Mathematics

http://www.aimspress.com/journal/Math

AIMS Mathematics, 9(1): 2181–2214.
DOI: 10.3934/math.2024109
Received: 13 August 2023
Revised: 01 November 2023
Accepted: 19 November 2023
Published: 21 December 2023

Research article

Exploring the complexity of natural languages: A fuzzy evaluative perspective on Greenberg universals

Antoni Brosa-Rodríguez[*], **M. Dolores Jiménez-López**[*] **and Adrià Torrens-Urrutia**[*]

Universitat Rovira i Virgili, Research Group on Mathematical Linguistics (GRLMC), 43002, Tarragona, Spain

* **Correspondence:** Email: antoni.brosa@urv.cat, mariadolores.jimenez@urv.cat, adria.torrens@urv.cat.

Abstract: In this paper, we introduced a fuzzy model for calculating complexity based on universality, aiming to measure the complexity of natural languages in terms of the degree of universality exhibited in their rules. We validated the model by conducting experiments on a corpus of 143 languages obtained from Universal Dependencies 2.11. To formalize the linguistic universals proposed by Greenberg, we employed the Grew tool to convert them into a formal rule representation. This formalization enables the verification of universals within the corpus. By analyzing the corpus, we extracted the occurrences of each universal in different languages. The obtained results were used to define a fuzzy model that quantifies the degree of universality and complexity of both the Greenberg universals and the languages themselves, employing the mathematical theory of evaluative expressions from fuzzy natural logic (FNL). Our analysis revealed an inversely proportional relationship between the degree of universality and the level of complexity observed in the languages. The implications of our findings extended to various applications in the theoretical analysis and computational treatment of languages. In addition, the proposed model offered insights into the nature of language complexity, providing a valuable framework for further research and exploration.

Keywords: linguistic universals; linguistic complexity; evaluative expressions; fuzzy grammar; linguistic gradience; linguistic constraints
Mathematics Subject Classification: 03B65

1. Introduction

This paper is placed in the area of mathematical linguistics and presents a model for calculating the complexity of natural languages.

Calculating the complexity of languages is very important for understanding their structures, use and evolution. The measurement of language complexity offers valuable insights into theoretical linguistics, language learning and teaching and the development of language technologies. Moreover, by calculating linguistic complexity, we gain a deeper and more comprehensive understanding of the richness and diversity of languages worldwide [1–7].

Despite the growing interest in linguistic complexity and the numerous research works dedicated to the study of complexity, we still lack a clear answer regarding the inherent differences in complexity among languages [8, 9]. Understanding the complexity of languages remains a challenging task due to the diverse types of complexity, the lack of standardized measures and the varying definitions used in the field [10–12]. As a result, to establish a universally applicable method for calculating linguistic complexity, understanding the different complexity among language remains a challenge.

This paper aims to provide a method for calculating the complexity of natural languages by establishing a relationship between language complexity [10] and universality [13]. Our objective is to determine the complexity of a natural language by considering its level of universality. The fundamental idea underlying our proposal is that languages sharing more universal features will exhibit a lower level of difficulty in learning one from another and vice versa. This approach eliminates the need to compare languages pairwise and considering all their rules in order to calculate their complexity. Instead, we determine the degree of universality exhibited by a language and utilize this information to calculate its complexity. This perspective offers an efficient and objective means of assessing linguistic complexity.

In the mathematical model proposed in this paper, both complexity and universality are considered as non-discrete concepts while in traditional linguistics, complexity and universality have been treated as discrete categories; we approach them as gradual or fuzzy categories. To accomplish this, we utilize a fuzzy model to handle evaluative expressions, viewing complexity and universality as “evaluations” of languages and their rules.

We consider it important to approach the concepts of complexity and universality from a fuzzy perspective for two reasons. On one hand, a linguistic feature cannot simply be classified as universal or non-universal; rather, it can exhibit various degrees of universality. This notion has significant implications when calculating the universality of languages, as it involves not only considering the number of universals a language possesses but also taking into account the degree or level of universality associated with those features/rules. For instance, a language with many high-level universals will have a higher level of universality than a language with numerous lower-level universals. On the other hand, a language cannot be simply categorized as complex or non-complex, but it can present different levels of complexity. These complexity levels are determined on the levels of universality found in the languages. We establish a relationship between the two concepts as follows: A high value of universality signifies that the language shares many characteristics with other languages, resulting in a relatively lower level of complexity (making it less challenging to learn). Conversely, a language with low levels of universality will possess many exclusive and non-shared features, leading to a higher level of complexity (making it more difficult to learn).

