Testing Greenberg's Linguistic Universals on the Universal Dependencies Corpora using a Graph Rewriting tool

Bibliographic Report

by

Arash Morteza | Seweryn Polec | Ludivine Robert

1st year MSc in Natural Language Processing

UE 705 EC1 Supervised Project

20 december 2019

Supervised by Karën Fort and Bruno Guillaume

Acknowledgments

We would like to thank our supervisors Karën Fort and Bruno Guillaume for their support and guidance for this project and providing us with correction advice. We extend our gratitude to Maxime Amblard for marking and giving his advice on the structure and contents of the bibliography report.

Contents

List of Figures

1. Introduction

This bibliography report contains an investigation into a theme of linguistic universals with specific focus on the well-known [Greenberg, 1963] paper on universals of language.

We wrote this report as part of our first year requirement for the Natural Language Processing (NLP) programme at the University of Lorraine. It is a very ambitious project for first year master students but our team collectively speaks fluently 5 languages (albeit all from the Indo-European family branch), so we can utilise our knowledge in this domain for the project evaluation.

We chose this topic due to our personal interest in languages and how languages are linked on a fundamental level. How certain features are shared by all languages independent of their geographical location and social change. This area of study expands to many different disciplines, among them psycho-linguistics, socio-linguistics, typology, computational linguistics, etc and as such is an important field of study that we are glad to have the opportunity to delve into.

We aim to produce quantifiable results for the tested Greenberg Universals (GU) using resources and tools we have in our hands, that are Univerals Dependencies (UD) and a Graph Rewriting tool (GREW).

In the first part of this report, we will introduce a short story of linguistics. Then, we present the [Greenberg, 1963] paper. In the second part, we describe UD and GREW. In the third part, we illustrate attempts that were already made to address this subject. In the end, we demonstrate the use of the tools and how we will proceed for the application phase of our project.

2. Linguistics through the ages

2.1 A quick overview of Linguistics

2.1.1 What is Linguistics?

Linguistics is a field of science which studies the human languages methodically. It is about finding the answers of principal questions: What is a language? How does it work and how is it constructed? How do humans communicate with each other? How did language evolve and change?

Human language is intertwined with biology and sociology[1]. Linguistics aim to expound the internal features of languages within its sub-studies such as morphology, syntax, phonetics, phonology, semantics. In this bibliography report, we will focus on NLP which is a sub-domain of computational linguistics and link it to historical studies of language universals.

2.1.2 Genesis and evolution

There were turning points in the history of Linguistics. The first revolution happened in 18th century and it was based on "Historicism", a thinking approach that emphasises a particular context, such as the historical period, geographic location or local culture. For instance it is believed that the basis of most European languages comes from an Indo-European mother tongue. Subsequently, a debate over the origin of languages, the family of languages, and whether these languages originate from several groups has begun. This revolution was complemented by the emergence of a group of linguists, later called the Nodostorians[2] they were

[1] The study of human social behavior

[2] The Neogrammarians were a German school of linguists, originally at the University of Leipzig, in the late 19th century who proposed the Neogrammarian hypothesis of the regularity of sound change.https://en.wikipedia.org/wiki/Neogrammarian(dec 15, 2019)

looking for rules for language alterations.

The second revolution in linguistics began as the concept ”Constructivism”, conceived of by Ferdinand de Saussure[3], he examined language as a system of universals and separated "synchronous linguistics" from "historical linguistics", and studied the language as a system of signs. The vision of Saussure has influenced linguistics for fifty years. Most of the linguistic activities of the constructivist period concern the first half of the twentieth century with a focus on word structure and the phonetic system.

The third revolution in the field of linguistics was initiated by Noam Chomsky[4], he sharply criticised and demonstrated inefficiencies of ”behavioral psychology” which attempts to describe language learning as a behavioral imitation. Chomsky presented his theories in universal grammar and child learning which created an important theoretical framework in linguistics called ”Generative order”. He believed that the language principles in humans are inherent, and genetically programmed in human brain at birth.

From about the 1960’s, some linguists such as Joseph Greenberg proposed new descriptions of linguistics. The fourth revolution of linguistic studies aligns with sociology, which contrasts with Chomsky’s approaches. That was the trigger that paved the way for the comprehensive study of a linguistic universal, which is a schema of the principles that consistently occurs throughout the natural languages. The particles like verbs and nouns, or consonants and vowels for spoken languages. It also concerns other mutual aspects of languages such as grammar and semantics.

2.2 Greenberg and his universals

2.2.1 The life of Greenberg

This section is based on a summary of the article [Croft, 2001].

Joseph Harold Greenberg (May 28, 1915 – May 7, 2001) was an American linguist, well-known for his studies in linguistic typology and the genetic classification of languages. Throughout his career he published a lot of articles and books

[3]Ferdinand de Saussure was a Swiss linguist and semiotician. His ideas laid a foundation for many significant developments in both linguistics and semiology in the 20th century.https://en.wikipedia.org/wiki/Ferdinand_de_Saussure(dec 15, 2019)

[4]Avram Noam Chomsky is an american linguist, philosopher, cognitive scientist, historian, social critic, and political activist. Sometimes called "the father of modern linguistics", he’s also a major figure in analytic philosophy and one of the founders of the field of cognitive science.https://en.wikipedia.org/wiki/Noam_Chomsky(dec 15, 2019)

in the field of linguistics such as sociolinguistics, psycholinguistics, phonetics, phonology, morphology and so on.

While linguistics was at the state of defining its core, Greenberg's discoveries had a profound impact into the establishment of linguistics as a field of science. He classified different types of languages of Africa, America, Australia and of other parts of the world.

Greenberg's research is particularly significant as he applied his developments for fourteen Languages in Oceania and shown that all of them had the same origin.

His major interest was in Linguistic Universals. Greenberg initiated language universals by analysing morphemes and words, language classification and subgrouping, evolution, diffusion, migration and the relationship between structure and function. He applied his method for word orders and morphological categories. By that, he based the methodology of what was known as the typological approach to grammar to derive major empirical results and offer an explanation. This is used widely today in typological analyses.

