Rediscovering Greenberg's Word Order Universals in UD

Kim Gerdes
LPP (CNRS)
Sorbonne Nouvelle, France
kim@gerdes.fr

Sylvain Kahane
Modyco (CNRS)
Université Paris Nanterre, France
sylvain@kahane.fr

Xinying Chen
University of Ostrava, Czech Republic
Xi'an Jiaotong University, China
xy@yuyanxue.net

Abstract

This paper discusses an empirical refoundation of selected Greenbergian word order universals based on a data analysis of the Universal Dependencies project. The nature of the data we work on allows us to extract rich details for testing well-known typological universals and constitutes therefore a valuable basis for validating Greenberg's universals. Our results show that we can refine some Greenbergian universals in a more empirical and accurate way by means of a data-driven typological analysis.

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.

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.

With the appearance of larger sets of treebanks, research has begun to test existing word order typology claims or hypothesis 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 at 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). 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]

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 140 treebanks of about 85 different lan-

[1] The development of treebanks is a 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 as the available translations of the New Testament in 986 languages. Such methods are not the subject of our paper but it is worth considering them for future works, knowing 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).

guages. UD has been developed with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a perspective of language typology (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., 2014), Google universal part-of-speech tags (Petrov et al., 2011), 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".[2]

One outstanding advantage of using this data set for language typology studies is the sheer size of the data set: we worked on UD 2.2, which includes 110 treebanks in over 70 languages. As all UD treebanks use the same annotation scheme, the database provides rich informative evidence that can be easily compared and interpreted across authentic texts of various languages.

Following Liu (2010), this paper aims to test well-known existing word-order universals based on the data analysis of a set of uniformly annotated texts of diverse languages. Even though the 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), 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.

The paper is structured as follows. In Section 2, we introduce dependency treebanks and explain amendments of the current annotation scheme that were necessary to obtain typologically relevant data. In Section 3, we discuss and compare some of Greenberg's (1963) Universals with our results. In the conclusion, we discuss the potential of using UD treebanks for future typological studies.

2 Material and Methods

Dependency trees encode the relations between words by means of an arrow that goes from the head to another element of the phrase (Tesnière 1959 [2015], Mel'čuk 1988). The direction of these arrows, which indicates the relative position of a phrase towards its governor, is the base of our measures. The dependency analysis[3] of the sentence "*Syntactic dependency treebanks help you understand typology*" has three head-initial relations (for example *understand* → *typology*) and three head-final relations (for example *treebanks* ← *help*), see Figure 1 for a graphical illustration. 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.

[2] UD introduction page http://universaldependencies.org/introduction.html consulted in August 2017.

[3] 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.

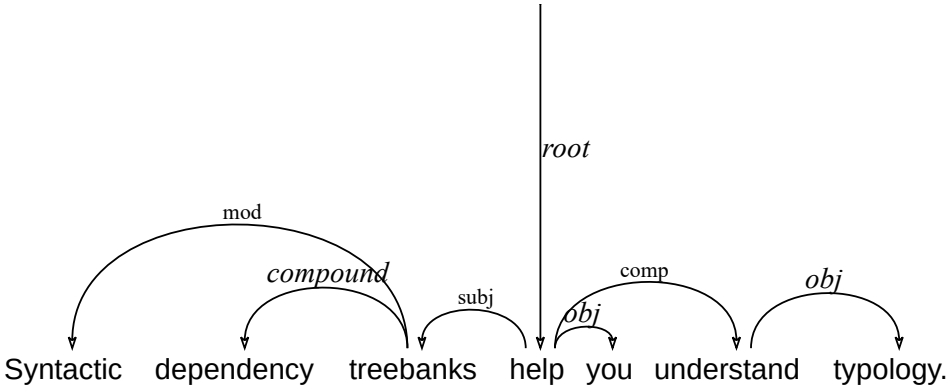

Figure 1: Example of an ordered dependency tree

Our study is based on Surface-Syntactic Universal Dependencies (SUD), a variant of the UD annotation scheme (Gerdes et al. 2018). SUD is better suited for word order studies as it is based on distributional criteria whereas UD favors relations between content words. 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. 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.