To show the effectiveness of our proposed calculation model, we present a proof of concept in which we use a mathematical theory of evaluative expressions from fuzzy natural logic (FNL) to assess the complexity of 143 languages based on their level of universality. The level of universality for each language is determined by measuring the degree of universality of eight Greenberg universals [14]. These eight out of 45 universals are filtered out since only these eight gather three essential criteria for our work: They can be formalized with the Universal Dependencies' (UD) annotation scheme, they can be formalized using the Grew tool and they are taking into account morpho-syntactic linguistic features, which is the domain that we want to explore. We analyze the presence of the eight universals in the languages within our corpus.

The results show that our mathematical model allows for the establishment of a trichotomous scale, representing the varying values of universality and complexity and revealing an inversely proportional relationship between the degree of universality and the observed level of complexity in the languages.

In summary, this paper introduces a model that explores the complexity of natural languages by linking it with the notion of universality and considers both complexity and universality as fuzzy concepts. Moreover, it presents a proof of concept that not only validates the effectiveness of our proposed model for calculating language complexity based on universality, but also shows the potential of using a mathematical model of fuzzy evaluative expressions in linguistic analysis.

The paper is organized as follows. Initially, we provide a concise presentation of the linguistic and formal background to contextualize the contributions presented in this article. In the linguistic part, we briefly introduce linguistic universals and discuss diverse approaches to tackling linguistic complexity. The subsequent section outlines the formal tools employed in this study: Universal dependencies, Grew-Match and evaluative expressions. Moving forward, we elaborate on the methodology and present the obtained results. Finally, we engage in a comprehensive discussion of the findings and the conclusions derived from this analysis.

2. Background

2.1. Linguistic background

2.1.1. Linguistic complexity

Linguistic complexity can be defined as a multifaceted concept that can be analyzed from different perspectives [15]. From a structural perspective, complexity is understood as a formal property of linguistic systems related to the number of elements. From a cognitive perspective, complexity is defined as the processing cost of linguistic structures. From a developmental point of view, complexity will be determined by the order in which linguistic structures emerge or are learned in the processes of acquisition and learning first and second languages.

The study of linguistic complexity has witnessed significant changes in recent years. From denying the possibility of calculating complexity during the 20th century, linguistics has now shifted its focus to a growing interest in understanding and measuring linguistic complexity. The resurgence in interest began around 2001 with an article by McWhorter in *Linguistic Typology*'s special issue [16]. The once prevailing dogma of equicomplexity, which claimed that all languages are equal in total complexity, has been questioned by researchers in the 21st century [17, 18]. The increasing number of works on complexity in theoretical and applied linguistics highlights the interest in finding a method to measure linguistic complexity [19–28].

Despite the growing interest in linguistic complexity studies in recent years and the general acknowledgment of varying levels of complexity among languages, accurately quantifying these differences remains a challenge. This difficulty may arise from the diverse interpretations of complexity within the field of natural language study.

Among the possible definitions of the term, one of the most recurrent dichotomies in the literature is the one that distinguishes *absolute* complexity and *relative* complexity [29]. Additional dichotomies in the literature include *global* complexity versus *local* complexity [29] or the distinction between *system* complexity and *structural* complexity [30].

The difference between absolute complexity and relative complexity is established as follows:

- *Absolute complexity* is defined as an objective property of the system and is calculated in terms of the number of parts of the system, the number of interrelationships between the parts or the length of the description of a phenomenon. This approach is common in typology studies [16,30].
- *Relative complexity* takes into account language users and is identified with the difficulty or cost of processing, learning or acquisition. It is common in sociolinguistic and psycholinguistic studies [31].

Researchers have proposed diverse measures to capture these two types of linguistic complexity, leading to a wide range of approaches. These measures encompass a vast range of formalisms, which can be categorized into two main types:

- Measures of absolute complexity, such as the count of categories or rules, description length, ambiguity, redundancy etc. [29].
- Measures of relative complexity, which grapple with the challenge of determining the type of task (learning, acquisition, processing) and the type of agent (speaker, listener, child, adult) to consider. For instance, complexity measures related to second language learning (L2) in adults [31,32] or processing complexity [33] are examples of assessments based on difficulty/cost considerations.

Moreover, researchers have explored other disciplines to find tools for calculating language complexity. Information theory, employing formalisms like Shannon entropy or Kolmogorov complexity [29, 30, 34], complex systems theory [35] or computational linguistics [36] are some instances of disciplines that have offered quantitative measures for evaluating linguistic complexity.