Greenberg studied diachronic typology, the fact that the universals of languages are able to change. He was interested in universals of synchronic language structure. The result of his research in cooperation with other linguistic scientists was a series of articles about the universals of human language. He found that limitations in cross-linguistic variations impacts the flow of language change and re-analysed synchronic typology as diachronic typology.

The other significant topics that Greenberg described on major papers concerned numeral constructions [Greenberg, 1977], gender markers [Greenberg, 1978], word order [Greenberg, 1980] and pronouns [Greenberg, 1988], [Greenberg, 1993], [Greenberg, 2000]. After his retirement, he paid lots of attention to "Genetic classification". Until the last days of his life, Greenberg worked on the Indo-European languages and their closest relatives, with focus on the grammar and lexical evidence.

2.2.2 Fourty-five Universals

In this section we discuss the extremely influential paper of [Greenberg, 1963]. He sought to discover the universal structures on which human language is based on with a functionalist point of view. In order to describe his universals, Greenberg decided to work on morpheme and word order. It was due to his previous experiences that he decided to place a focus on this particular aspect of grammar. The universals were classified into 3 large parts:

– Typology: which groups the first 7 universals

– Syntax: which groups 18 universals
– Morphology: which groups the last 20 universals

Greenberg's universals follow an order which links the previous with the following, it means that they are related to each other and are not randomly assigned.

All of the 45 universals of Greenberg are implicational as explained by [Köhler et al., 2005] and Bisang[5]. This means that universals show a dependency between two logically independent statements A and B.

An implicational universal is one which say that:

(a) "If a language has a characteristic A, then it also has a characteristic B."

There's also another but less important formulation that expresses what an implicational universal is, which is : "If a language does not have characteristic B. However, if a language does not have characteristic A, then it can either have or not have, characteristics B."

To explain and demonstrate his claims, Greenberg used a sample of 30 languages that he personally knew or those with a reasonably available grammar at the time.

To detail the rules, he employed three sets of criteria involving basic factors in typology. These criteria consist of:

1 Preposition and Postposition (Pr/Po): adpositions are elements of grammar that attach themselves to a word to show a word's relationship to another nearby word, in English: "The book by Greenberg" "by" is a preposition in this as it occurs before "Greenberg" and links the name to the book to express a relationship.
2 Order of subject-object-verb (SVO,SVO,SOV,VOS,OVS,OSV): the order is identified by analysing position of subject and object in relation to the verb in a sentence, for instance English is an SVO language as the subject occurs before the verb which occurs before the object for example the sentence "We read Greenberg's paper"
3 Qualifiyng adjectives related to the noun (NA/AN): it means if an adjective precedes the noun or follows it For example in English "The yellow book" the adjective precedes the noun but in French "Le livre jaune" it follows

[5] https://www.phil-fak.uni-duesseldorf.de/summerschool2002/Bisang3.PDF(dec 12, 2019)

2.2.3 Presentation of Greenberg Universals

As a first step in our work, we studied and rewritten all the universals in our words as we understood them. After that, they became easier to comprehend at our level. But it was just a simple interpretation of them and not necessarily what Greenberg thought. We have tested the universals with examples in five different languages that are known to the members of our team: English, French, Polish, German and Persian.

An example of that can be shown with universal 14 in 2.1 where we illustrated the universal giving a sentence that conforms to it in the 5 languages (respectively). This one states that: "In conditional statements, the conditional clause precedes the conclusion as the normal order in all languages" [Greenberg, 1963].

(1) If grandma is there, then I will go.
/ɪf ˈgrænmɑː ɪz ðeə, ðɛn aɪ wɪl gəʊ/

(2) Si ma grand-mère est là, alors j'irai.
(If my grand-ma is there, then I will go)
/si ma gʀɑ̃ mɛʀ ɛ la alɔr ʒiʀe/

(3) Jeśli babcia tam jest to tam pojadę.
(If grand-ma there is then I will go)
/jɛɕli babt͡ɕa tam jɛst tɔ tam pɔjadɛ̃/

(4) Wenn meine Großmutter dort ist, werde ich ankommen.
(If my grand-ma there is, will I go)
/vɛn maɪnə groːsmʊtɐ dɔʁt ɪst veːɐdə ɪç ank ɔmən/

(5) اگر مادربزرگ آنجا باشد؛ من خواهم رفت
(go will I be there grand-ma If)
/ægær mɒːdær boɜorg ɒːndʒɒː bɒːʃæd; mæn xɒːhɒːm rɒːft/

Figure 2.1: Universal 14: examples in five languages

But all of them are not like this one, most of the time we found difficulties to understand and describe them. As an example we have universal 22. This universal states that : "If in comparisons of superiority, the only order, or one of the alternative orders, is standard marker adjective, then the language is postpositional. With overwhelmingly more than chance frequency, if the only order is adjective marker standard, the language is prepositional" [Greenberg, 1963]. It is a fairly long definition that contains ambiguous terms. We not have clear explanation as

concerned the features, "adjective", "marker" and "standard", hence it is not simple to understand the universal correctly.

All of the universals can be seen in appendix A.

2.3 Linguistic Universals and NLP

In the area of NLP, as demonstrated by [Raskin, 1987], it is widely known that linguistic knowledge is not totally required for all the tasks, but is recommended in order to understand the problem better and hence produce a project with better qualities and results. We will not consider all linguistic features but will concentrate on those that we discussed that being adpositions, word order and adjective-noun pairing.

Some difficulties with ambiguity may be encountered depending on the task, which is noticeable with general linguistic features, therefore a basic knowledge of linguistics is expected [Bender, 2009].

Also, we have to consider that there are currently over 6,000 languages in existence, and more when considering dead languages, sub-languages and difficulties in determining a language from a dialect, due to our current possibilities and resources we can only consider a fraction of them. Indeed, testing and analysis is not just a matter of ”word” extraction; the linguistic knowledge about the grammatical structures of one or more natural languages must be taken into account.