From these 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. For instance, we were interested in a separation of the object relation (*comp:obj* in SUD) into V pronO (*VERB-comp:obj>PRON*) and V nomO (*VERB-comp:obj>NOUN*) (pronominal vs nominal object) and of the subject relation into *-subject>PRON* and *-subject>NOUN* (pronominal vs nominal subject). For each relevant *POS-relation>POS* triple (as well as *POS-relation>*, *-relation>POS*, and *relation*) and each of the UD languages (merging all treebanks of the same language),[4] we computed the number of head-initial and head-final dependencies.

The scatter plot of Figure 2 shows the percentage of head-initial head-daughter dependencies, that is, dependencies that link a head with a constituent that is subordinated to it. We do not consider coor-

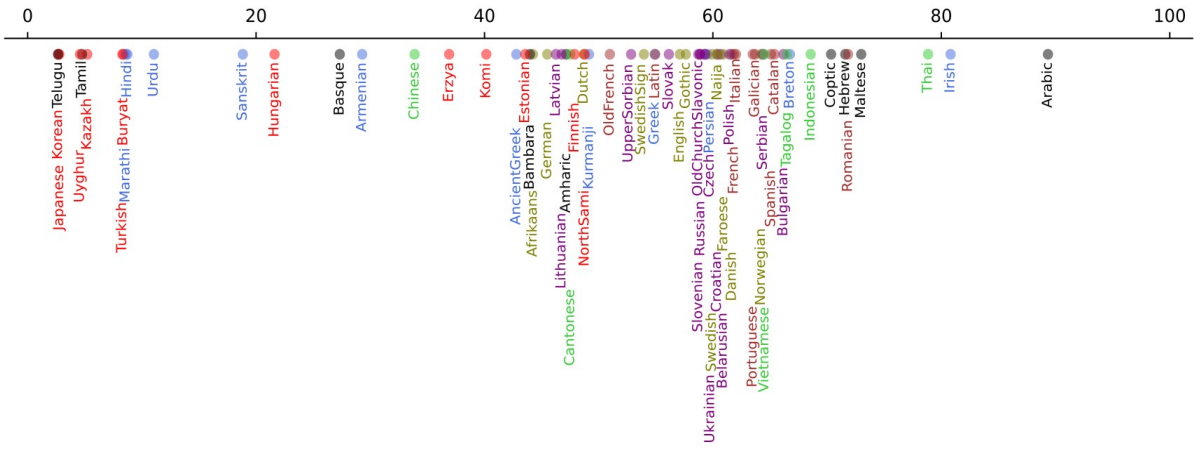

Figure 2: Percentage of head-initial head-daughter dependency relations in the UD treebanks ranging from 3% for Japanese to 89% for Arabic.

[4] 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.

dination for instance, although coordination can also be encoded with the same formal device of "dependency". 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 list of SUD/UD relations we eliminated includes *conj*, *appos*, *reparandum*, *fixed*, *flat*, *compound*, *list*, *parataxis*, *orphan*, *goeswith*, *punct*, *root*, *dep*, and *clf*. We decided to keep the *det* relation for determiners, even if 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 relation is that it has been used even in some languages, such as Japanese, which do not have clear determiners, for closed classes of adjectives which have a similar meaning as English determiners. (We consider that a language has clear determiners when the noun cannot be used alone in some argument positions.)

3 Results and Discussion

"*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 Greenbergian universal means that languages with dominant ADJ-NOUN order (that is, with a dominant head-final *NOUN-dependent>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-dependent>ADJ* relation in Figure 3, 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%.[5] 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.[6]