Most studies on linguistic complexity have primarily focused on absolute complexity [2,20,27,29], while relative complexity [3,23,28], though conceptually consistent, has not been thoroughly explored. Approaching complexity analysis from a relative perspective poses challenges in determining the specific task (learning, acquisition, processing) and the type of agent (speaker, listener, child, adult) to consider. Some authors argue that relative complexity should be examined within the context of adult (user) second language learning (L2) [31,32]. However, many studies that explore complexity in L2 processes primarily measure the complexity of the target language, neglecting the potential influence of the learner's mother tongue on relative complexity [37–42]. Observational and experimental methods used to calculate complexity in L2 processes may encounter difficulties related to the impact of extralinguistic factors, which could influence the process and affect complexity measurements. As a result, the objectivity of such analyses for cross-linguistic comparison may be compromised, as the perceived difficulty might depend on the specific speakers considered in the experiments. Moreover, the lack of standardized definitions and measures has led to inconsistent and noncomparable results in this area [40].

In this paper, we argue that a comprehensive assessment of linguistic complexity requires considering both absolute and relative complexity together. We want to emphasize the impact of the speakers' mother tongue in studies focusing on relative complexity, particularly in the context of L2. The mother tongue plays a significant role in either facilitating or complicating the learning process of the target language, thus influencing the determination of relative linguistic complexity. Additionally, we would like to show the essential methodological advantages offered by mathematical tools in objectively calculating the relative complexity of natural languages.

2.1.2. Language universals

A language universal can be defined as "a grammatical characteristic that can be reasonably hypothesized to be present in all or most human languages" [13]. The study of universals in a cross-linguistic way is within the discipline known as linguistic typology, which is "the systematic cross-linguistic comparison that aims to discover the underlying universal properties of human language" [43]. This discipline with its roots in the 19th century underwent a revolution in 1963 thanks to the paradigm shift proposed by Greenberg [14].

The study presented by Greenberg [14] takes the form of the formulation of 45 different universals in language thanks to the comparison of different grammatical features extracted from the grammars of 30 different languages varied and representative of the languages of the world.

In the years following Greenberg, the search for such universals continued. The main difference with this pioneering work was the inclusion of many more languages in the selection in order to try to formulate more reliable universals. The search for new typological conditions leading to new universals was also attempted, and a new research methodology was explored as opposed to grammars (second-hand data): Questionnaires [44]. The results were not very different from those achieved by Greenberg years earlier, and interest in the subject gradually diminished. However, in recent years, we have observed a new boom in typological studies working with universals. The main trigger of such a change in trend may be the new possibilities opened up by cross-disciplinary collaboration with natural language processing and [45, 46]. This collaboration has made it possible to have new data to work with new methodologies (linguistic corpora and a quantitative approach with real texts) and tools that allow effective processing of a large amount of data previously unattainable, also giving rise to new metrics [47] and tools offering visualizations of previously unknown data as Typometrics [48].

In the literature, universals are usually classified taking into account two criteria: *Frequency* and *extension* [49]. Considering their extension, we distinguish two types of universals:

- *Unrestricted universals* are understood as descriptive generalizations of the languages of the world. They are typical universals formulated in the field of generativism, a perspective not considered in this paper, with structures such as: "In all languages, Y" [13].
- *Implicational universals* are a parameter globally favored under certain structural conditions. In this case, we find rules in a conditional structure, as in: "In all languages, if there is X, then there is Y".

Considering their frequency, universals can be [49]:

- *Absolute universals* are formulations applicable to all the languages of the world. Absolute universals are formulated as: "All languages have Y".
- *Statistical universals* are formulations that exhibit a high frequency of adoption in the languages of the world without being absolute. Statistical universals usually are a formulation similar to: "Almost all languages have Y".

The most frequent and fruitful universals in Greenberg's proposal tend to be implicational and statistical universals. The universals proposed by Greenberg could be divided into three main groups:

(1) Syntactic universals about word order. For example, universal 1: "In declarative sentences with nominal subject and object, the dominant order is almost always one in which the subject precedes the object."

(2) Morphological universals about word inflection and derivation. For example, universal 29: "If a language has inflection, it always has derivation".

(3) Morphological universals about word features. For example, universal 36: "If a language has the category of gender, it always has the category of number".

In this paper, we work with Greenberg universals, specifically with the above third group of morphological universals about word features. This means that we will only consider the universals that refer to the linguistic domain of this group: The morphological characteristics present (or not) in the different languages and their correlation.