The variations in language structures can have consequences on the NLP applications that developers want to create. It is for these reasons that qualitative and quantitative analysis studies of linguistic features are required. And we have to learn more about universals, taking into account a large sample of languages even those without a lot of speakers.

3. Constraints for resources and tool

3.1 Sets of corpora

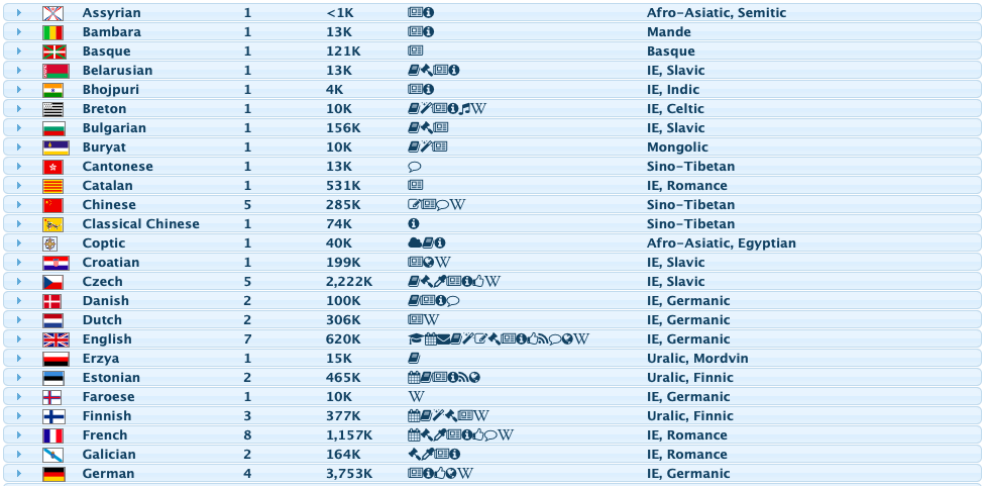
3.1.1 Universal Dependencies

UD project[1] is composed of several corpora annotated in accordance with common guidelines and infrastructure. The project arose to remedy the issue of divergent annotation in the same language and beyond that resulted from differing annotation standards [Nivre et al., 2016]. As such UD project offers a consistent annotation standard across languages that is necessary for the testing of GU. The current version that we will use is UD 2.5 released on November 15, 2019. It groups 157 treebanks in 90 languages.

Concerning the available corpora on UD, we can see on the 3.1 that we have lot of data to work with. It is structured like this: the 1st column enumerates the languages, the 2nd the number of corpora available in those languages, the 3rd the number of words and the 4th the features. We remark that the various corpora are not homogeneous. UD groups a large number of corpora and they are annotated by multiple annotators. However, the size and the contents for each language are different. For instance, some languages like Assyrian has only 1 corpora with less than 1K of words which is a smallest one, while there exist very complete databases like German that has 4 corpora with more than 3,500K words.

The project provide a universal inventory of categories to ease consistent annotation of similar grammatical forms across languages such as 17 part of speech

[1] https://universaldependencies.org



Assyrian	1	<1K	Afro-Asiatic, Semitic
Bambara	1	13K	Mande
Basque	1	121K	Basque
Belarusian	1	13K	IE, Slavic
Bhojpuri	1	4K	IE, Indic
Breton	1	10K	IE, Celtic
Bulgarian	1	156K	IE, Slavic
Buryat	1	10K	Mongolic
Cantonese	1	13K	Sino-Tibetan
Catalan	1	531K	IE, Romance
Chinese	5	285K	Sino-Tibetan
Classical Chinese	1	74K	Sino-Tibetan
Coptic	1	40K	Afro-Asiatic, Egyptian
Croatian	1	199K	IE, Slavic
Czech	5	2,222K	IE, Slavic
Danish	2	100K	IE, Germanic
Dutch	2	306K	IE, Germanic
English	7	620K	IE, Germanic
Erzya	1	15K	Uralic, Mordvin
Estonian	2	465K	Uralic, Finnic
Faroese	1	10K	IE, Germanic
Finnish	3	377K	Uralic, Finnic
French	8	1,157K	IE, Romance
Galician	2	164K	IE, Romance
German	4	3,753K	IE, Germanic

Figure 3.1: Corpora in UD: short set

tags[2], 47 morphological features[3] and 37 syntactic dependencies[4]. Further allowing language specific constructions if needs be; the project is publicly accessible with many existing treebanks being converted to UD. The reason we utilise UD is that it is a vast collection of treebanks for various languages not dependent on a specific language family and because of its unified structure we can be sure that we will be able to obtain results that will be consistent. Hence, allow us to extrapolate and draw reliable conclusions from.

Language annotations

A language tag is an indicator of a grammatical feature present within a sentence or a word. A tagset hence is a collection of these tags. UD maps a diverse tagset to a common standard. Each word depends either on another word or on a root of the sentence. The categorisation is made following three principles: content words are related by dependency relations, function words attach to the content word that they specify further, and punctuation is attached to the head of the phrase or clause where it is present [Nivre et al., 2016]. This diverse tagset provides

[2] https://universaldependencies.org/u/pos/index.html
[3] https://universaldependencies.org/u/feat/index.html
[4] https://universaldependencies.org/u/dep/

us an opportunity to indicate how linguistic universals given by Greenberg can be represented, on a set of naturally occurring text corpora with punctuation and dependency standards.