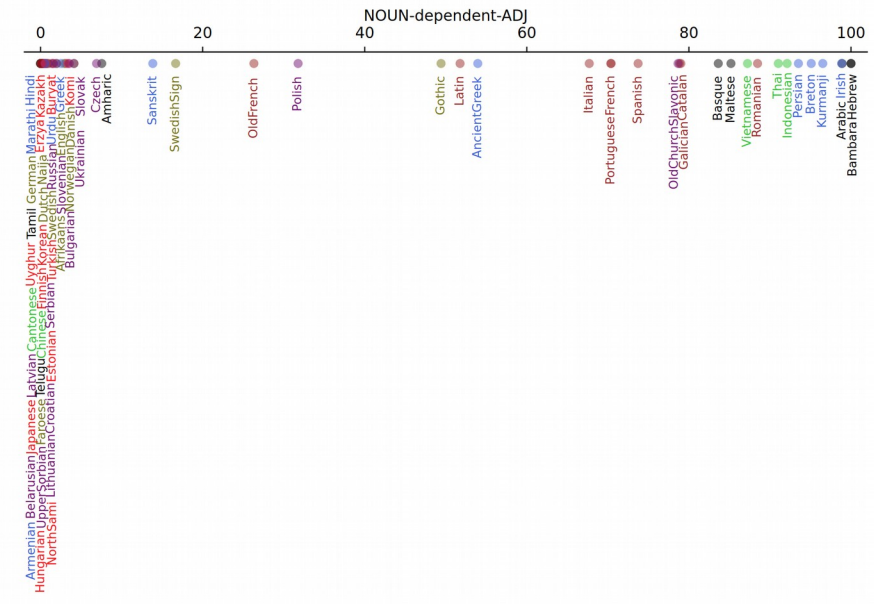

Figure 3: Language distribution for the direction of the NOUN-dependent-ADJ relation.

[5] 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.