2.2. *Formal background*

2.2.1. Universal dependencies

Universal Dependencies (UD) [50] is an open repository of homogeneously annotated multilingual corpora. This means that in this resource, we find large collections of different real texts (241) corresponding to different languages (143), in the 2.11 version. The main differentiating aspect of this resource compared to others is the homogeneous annotation. This means that the labeling of the texts in the different languages has been done using the same methodology and the same labels, which facilitates comparison and the drawing of conclusions in multiple languages. For this purpose, the labels and methodology proposed by Google [51] in relation to part of speech are used. On the other hand, for the syntactic analysis of the different sentences, the guidelines and terminology of the Stanford Dependencies [52] are used.

Multiple researchers from all over the world are updating the database with more texts or more languages. Although most languages are Indo-European and it is still difficult to avoid such bias, there is a remarkable effort to try to include languages with a marginal or nonexistent representation in the linguistic tradition. This can provide very interesting data, especially in studies such as the one presented here.

Moreover, the computational annotation of most of the morphological and syntactic data of the analyzed texts allows a quantitative and more objective approach to linguistic phenomena. The claims that can be formulated have a mathematical backing and are more fine-grained. In addition, it also allows new metrics to be obtained in an automatic and efficient way [53].

2.2.2. Grew-Match

The morphological and syntactic data in the texts of the different UD languages are annotated in text, which is not processed or normalized. In order to know the invariance of occurrences, all of this data must be automatically cross-checked. Therefore, Grew-Match is ideal since it is able to manage non-normalized information text from UD efficiently [54].

This tool has both an online interface and a Python implementation. It allows both the query of linguistic occurrences in a given corpus and a comparison of the results in multiple languages. Both quantitative results and qualitative examples of occurrences can be accessed at the same time. In order to carry out the queries, one must know the tool's formal language, which will be the syntax that supports the labels of the UD annotation system.

However, the tool also has an alternative labeling system to the original UD, named Surface Universal Dependencies (SUD), which is the one we use. In this case, it is an updated and improved version of syntactic annotation. It offers a representation with a higher weight of syntactic criteria and a lower semantic weight when deciding which word acts as head and which word acts as dependent.

The queries that can be performed are unrestricted and can contain different complex structures within themselves. The result of the query makes it possible to obtain quantitative data on the occurrences of specific linguistic structures in real texts, which allows comparison and the formulation or revision of universals. In addition, the mere formalization to be carried out is already of great interest since it will enable us to offer a linguistic formalization.

2.2.3. Evaluative expressions

We propose to compute the assumed concepts of linguistic *universality* and *complexity* in a continuum with natural language words.

Fuzzy natural logic (FNL) is based on six fundamental concepts, which are the following: The concept of *fuzzy set*, *Lakoff's universal meaning hypothesis*, the *evaluative expressions*, the concept of *possible world* and the concepts of *intension* and *extension*. The most remarkable aspect of this work is the theory of *evaluative linguistic expressions*.

An evaluative linguistic expression is defined as an expression used by speakers when they want to refer to the characteristics of objects or their parts [55–61], such as *length*, *age*, *depth*, *thickness*, *beauty*, *kindness*, among others. We will consider "*universality*" and "*complexity*" as evaluative expressions.

FNL assumes evaluative linguistic expression with the general form of:

$$\langle \textit{intensifier} \rangle \langle \textit{TE-head} \rangle. \tag{2.1}$$

$\langle \textit{TE-head} \rangle$ (head of a trichotomous evaluative linguistic expressions) can be grouped to form a *fundamental evaluative trichotomy* consisting of two antonyms and a middle term; for example, $\langle good, normal, bad \rangle$. We will consider the trichotomy of $\langle low, medium, high \rangle$.

FNL has been applied in linguistics in the work of Torrens-Urrutia et al. [62–64]. In [65], the study of linguistic universals and complexity through the use of fuzzy evaluative expressions displays the membership scale of universality in linguistic rules recognizing:

- *High Satisfied Universal*. Linguistic rules that trigger a *high* truth value of satisfaction in a set of languages therefore, found satisfied in quasi-all the objects of a set.
- *Medium Satisfied Universal*. Linguistic rules that trigger a *medium* truth value of satisfaction in a set of languages.
- *Low Satisfied Universal*. Linguistic rules that trigger a *low* truth value of satisfaction in a set of languages.