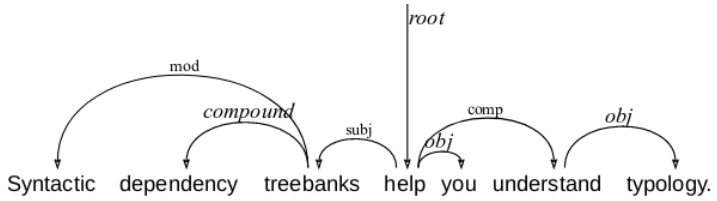


Figure 3.2: Extract from [Gerdes et al., 2019] of an ordered dependency tree

The languages within treebanks can be arranged on a spectrum with absolute head-initial and head-final patterns at both ends. [Gerdes et al., 2019] states that the treebank based methods will be able to provide a complete and fine-grained typological analysis without having to rely on focus on basic word order phenomena. Data will be obtained using extracted patterns that will be generated by us and processed using the GREW tool which we are going to describe in the next section 3.2.

Language analysis limitations

Our aim is to satisfy individual language analysis while also being appropriate for linguistic typology (cross-linguistic parallelism). We will test GU based on data analysis of a set of uniformly annotated texts present within a diverse set of languages. As we mentioned above, some languages in UD are not well balanced or they have a small sample size. Due to new UD methodology, we are able to tackle the questions of universals within languages and discuss the potential of UD for future typological studies.

Due to the vast differences between various corpora in UD we will have to be selective on which ones do we include for pattern processing. We cannot confirm whether the given corpora will produce results that are a true representation of a certain language. For instance, we have no access to the words and sentences of some corpora (according to copyright laws) or, some of them are exclusively designed for a certain manner (for example corpora of questions). These types do not represent the broad use of the language, therefore we will separate them from the rest. We will also decide on the minimum size requirements when selecting our corpora.

3.1.2 Surface-Syntactic Universal Dependencies

When testing universals relating to syntax, we are interested in using the Surface-Syntactic Universal Dependencies (SUD). This is because SUD gives a comprehensive view of all constructions of one language in contrast to UD that relies on maximising parallelism between languages by reducing language differences and hence is less suitable for typological research on syntax [Gerdes et al., 2018]. This means that for the universals concerned with syntax we will opt for the SUD annotation scheme. A concern about using UD for testing of GU was raised by [Gerdes et al., 2019]. The authors of that paper highlight a problem with accuracy when using the standard UD notation for testing of universals 19 and 25, namely with Japanese where UD has a number of head initial relations whereas in SUD it is nearly completely head final.

Furthermore SUD aims to define the dependency labels and links on purely syntactic criteria which allows for bi-directional transformation from UD to SUD [Chen and Gerdes, 2017] mention that this transformation ensures inter and intra language coherence of UD treebanks.

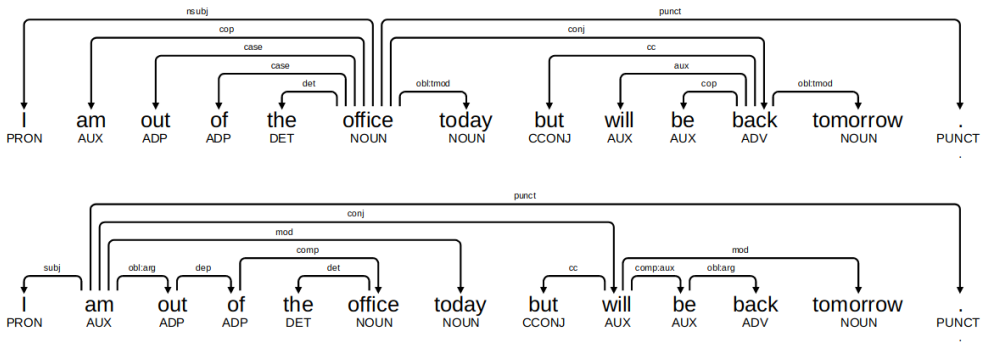


Figure 3.3: The same sentence annotated in UD (Top) and SUD (Bottom). Figure 1 extract from [Gerdes et al., 2018]

It is important to mention that UD categorises the head by parting with the surface syntax criteria and then applies the criterion of "content word as head" this is contrasted with the surface syntax which utilises a distributional criteria of each individual word. The rule for this criterion is that the surface syntactic head determines the distribution of the unit [Gerdes et al., 2018], this seen in figure 3.3.

This standard follows standard distributional criteria for head positioning and relation labeling and therefore places itself as closer to more traditional constituency-based surface syntax as well as to dependency-based surface syntax. This signifies

that the SUD can be used by annotators who are trained in more traditional forms of syntax. We have to mention though that SUD is not yet operable on all corpora and as such we would have to make a judgement as to how we will utilise it.

3.2 A graph rewriting tool: GREW

GREW[5] is dedicated for use in NLP. All structures are considered as graphs. Graphs are a natural way of representing the deep syntax and the semantics of natural languages.

GREW creates a graph by operating on given keywords of the query language which is called a pattern. The nodes in a pattern can be filtered by selected values without any constraint and be bound to a defined name. It is searched among the nodes and edges of a graph. GREW allows for corpora annotations and transformations to be presented within a common framework. Currently however there is no standard model for graph rewriting, so the authors of this tool created one specifically catered to NLP. We aim at implementing it into our project with the use of a ”grew-daemon” tool which allows us to, via the use of a specified pattern and corpora, generate and output the results of patterns processed on specified corpora. Graph Rewriting itself allows for a combination of efficiency along with linguistic readability for producing representations at the desired linguistic level [Bonfante et al., 2018].

GREW itself is written in the Ocaml programming language[6] which allows for Python binding. We will use Python for programming as it is most commonly used for NLP and we have experience in dealing with the NLTK toolkit. This knowledge will allow us to operate and understand the GREW tool better.

The GREW library contains a syntax for describing patterns and the corresponding matching function. It works by splitting the pattern matching code from the patterns themselves. This means that programmers that use GREW are able to define their own patterns that can then be modified without changing the code whatsoever. This represents a major advantage in terms of design as well as long-term maintenance [Bonfante et al., 2018]. This is obviously a big advantage for us as it provides long term support for the given patterns and allows for easier testing.

