Comrie. 2005. *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library.

Hawkins, J. A. 1983. *Word order universals: Quantitative analyses of linguistic structure*. New York: Academic Press.

Hudson, R. 1984. *Word Grammar*. Oxford: Basil Blackwell.

Hudson, R. 1998. Functional control with and without structure-sharing. *Typological studies in language* 38. 151–170. DOI: *https://doi.org/10.1075/tsl.38.11hud*

Hunter, J. D. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* 9(3). 90–95. DOI: *https://doi.org/10.1109/MCSE.2007.55*

Isačenko, A. V. 1939. Versuch einer Typologie der slavischen Sprachen. *Linguistica Slovaca* 1. 64.

Justeson, J. S., and L. D. Stephens. 1984. On the relationship between the numbers of vowels and consonants in phonological systems. *Linguistics*, 22. 531–545. DOI: *https://doi.org/10.1515/ling.1984.22.4.531*

Krámský, J. 1959. A quantitative typology of languages. *Language and speech* 2(2). 72–85. DOI: *https://doi.org/10.1177/002383095900200202*

Krámsky, J. 1972. On Some Problems of Quantitative Typology of Languages on Acoustic Level. *Prague Studies in Mathematical Linguistics* 3. 15–26.

Lehmann, W. P. 1973. A Structural Principle of Language and its Implications. *Language* 49. 47–66. DOI: *https://doi.org/10.2307/412102*

Li, C. N. and S. A. Thompson. 1989. *Mandarin Chinese: A functional reference grammar*. Univ of California Press.

Liu, H. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua* 120(6). 1567–1578. DOI: *https://doi.org/10.1016/j.lingua.2009.10.001*

Liu, H., Y. Zhao and W. Li. 2009. Chinese syntactic and typological properties based on dependency syntactic treebanks. *Poznań Studies in Contemporary Linguistics* 45(4). 509–523. DOI: *https://doi.org/10.2478/v10010-009-0025-3*

Mel'čuk, I. A. 1988. *Dependency syntax: theory and practice*. New York: SUNY press.

Mithun, M. 1987. Is basic word order universal? In R. Tomlin (ed.), *Grounding and Coherence in Discourse* [Typological Studies in Language, 11]. Amsterdam: John Benjamins. 281–328. Reprinted 1992 in Payne, D. (ed.), *The Pragmatics of Word-Order Flexibility* [Typological Studies in Language 22]. Amsterdam: John Benjamins. 15–61. DOI: *https://doi.org/10.1075/tsl.22.02mit*

Næss, Å. 2006. Bound Nominal Elements in Äiwoo (Reefs): A Reappraisal of the "Multiple Noun Class Systems". *Oceanic Linguistics* 45(2). 269–296. DOI: *https://doi.org/10.1353/ol.2007.0006*

Nichols, J. 1992. *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press. DOI: *https://doi.org/10.7208/chicago/9780226580593.001.0001*

Nivre, J., M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajic, C. D. Manning, R. T. McDonald, S. Petrov, S. Pyysalo, N. Silveira and R. Tsarfaty. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. *Proceedings of the conference on Language Resources and Evaluation (LREC)*, 1659–1666.

Osborne, T. and K. Gerdes. 2019. The status of function words in dependency grammar: A critique of Universal Dependencies (UD). *Glossa: a journal of general linguistics* 4(1): 17. 1–28. DOI: *https://doi.org/10.5334/gjgl.537*

Östling, R. 2015. *Bayesian Models for Multilingual Word Alignment*. Doctoral dissertation, Stockholm University.

Perfors, A., J. Tenenbaum, E. Gibson and T. Regier. 2010. How recursive is language? A Bayesian exploration. *Recursion and human language*, 159–175. DOI: *https://doi.org/10.1515/9783110219258.159*

Perkins, R. D. 1989. Statistical techniques for determining language sample size. *Studies in Language* 13(2). 293–315. DOI: *https://doi.org/10.1075/sl.13.2.04per*

Perkins, R. D. 2001. Sampling procedures and statistical methods. In M. Haspelmath, E. König, W. Oesterreicher, W. Raible (eds.), *Language Typology and Language Universals: An International Handbook* 1. 419–434. Berlin: De Gruyter.

Petrov, S., D. Das and R. McDonald. 2012. A universal part-of-speech tagset. *Proceedings of the conference on Language Resources and Evaluation (LREC),* 2089–2096.

Sapir, E. 1985. *Selected Writings in Language, Culture, and Personality*, University of California Press.

Schmidt, P. W. 1926. *Die Sprachfamilien und Sprachenkreise der Erde: Atlas von 14 Karten.* Winter, Heidelberg.

Song, J. J. 2001. *Linguistic Typology: Morphology and Syntax*. Pearson Education.

Stein, A. and S. Prévost. 2013. Syntactic annotation of medieval texts: the Syntactic Reference Corpus of Medieval French (SRCMF). In P. Bennett, M. Durrell, S. Scheible and R. Whitt (eds.), *New Methods in Historical Corpus Linguistics,* 75–82. Corpus Linguistics and International Perspectives on Language, CLIP Vol. 3. Tübingen: Narr.

Steinthal, H. 1850. *Klassifikation der Sprachen, dargestellt als die Entwickelung der Sprachidee* (ib. 1850), which appeared in 1860 under the title *Charakteristik der Hauptsächlichre*, edited and enlarged by the author and Misteli, as the second volume of *Abriss der Sprachwissenschaft* (ib. 1893).

Sun, C. F. and T. Givón. 1985. On the so-called SOV word order in Mandarin Chinese: A quantified text study and its implications. *Language*, 329–351. DOI: *https://doi.org/10.2307/414148*

Tesnière, L. 1945. *Petite grammaire russe*. Paris: Didier.

Tesnière, L. 1959. *Eléments de syntaxe structurale*. Paris: Klincksieck. [Transl. by Osborne, T., Kahane, S. (2015) *Elements of structural syntax*. Benjamins].

Weil, H. 1844. *De l'ordre des mots dans les langues anciennes comparées aux langues modernes*, Doctoral dissertation. Paris; 3e éd., 1879. Réimpr. de la 3e éd., Paris: Didier érudition, 1991, IX–101, ISBN 2-86460-166-4.

Whaley, L. J. 1996. *Introduction to typology: the unity and diversity of language*. Sage Publications. DOI: *https://doi.org/10.4135/9781452233437*

Wong, T. S., K. Gerdes, H. Leung and J. Lee. 2017. Quantitative Comparative Syntax on the Cantonese-Mandarin Parallel Dependency Treebank. *Proceedings of the conference on Dependency Linguistics (Depling),* 266–275.

Zeman, D. 2008. Reusable Tagset Conversion Using Tagset Drivers. *Proceedings of the conference on Language Resources and Evaluation (LREC),* 213–218.

TO CITE THIS ARTICLE:
Gerdes, Kim, Sylvain Kahane and Xinying Chen. 2021. Typometrics: From Implicational to Quantitative Universals in Word Order Typology. *Glossa: a journal of general linguistics* 6(1): 17. 1–31. DOI: *https://doi.org/10.5334/gjgl.764*

Submitted: 30 July 2018
Accepted: 14 July 2020
Published: 09 February 2021