The value of complexity is usually computed as a negation of the value of universality, defining its correlation and by using $IF - THEN$ rules such as:

We characterize fuzzy $IF - THEN$ rules for complexity as follows:

- IF *a rule* is *a high universal* THEN *the value of complexity* is *low*.
- IF *a rule* is *a medium universal* THEN *the value of complexity* is *medium*.
- IF *a rule* is *a low universal* THEN *the value of complexity* is *high*.

Similarly, we can express:

- IF *the value of complexity* is *high* THEN *the rule* is *a low universal*.
- IF *the value of complexity* is *medium* THEN *the rule* is *medium universal*.
- IF *the value of complexity* is *low* THEN *the rule* is *high universal*.

The membership scale of complexity in linguistic rules is [66]:

- *Low Complexity*. Linguistic rules that have a *high* truth value in terms of weight in a set of linguistic rules.
- *Medium Complexity*. Linguistic rules that have a *medium* truth value in terms of weight in a set of linguistic rules.
- *High Complexity*. Linguistic rules that have a *low* truth value in terms of weight in a set of linguistic rules.

A *possible world* is defined as a specific context in which a linguistic expression is used. In the case of evaluative expressions, it is characterized by a triple $w = \langle v_L, v_S, v_R \rangle$. Without loss of generality, it can be defined by three real numbers $v_L, v_S, v_R \in \mathbb{R}$, where $v_L < v_S < v_R$.

Intension and extension: Our intension will simply be the membership degree [0-1], while our extension will depend on the number of languages we consider in a representative set for evaluating universality and complexity.

Figure 1 is the representation of which we will base our work in this paper for interpreting values with evaluative expressions.

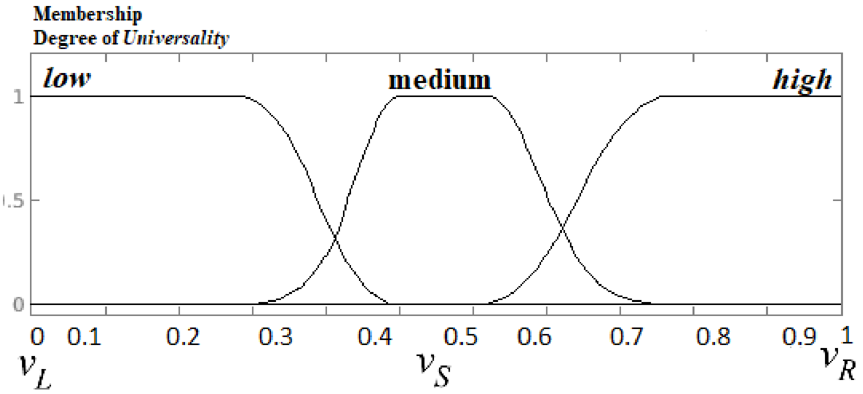

Figure 1. Linguistic universality as an evaluative expression.

We have established a theoretical partition of the possible world:

- Being impossible to find a real number scale for a context, such as what happens when evaluating temperature o speed, we establish an abstract context of a degree 0-1, usually understanding the y-axis as a membership degree of universality and the x-axis as the number of possible languages of an evaluative set.

- We respect the structure of an evaluative expression, two antonyms in each contrary set and one middle term; the middle term shares space with each antonym representing a transition between sets.
- Our strict theoretical tripartition can be defined as:
  - *small* 0-0.4.
  - *medium* 0.41-0.6.
  - *big* 0.61-1.

3. Materials and methods

Figure 2 represents the process of our research regarding its materials and methods.
Figure 2 has to be interpreted in two main parts:

- The materials:
  - Greenberg's universals (in blue), Grew Tool (in orange) and Universal Dependencies (in green).
- The methods to calculate:
  - Theoretical universality and complexity of Greenberg's universals and languages (in pink).
  - Relative universality and complexity of Greenberg's universals and languages (in yellow).
  - Application of the theory of fuzzy evaluative expressions for computing with words results of universality and complexity of natural languages.

Regarding the materials, we distinguish three steps:

- We have made a selection of Greenberg's universals.
- We have formalized Greenberg's universals with the Grew Tool.
- We have prepared a dataset of 146 languages annotated with Universal Dependencies in which we have been searched for the satisfaction, violation or non-applicability of Greenberg's universals formalised with the Grew Tool.

Regarding the method, we distinguish three main parts:

1) Method for computing theoretical weight of universality and complexity of Greenberg's universals.
2) Method for computing relative weight of universality and complexity of Greenberg's universals.
3) Method for computing theoretical and relative weight of universality and complexity of language, expressing such results in words with the theory of the fuzzy evaluative expressions.