Some previous work on GREW involved developing two treebank conversion grammars, one SUD->UD and one UD->SUD conversion tool which is shown in [Gerdes et al., 2018] and 3.3. The authors created a set of rules and a strategy

[5] http://grew.fr
[6] http://ocaml.org

that described how the rule applications must be ordered. We may be pressed to adopt the same set of rules and strategy when working with SUD.

The reason we use GREW is that we want to produce patterns that relate to the GU. It allows us to traverse across multiple corpora of various languages, to find and statistically correlate various universals. The aim is to making a case for denying or confirming them (an example of a pattern and how we processed it using GREW is present in chapter 5 and in the following GREW-Match subsection).

3.2.1 GREW-Match

Furthermore we will use an online[7] graph interface for GREW to evaluate any erroneous results or check our findings. To demonstrate how GREW-Match works we wrote a pattern (3.4) that processes the SVO word order.

```
SVO pattern
{V[upos= "VERB"]; V-[nsubj|csubj]->N1; V-[obj]->N2;  N1<<V; V<<N2}
```

Figure 3.4: The GREW pattern for identifying SVO relation in tested corpora

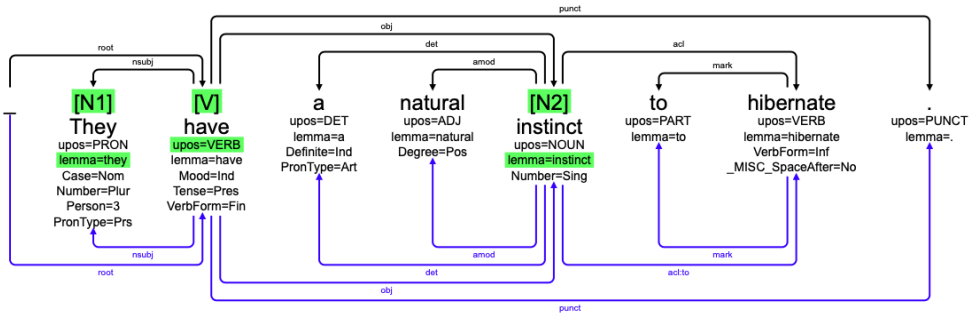


Figure 3.5: Simple output of GREW-match

GREW-Match produces a graph (3.5). For each word it lists the grammatical properties and then indicates the relationship of each word to another via use of arrows that carry the relation type thence for this reason GREW-Match is a powerful tool for anomaly detection.

[7] http://match.grew.fr

4. Experiments in NLP about linguistic typology

4.1 Typological information

Studying linguistic universals leads to taking into account typological linguistics, which is the study and classifications of languages according to their structural and functional features.

There exist sets of databases of typological information. Among them the World Atlas of Languages Structures (WALS)[1] [Dryer and Haspelmath, 2013] which provides information on the location, linguistic affiliation and basic typological features of a great number of the world's languages. But with limitations (for example the data are not annotated on WALS).

In the domain of linguistic typology, we can find several influential points for the creation of NLP models, like: multilingual syntactic parsing, POS tagging, phonological modeling, language learning etc.

It is shown by [Ponti et al., 2018] and [O'Horan et al., 2016], that the existence of linguistic typology must be taken into account and applied to NLP systems.

Furthermore as it relates to other domains, NLP could have an important place in increasing the resources or documentation on typology.

4.2 Existing work on Greenberg's Universals

We have discovered that different ways of studying linguistic typology are used. Some authors have readily available work on the question of universals and utilising different resources: networks, quantitative/statistical analysis and different databases: WALS, UD, for instance.

[1] https://wals.info

4.2.1 Questionable procedures

Since networks are known to explain and understand complex systems, the authors of [Sharma et al., 2019] have tried to analyse GU with the concept of language networks which allows for better representation of linguistic knowledge.

They encoded relationships between items to construct the language network 4.1 so that this network could represent generalizations and since GU are sets of generalisations, they can be derived from language networks.

A network point of view can be useful to observe the relations from another perspective.

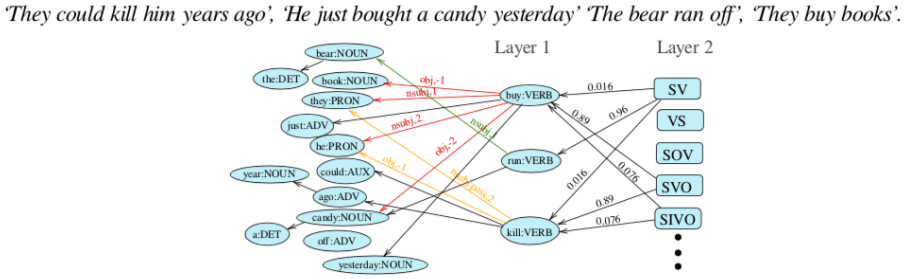


Figure 4.1: The language network for 4 sentences. Figure 2 extract from [Sharma et al., 2019] with courtesy of Kaivalya Swami

We can note that this work is very difficult to interpret and digest using our present knowledge. Hence we want not utilise this approach in our project. We prefer to utilise resources and strategies of UD and GREW. This is because there is more documentation and corpora data letting us test our patterns on multiple languages. We cannot utilise the same approach with a network due to resource limitations as we would have to train the net on each language. The results may prove unusable due to noise and this may limit the exploitation of our proposed state of the art by other researchers.

4.2.2 As a reference work for us

The paper [Gerdes et al., 2019] on using a data driven topological analysis intends to provide a more empirical and more accurate way in which GU can be refined. It discusses the use of UD and GREW in order to test the universals 19 and 25.