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

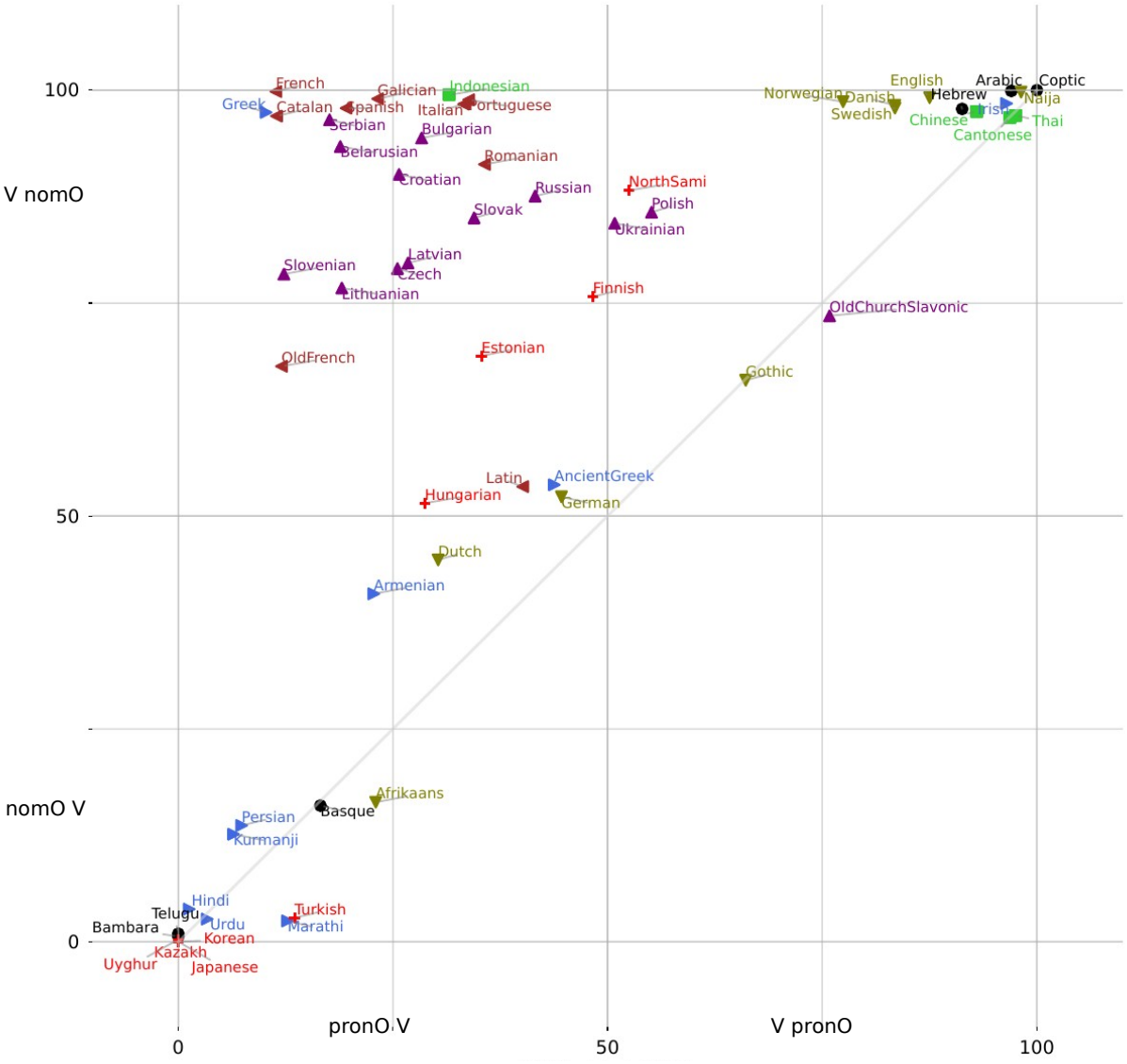

Figure 4: Scatter plot of the percentage of V pronO compared to V nomO

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.

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 NOUN-ADJ order (i.e. with a *NOUN-dependent>ADJ* relation 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 distribution of languages considering a special feature (the distribution of *ADJ*s to-

wards the NOUN) in qualitative terms, which is not straightforward. We believe that a diagram such as Figure 3 can be a more satisfying alternative to such descriptions since it provides many more details.

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

Universal 25 is 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". Therefore, 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.

Let us see now how Universal 25 is related with the scatter plots in Figure 4. We can remark that Greenberg's statement is not totally clear. What does it 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 might be, we can translate the statements of Universal 25 into more satisfying, quantitative statements and see whether the implication is verified on our data. 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 5a. Let us recall that the negation of a property A → B is A & ¬B. Thus, Universal 25' claims that there is no language with V pronO ≥ 75% and V nomO < 75%, that is, that the corresponding rectangle in Figure 5a (hatched in gray) is empty of any language.

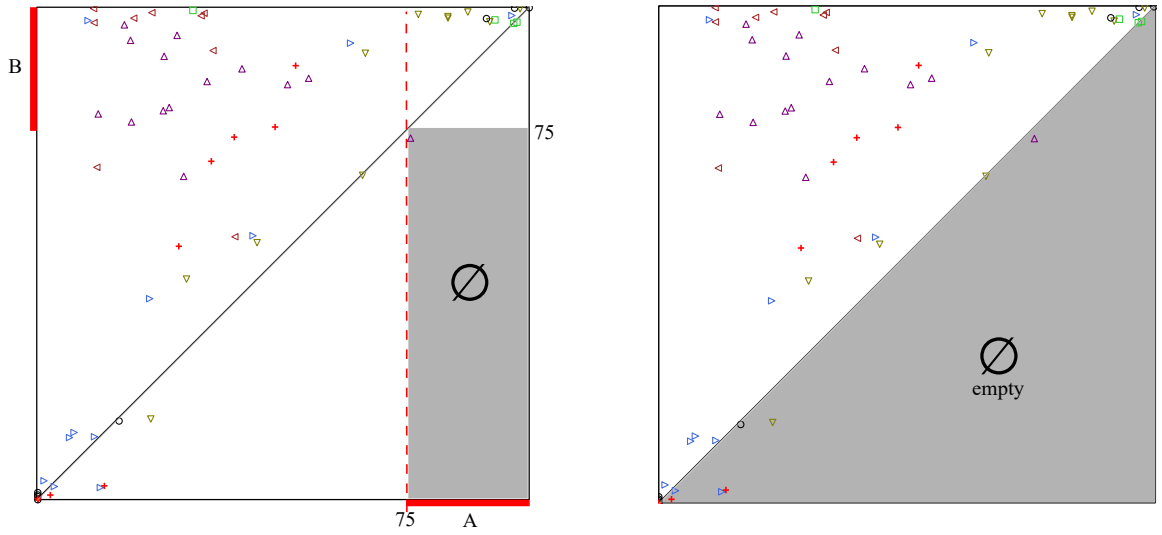

Figure 5: Universal 25'

a. V pronO ≥ 75% → V nomO ≥ 75%
b. V nomO ≥ V pronO (that is, for every a, V pronO ≥ a → V nomO ≥ a)

Yet, we do not know the relevant threshold *a*. 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. We believe that qualitative universals such as Universal 25 (V pronO → V nomO) should be interpreted by means of quantitative universals such as "V pronO ≥ *a* → V nomO ≥ *a*" for some *a*, so that we can obtain more accurate claims for language universals.