Regarding these three parts, we will obtain the following results:

1) Evaluation w.r.t. the weight of theoretical universality and complexity of Greenberg's universals.
2) Evaluate w.r.t. the weight of relative universality and complexity of Greenberg's universals.
3) Taking into account the results of point one and two, we are able to:

- Evaluate theoretical, universality and complexity of languages, language families and languages grouped by basic word order dominance through Greenberg's universals.
- Evaluate theoretical and relative complexity of languages, language families and languages grouped by basic word order dominance through Greeneberg's universals.

We explain each of these parts with more detail in the following Subsections 3.1 and 3.2.

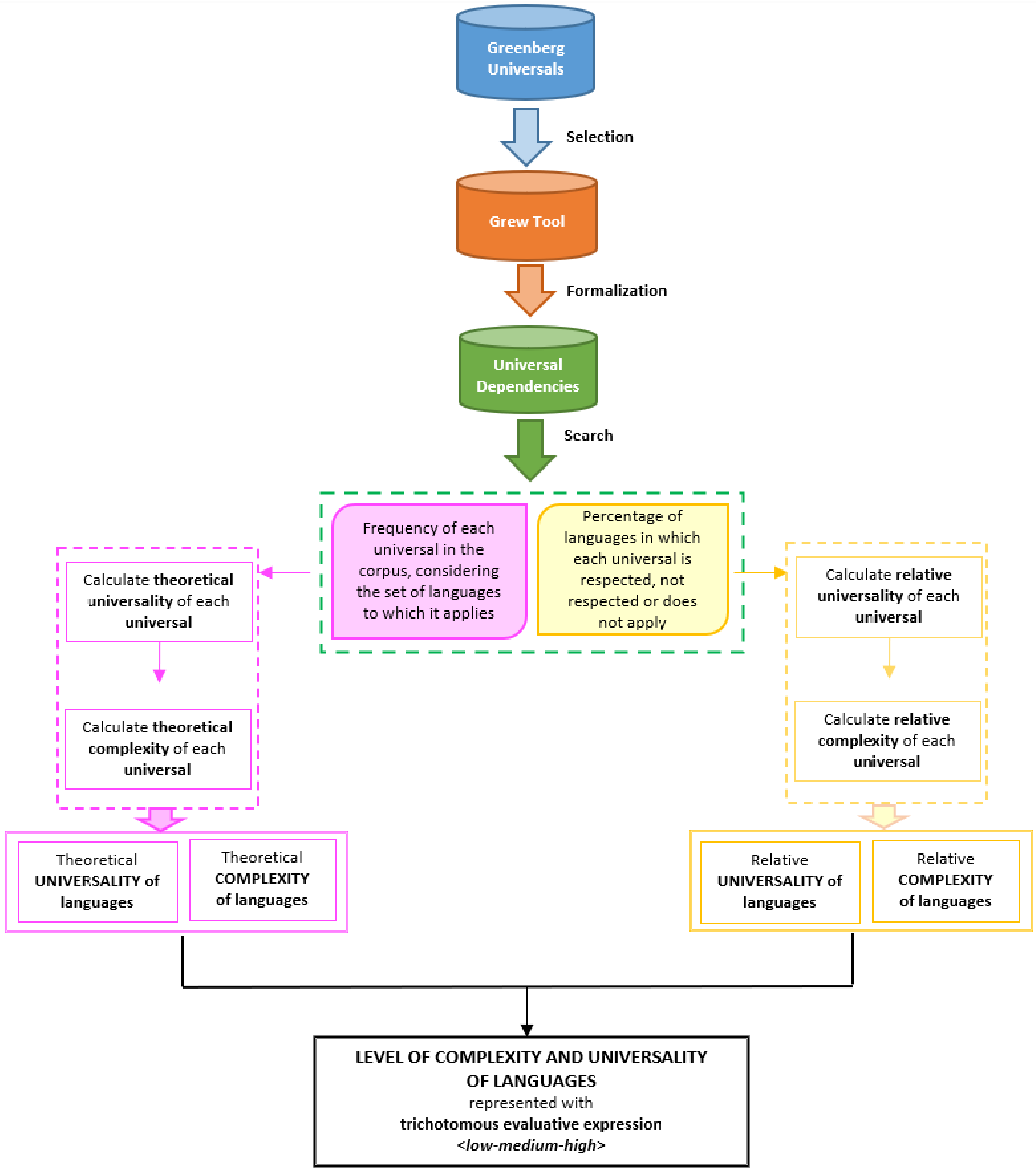
Greenberg Universals
Selection
Grew Tool
Formalization
Universal Dependencies
Search
Frequency of each universal in the corpus, considering the set of languages to which it applies
Percentage of languages in which each universal is respected, not respected or does not apply
Calculate theoretical universality of each universal
Calculate theoretical complexity of each universal
Calculate relative universality of each universal
Calculate relative complexity of each universal
Theoretical UNIVERSALITY of languages
Theoretical COMPLEXITY of languages
Relative UNIVERSALITY of languages
Relative COMPLEXITY of languages
LEVEL OF COMPLEXITY AND UNIVERSALITY OF LANGUAGES
represented with
trichotomous evaluative expression
<low-medium-high>