The authors of the paper attempted to statistically confirm that their method managed to confirm the aforementioned universals. Yet they mention that these

universals, as well as others in the original Greenberg text, are rather vague in themselves and purely implicational. This makes sense due to the conditions of computing power in the 1960's the universals were produced by Greenberg without considering computational processing. The authors urge for these universals to be rewritten into a modern empirically verifiable format and offer a method for accomplishing such task, with certain universals becoming quantified. For instance typological universals contain a aforementioned logical clause of "If a language has a characteristic A, then it also has a characteristic B". This can be used to describe universal restrictions on human languages that then shape the clouds of languages on scatter plots of various properties as seen on 4.2 which analyses universal 25 which states: "Almost every language has a higher proportion of nominal objects than of pronominal objects on the right of the verb"A

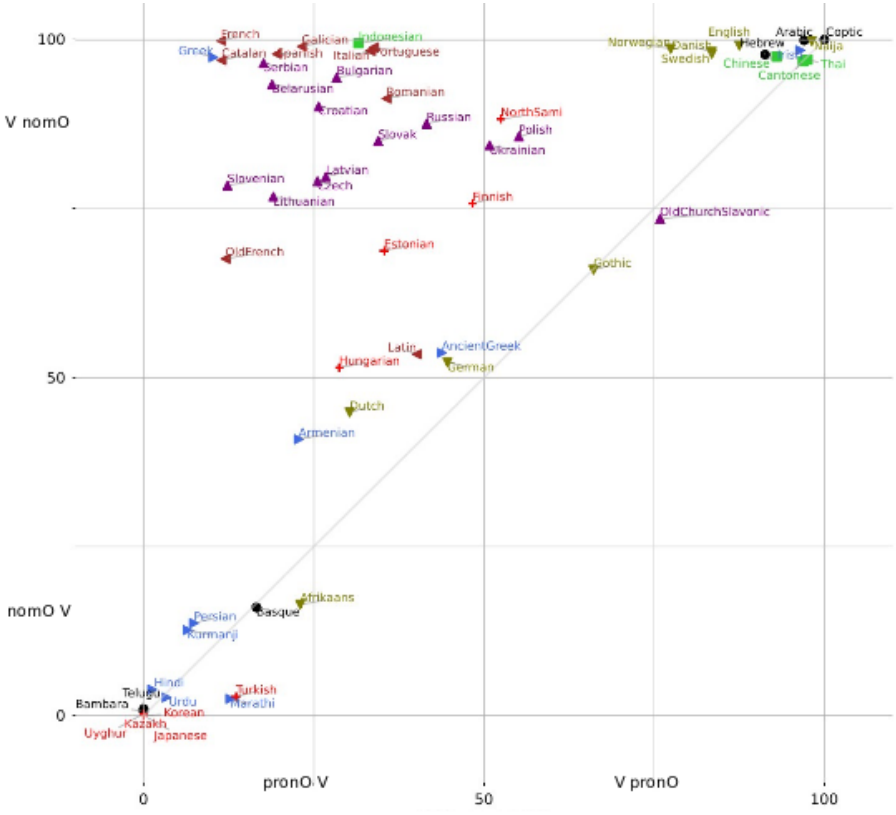


Figure 4.2: Scatter plot of the percentage of verb pronominal compared to verb nominal. Extract from [Gerdes et al., 2019]

The authors mention that the universal 25 was written as a qualitative measure. The authors due to the vague phrasing of the universal were not certain whether Greenberg's intention was that this universal is to be very strictly enforced (i.e universal 25 is to always apply) or whether it is to apply most of the time. As a result the authors decided to set a threshold for quantitative testing.

Despite the results generally confirming to a given threshold, it is not certain what percentage would reflect initial Greenberg's intention, resulting in a high percentage and hence high accuracy but high exclusion or the opposite. Authors therefore decided to redefine the universal once again to exclude considering a specific threshold: "Almost every language has a higher proportion of nominal objects than of pronominal objects on the right of the verb" is the redefinition, which conforms to the provided data. Authors reveal that indeed working with quantitative data allows for completely new universals and their considerations to apply.

It is our belief that using the same reasoning we must in some circumstances redefine universals for quantitative testing using UD and the GREW tool, we shall use the definitions used by authors of [Gerdes et al., 2019] for the universals 19 and 25.

Furthermore we may use their method to map head-initial head-daughter dependencies.

5. First moves and Roadmap

We first have to review the GU, see which ones are possible to test and which ones are not. In fact, some of them are not testable because we do not possess all the resources for a portion of linguistics features. So we need to establish an order from the easiest to hardest for us to test. We already know that we can test the universal 25 "If the pronominal object follows the verb, so does the nominal object" due to the example that we have with the [Gerdes et al., 2019] paper. In addition, we cannot test the universal 9 "With well more than chance frequency, when question particles or affixes are specified in position by reference to the sentence as a whole, if initial such elements are found in prepositional languages, and if final, in postpositional" because it requires morphological information that we are not in a position to have with the available resources.

We plan firstly, on exploiting GREW by producing our own patterns. Then using the command prompt, executing the patterns on a compiled set of corpora. The specified corpora will be given in a json file. We shall of course exclude all corpora that do not fit our requirements as discussed in 2.3.

```
SOV pattern
{V[upos=VERB];N1[];N2[];V-[nsubj]->N1;V-[obj]->N2; N1<<N2; N1<<V;N2<<V}
```

Figure 5.1: The Grew pattern for identifying SOV relation in tested corpora

We already presented an example of what a pattern looks like (3.4), we now present a modified version of it (5.1) for the SOV word order.

For the pattern structure 5.1:

- The first part defines the variables as a verb related to a first subject word and to a second object word;
- The latter part defines the order of these variables, how they are to be positioned in a sentence.

For inducing the other patterns, modification of the latter part which defines the order of the variables is needed.

Results 5.2 show that the five languages tested seem to have a basic SVO word order, French however seems to also have a significant percentage of SOV as well, Persian has a strong preference for SVO but also correlates highly with VSO, German has a significant amount of SOV, VSO and VSO sentences, Polish on the other hand has multiple examples in all word orders.