Another direction is to not consider a particular threshold at all. For our example, we do not need to propose a threshold, because for almost all languages, we have V pronO ≥ *a* → V nomO ≥ *a* for every *a*, which is equivalent to V nomO ≥ V pronO, which gives us the following universal:

Universal 25": Almost every language has a higher proportion of nominal objects than of pronominal objects on the right of the verb.

This last statement is verified on our data and corresponds to a near empty triangular form in Figure 5b. Universal 25" has no equivalent in terms of qualitative universals à la Greenberg. Thus working with quantitative data opens up the door to completely new universals.

4 Conclusion

Our results roughly confirm Greenberg's word order universals 19 and 25 in that these two universals are coherent with the empirical analysis based on the treebanks of more than 70 languages in UD. However, we also can see obvious limitations of Greenberg's universals in our discussion. To be more specific, Greenberg's universals remain to a certain extent vague, since they are purely implicational, and should be updated into a more accurate and empirically verifiable description, going along with the growing treebank data resources and computing power that are available our days.

In this pilot study, we present one way of accomplishing this task. Commonly, typological universals declare or can be interpreted as the impossibility (or statistical rareness) of languages with certain properties. As we have shown in our study, some of Greenberg's universals about word order have this type of configurational interpretation. By introducing more informative quantitative descriptions with broader conditions, we can establish more sophisticated quantitative universals which provide more accurate descriptions and actually can generalize Greenberg's universals. For example, Universal 25 is in fact (almost) true for every *a*, giving us a triangular pattern, which paves the way for other types of universals, where we would actually describe universal restrictions on human languages as the shapes that the clouds of language take on scatterplots of various properties.

Reference

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

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

Bell, A. (1978). Language samples. In J. H. Greenberg, C. A. Ferguson, E. A. Moravcsik (eds.), *Universal off Human Languages, Vol. I: Method-Theory, 123-156.*

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

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.

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

De Marneffe, M.-C., T. Dozat, N. Silveira, K. Haverinen, F. Ginter, J. Nivre, C. D. Manning (2014). Universal Stanford dependencies: A cross-linguistic typology. *Proceedings of LREC*. Vol. 14.

Dik, B. (2010). Language Sampling, in Song, J.J. (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.

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

Gerdes, K., B. Guillaume, S. Kahane, G. Perrier. (2018). SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. *Proceedings of Universal Dependencies Worksho*p.

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*, Cambridge: MIT, 73-113.

Haspelmath, M., M. S. Dryer, D. Gil, 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.

Lehmann, W. P. (1973) A Structural Principle of Language and its Implications. *Language*, 49, 47-66.

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

Liu, H., Y. Zhao, W. Li (2009). Chinese syntactic and typological properties based on dependency syntactic treebanks. *Poznań Studies in Contemporary Linguistic*s, 45(4), 509-523.

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 in D. Payne (ed.) (1992). *The Pragmatics of Word-Order Flexibility* [Typological Studies in Language, 22], Amsterdam: John Benjamins. 15-61.

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

Östling, R. (2015) *Bayesian Models for Multilingual Word Alignmen*t, Doctoral dissertation, Stockholm University.

Petrov, S., D. Das, R. McDonald (2011). A universal part-of-speech tagset. *arXiv preprint*, arXiv:1104.2086.

Perkins, R. D. (1989), Statistical techniques for determining language sample size. *Studies in Language*,13(2), 293-315.

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*. Vol. 1. Berlin: De Gruyter, 419-434.

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

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

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

Whaley, L. J. (1996) *Introduction to typology: the unity and diversity of language*. Sage Publications.

Zeman, D. (2008) Reusable Tagset Conversion Using Tagset Drivers. *Proceedings of LREC*.