Figure 2. Diagram of materials and methods to compute degree of theoretical and absolute universality and complexity of Greenberg's universals and natural languages.

3.1. Materials

3.1.1. Data: Greenberg universals

To find a balance between Greenberg's universals, and the data available, we have to disregard those Greenberg's universals hardly evaluable under the combination of the Grew Tool and UD Dependencies (UD) corpus.

UD corpus is one of the most reliable, well-annotated, vast and accessible available data we can work with for linguistic studies. However, all 45 Greenberg's universals cannot be computationally analyzed in UD. Therefore, the data and its formalism condition the Greenberg's universals to be evaluated and used in our research.

We disregard:

- Greenberg's universals that cannot be formalized with UD annotation schemes since this information is not labeled in this repository. These universals correspond mainly to the morphological universal groups regarding rules of inflection and derivation of words.
- Those universals related to intonation and similar aspects not covered in UD or in any corpora that gathers written language as linguistic data.
- Those universals from the group of syntactic universals evaluating word order [73, 74]. Additionally, the influence of the word order constraints evaluating the interplay between complexity and universality has already been shown in [65]. Therefore, we characterize those universals that fall under the morpho-syntactic domain since they are the ones that have been less carefully studied to evaluate the tandem complexity-universality through their features.

Therefore, we are left with the group of morphological universals considering word features since they adapt to our materials, and we have yet to find any information and previous research on them being evaluated and used to compute linguistic universality and complexity. Therefore, we work with the following universals:

- *Universal 30*. If the verb has categories of person-number or if it has categories of gender, it always has tense-mode categories.
- *Universal 31*. If either the subject or object noun agrees with the verb in gender, then the adjective always agrees with the noun in gender.
- *Universal 32*. Whenever the verb agrees with a nominal subject or nominal object in gender, it also agrees in number.
- *Universal 34*. No language has a trial number unless it has a dual. No language has a dual unless it has a plural.
- *Universal 36*. If a language has the category of gender, it always has the category of number.
- *Universal 40*. When the adjective follows the noun, the adjective expresses all the inflectional categories of the noun.
- *Universal 42*. All languages have pronominal categories involving at least three persons and two numbers.
- *Universal 43*. If a language has gender categories in the noun, it has gender categories in the pronoun.

3.1.2. Formalization of Greenberg's universals with Grew tool

To obtain the results of each universal for each language in the UD corpora, we must convert Greenberg's natural language formulations into a more abstract formalization compatible with the terminology used in the systems mentioned above. This is possible thanks to the use of Grew tool 2.2.2.

Due to a lack of space, we provide two examples of the functioning of the formalization process of universals. In the rest of the cases, the functioning is the same. We must carry out the query in Grew-Match (freely available), which will allow us to obtain the occurrences of each language in relation to the analyzed feature. Once this data is obtained, it can be determined whether the universal is fulfilled in the different languages. In the case of Universal 30:

– If the verb has categories of person-number or if it has categories of gender, it always has tense-mode categories.

If the universal is correct in each of the languages, these conditions must be met:

(1) There must be the same amount or less of person-number than tense-mode.

(2) There must be the same amount or less of gender than tense-mode.

Therefore, *Universal 30* states that we cannot (or shouldn't) find verbs with person-number or gender that do not have tense-mode. Thus, the world languages' verb forms may possess both features (having tense-mode and gender or person-number), one of the features (having tense-mode) or none of those features. In formal terms, we could propose this universal as:

$$U30 = A \vee B \implies C, \tag{3.1}$$

where *A* is understood as "person-number", *B* is understood as "gender" and *C* is understood as "tense-mode."