Corpus	# sentences	SVOproc	SOVproc	VSOproc	VOSproc	OVSproc	OSVproc	
UD_English-EWT-----		16622	6460	0	1	1	1	396
UD_French-Sequoia		3099	902	70	1	1	23	64
UD_Polish-LFG-----		17246	1398	224	66	115	229	87
UD_German-GSD-----		15590	2110	2173	1333	230	192	416
UD_Persian-Seraji		5997	18	1626	0	1	9	76

Figure 5.2: The processed result for the word order using grew on 5 different language corpora

English seems to show erroneous values for OSV with almost 400 instances of this order being present, analysing this phenomena on the online match.grew interface we deducted that often question sentences were missing appropriate punctuation and such were marked as standard sentences (if there was no '?' at the end).

We will use the previous patterns to categorise languages and then select those that fulfill the criteria mentioned by Greenberg (adpositions ; word order; adjective-noun position).

For our work a significant part of it will involve defining new patterns and testing them on the corpora. Then experimenting on them and if needs be, redefine them (to interpret them for sufficient quantitative analysis)

For our work a significant part of it will involve defining new patterns, testing them, experimenting on them and for those Greenberg's universals that are not interpreted for quantitative analysis, redefine them.

At a later time the data will be evaluated using a python program that outputs the given results in a csv file. For each universal, we will check the output using appropriate statistical measures and when we are satisfied we will record them and give a reason as to the resulting data patterns, mentioning any irregularity and if possible explaining it.

We plan to use some graphical processing to possibly produce a graph for universals that can be represented this way, this will make our results look more presentable and easier to digest by the reader.

6. Conclusion

Our aim for this project is to produce a set of patterns whereby we are able to investigate statistically various GU, and to make a case for confirming them or (if results significantly suggest) deny them. For universals that are hard to define in a way that allows for quantitative measures we will redefine them. Overall for this project we will utilise graph rewriting software GREW and the available corpora from the UD project. We shall finally produce a presentation and a lab report by the end of the project detailing in depth the process and whether we managed to confirm or deny selected universals using our patterns.

We remark on the idea of linguistic universals as being highly interesting and relevant for modern research. Current technology and computational power allows us to utilise a large set of uniformly annotated data for multiple languages. This means that we may derive new typological universals and finally have the capabilities to test those that were hypothesised in the past. It is worthy to mention that this project has a long term significance because new corpora are consistently added on the UD project with biyearly updates, for instance a significant update mentioned by one of the papers [Gerdes et al., 2018] suggests that SUD scheme could be adapted for each language leading to more homogenous annotation in UD 3.0. This means that our findings are not final, but given our framework and state of the art we expect that other researchers will be able to utilise our work and test the patterns on future UD versions and new language corpora.

Bibliography

[Bender, 2009] Bender, E. M. (2009). Linguistically naïve!= language independent: why nlp needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32.

[Bonfante et al., 2018] Bonfante, G., Guillaume, B., and Perrier, G. (2018). *Application de la réécriture de graphes au traitement automatique des langues*, volume 1 of *Série Logique, linguistique et informatique*. ISTE editions.

[Chen and Gerdes, 2017] Chen, X. and Gerdes, K. (2017). Classifying languages by dependency structure. typologies of delexicalized universal dependency treebanks. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 54–63, Pisa,Italy. Linköping University Electronic Press.

[Cooreman and Goyvaerts, 1980] Cooreman, A. and Goyvaerts, D. (1980). Universals in human language. a historical perspective. *Revue belge de Philologie et d'Histoire*, 58(3):615–638.

[Croft, 2001] Croft, W. (2001). Obituary: Joseph harold greenberg. *Language*, 77(4):815–830.

[Croft et al., 2017] Croft, W., Nordquist, D., Looney, K., and Regan, M. (2017). Linguistic typology meets universal dependencies. In *TLT*, pages 63–75.

[Dryer and Haspelmath, 2013] Dryer, M. S. and Haspelmath, M., editors (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

[Gerdes et al., 2018] Gerdes, K., Guillaume, B., Kahane, S., and Perrier, G. (2018). SUD or surface-syntactic universal dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Workshop on Universal*

Dependencies (UDW 2018), pages 66–74, Brussels, Belgium. Association for Computational Linguistics.

[Gerdes et al., 2019] Gerdes, K., Kahane, S., and Chen, X. (2019). Rediscovering greenberg's word order universal in ud. In *Proceedings of the Universal Dependencies Workshop (UDW)*, SyntaxFest, Paris.

[Greenberg, 1993] Greenberg, J. (1993). The second person is rightly so called. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, pages 9–9.

[Greenberg, 1963] Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. *Universals of Language*, pages 73–113.

[Greenberg, 1977] Greenberg, J. H. (1977). Numeral classifiers and substantival number: Problems in the genesis of a linguistic type. In Makkai, A., Makkai, V. B., and Heilmann, L., editors, *Linguistics at the crossroads*, number 4 in Universita degli studi di Bologna centro interfacolta di linguistica teorica e applicata, page 276?300. Liviana Editrice, Bologna.

[Greenberg, 1978] Greenberg, J. H. (1978). How does a language acquire gender markers. *Universals of human language*, 3:47–82.

[Greenberg, 1980] Greenberg, J. H. (1980). Circumfixes and typological change. In *Papers from the Fourth International Conference on Historical Linguistics, Stanford, March 26–30 1979*, page 233. John Benjamins Publishing Company.

[GREENBERG, 1981] GREENBERG, J. H. (1981). Nilo-saharan moveable-k as a stage III article (with a penutian typological parallel). *Journal of African Languages and Linguistics*, 3(2).

[Greenberg, 1988] Greenberg, J. H. (1988). The first person inclusive dual as an ambiguous category. *Studies in Language*, 12(1):1–18.

[Greenberg, 2000] Greenberg, J. H. (2000). 24. from first to second person: The history of amerind. In *Functional Approaches to Language, Culture and Cognition*, page 413. John Benjamins Publishing Company.

[Köhler et al., 2005] Köhler, R., Altmann, G., and Piotrowski, R. G. (2005). *Quantitative Linguistik / Quantitative Linguistics*. Walter de Gruyter. pages 554-578.

[Nivre et al., 2016] Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

[O'Horan et al., 2016] O'Horan, H., Berzak, Y., Vulić, I., Reichart, R., and Korhonen, A. (2016). Survey on the use of typological information in natural language processing. *arXiv preprint arXiv:1610.03349*.

[Ponti et al., 2018] Ponti, E. M., O'Horan, H., Berzak, Y., Vulić, I., Reichart, R., Poibeau, T., Shutova, E., and Korhonen, A. (2018). Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, (Just Accepted):1–43.

[Raskin, 1987] Raskin, V. (1987). Linguistics and natural language processing. *Machine translation: Theoretical and methodological issues*, pages 42–58.

[Sharma et al., 2019] Sharma, K., Swami, K., Shete, A., and Husain, S. (2019). Can greenbergian universals be induced from language networks? In *Proceedings of the Universal Dependencies Workshop (UDW)*, SyntaxFest, Paris.

A. Universals

The 45 universals - extracts from [Greenberg, 1963] - are reprinted below.

1. In declarative sentence with nominal subject and object, the dominant order is almost always one in which the subject precedes the object
2. In language with prepositions, the genitive almost always follows the governing noun while in languages with postpositions it almost always precedes
3. Language with dominant VSO order are always prepositional
4. With overwhelmingly greater than chance frequency, languages with normal SOV order are postpositional
5. If a language has dominant SOV order and the genitive follows the governing noun, then the adjective likewise follows the noun
6. All language with dominant VSO order have SVO as an alternative basic or as only alternative basic order
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
8. When a yes-no question is differentiated from the corresponding assertion by an intonational pattern, the distinctive intonational features of each of these patterns are reckoned from the end of the sentence rather than from the beginning
9. With well more than chance frequency, when question particles or affixes are specified in position by reference to the sentence as a whole, if initial such elements are found in prepositional languages, and if final, in postpositional
10. Questions particles or affixes, when specifies in position by reference to a particular word in the sentence, almost always follow that word. Such particles do not occur in languages with dominant order VSO
11. Inversion of statement order so that verb precedes subject only occurs in languages where the question word of phrase is normally initial. This same

inversion occurs in interrogative word questions

12. If a language has dominant order VSO in declarative sentence, it always puts interrogative words or phrases first in interrogative word questions; if it has dominant order SOV in declarative sentences, there is never such an invariant rule
13. If the nominal object always precedes the verb, then verb forms subordinate to the main verb also precede it
14. In conditional statements, the conditional clause precedes the conclusion as the normal order in all languages
15. In expressions of volition and purpose, a subordinate verbal form always follows the main verb as the normal order except in those languages in which the nominal object always precedes the verb
16. In languages with dominant order VSO, an inflected auxiliary always precede the main verb. In languages with dominant order SOV, an inflected auxiliary always follows the main verb
17. With overwhelmingly more than chance frequency, language with dominant order VSO have the adjective after the noun
18. When the description adjective precedes the noun, the demonstrative, and the numeral with overwhelmingly more than chance frequency, does likewise
19. When the general rule is that the descriptive adjective follows, there may be minority of adjective which usually precede, but when the general rule is that descriptive adjectives precede, there are no exceptions
20. When any or all of the items (demonstrative, numeral, and descriptive adjective) precede the noun, they are always found in that order. If they follow, the order is either the same or its exact opposite
21. If some or all adverbs follow the adjective they modify, then the language is one which the qualifying adjective follows the noun and verb precedes its nominal object as the dominant order
22. If in comparisons of superiority, the only order, or one of the alternative orders, is standard marker adjective, then the language is postpositional. With overwhelmingly more than chance frequency, if the only order is adjective marker standard, the language is prepositional
23. If in apposition the proper noun usually precedes the common noun, then the language is one in which the governing noun precedes its dependent genitive. With much better than chance frequency, if the common noun usually precedes the proper noun, the dependent genitive precedes its governing noun
24. If the relative expression precedes the noun either as the only construction or

as an alternative construction, either the language is postpositional, or the adjective precedes the noun or both

25. If the pronominal object follows the verb, so does the nominal object
26. If a language has discontinuous affixes, it always has either prefixing or suffixing or both
27. If a language is exclusively suffixing, it is postpositional: if it is exclusively prefixing, it is prepositional
28. If both the derivation and inflection follow the root, or they both precede the root, the derivation is always between the root and the inflection
29. If a language has inflection, it always has derivation
30. If the verb has categories of person-number or if it has categories of gender, it always has tense mode categories
31. If either the subject or object noun agrees with the verb in gender, then the adjective always agrees with the noun in gender
32. Whenever the verb agrees with a nominal subject or nominal object in gender, it also agrees in number
33. When number agreement between the noun and verb is suspended and the rule is based on order, the case is always one in which the verb precedes and the verb is in the singular
34. No language has trial number unless it has a dual. No language has a dual unless it has a plural
35. There is no language in which the plural does not have some non-zero allomorphs, whereas there are languages in which the singular is expressed only by zero. The dual and the trial are almost never expressed only by zero
36. If a language has the category of gender, it always has the category of number
37. A language never has more gender categories in non-singular numbers than in the singular
38. Where there is a case system, the only case which ever has only zero allomorphs is the one which includes among its meaning that of the subject of the intransitive verb
39. Where morphemes of both number and case are present and both follow or both precede the noun base, the expression of number almost always comes between the noun base and the expression of case
40. When the adjective follows the noun, the adjective expresses all the inflectional categories of the noun. In such cases the noun may lack overt expression of one or all of these categories
41. If in language the verb follows both the nominal subject and nominal object as the dominant order, the language almost always has a case system

42. All language have pronominal categories involving at least three persons and two numbers
43. If a language has gender categories in the noun, it has gender categories in the pronoun
44. If a language has gender distinctions in the first person it always has gender distinctions on the second or third person, or in both
45. If there are any gender distinctions in the plural of the pronoun, there are some gender distinctions in the singular also