First of all, we have to formalize the presence of person-number in the verbs of the world languages, for which we propose the formalization (3.2):

$$\begin{gathered} \%30 - V - Person - number \\ pattern\{V[upos = VERB, Person, Number]\}. \end{gathered} \tag{3.2}$$

Equation (3.2) reads as follows. In the first line, headed by the symbol %, we find the title of the formalization or the information to identify it. In this case, we have named it *30-V-Person-number*, which refers to the number of universals that such a structure contains, the part of speech affected (verb) and the features that such an object has (person-number). Subsequently, we open the call to the occurrence filter through the word *pattern*, which will restrict the search to the structure enunciated within "{" and "}". Within these symbols, we activate an element that we name *V* for simplification, corresponding to the characteristics within the symbols "[" and "]". In these symbols, we indicate the part of speech that we want to restrict *(upos=VERB)* and, subsequently, separated by "," we indicate the characteristics that this verb must have. In this case, we ask for any value corresponding to the characteristics of *Person* and *Number* to be active.

To formalize the presence of gender in verbs, we use the pattern (3.3), and for the formalization of tense-mode, we employ the pattern (3.4):

$$\begin{gathered} \%30 - V - gender \\ pattern\{V[upos = VERB, Gender]\} \end{gathered} \tag{3.3}$$

$$\begin{gathered} \%30 - V - Tense - Mood \\ pattern\{V[upos = VERB, Tense, Mood]\}. \end{gathered} \tag{3.4}$$

Equations (3.3) and (3.4) read their structure as Eq (3.2). Only the characteristics displayed for the verbs are changed to the desired ones.

Once the first universal has been exemplified, we must apply the same type of searches using Grew's syntax to search for the characteristics we want. For example, in the case of universal 36:

– If a language has the category of gender, it always has the category of number.

We must formalize both categories independently. That is, what we show in (3.5) and (3.6):

$$\begin{gathered} \%36 - Gender \\ pattern\{Gender[Gender]\} \end{gathered} \tag{3.5}$$

$$\begin{gathered} \%36 - Number \\ pattern\{Number[Number]\}. \end{gathered} \tag{3.6}$$

Once we obtain the occurrences in the different languages, we will know whether such categories apply to each analyzed language. We understand this universal as a simple implication:

$$U36 = A \implies B. \tag{3.7}$$

If we find A (Gender) in a language, we will always find B (Number). This implication tells us that we will not find languages with gender that do not also have number (something that is possible the other way around).

Through these two examples of formalization, the rest of the universals we work with can be understood and retrieved. If the premise of the implication of the universal is double, this structure can be rescued from the universal 30. If the premise is simple, the structure can be rescued from the universal 36. The only universal without the implicational structures shown above is universal 42, which we can formalize as:

$$U42 = A + B. \tag{3.8}$$

This can be understood as meaning that any language contains A and B. By A we mean three features or more of pronominal person and by B we mean two features or more of pronominal number.

3.1.3. Data: Set of languages and universal dependencies

A selection of languages is mandatory, as it is impossible to analyze the totality of the world's languages for two main reasons. First, we are still determining exactly how many languages there are in the world, as languages are constantly being born and dying without our knowledge [67]. Second, there has yet to be an agreement on the distinction between dialects and languages [68]. Therefore, we do not know the totality of the world's languages [69]. Additionally, for many non-Indo-European languages, even though we have a label for them, we don't have reliable and normalized scientific data [70].

Therefore, studies of linguistic universality and complexity have to make a representative selection of the world's languages that will allow us to extrapolate the data obtained. Depending on the type of study carried out, this representation may have different characteristics [71]:

- *Convenience Sampling*. If data availability is inadequate, a perfect balance cannot be guaranteed. However, the results may be indicative of a clear universal trend.
- *Variety Sampling*. There is a wide availability of language data, yet the phenomenon needs to be better studied. Representative languages of different linguistic types and genetic backgrounds and areas are selected, also including languages that are characteristically untypical examples in the language set.
- *Probability Sampling*. If the availability of data from different languages is reliable, normalized and correct and we want to know the representativeness of a phenomenon, we must balance the selection to maintain an equilibrium of linguistic type, linguistic family and area.

In our case, we have created a dataset of convenience sampling. We have worked with 241 corpora corresponding to 143 different languages of UD 2.2.1. We analyze the totality of the available corpus for two reasons:

(1) First, given the pioneering nature of the study, it is interesting to extend the results to the maximum possible number of languages to gain information on those not represented.

(2) Second, recent studies guarantee that with a varied number of languages, it is not necessary to establish any corrective measure for sampling, as the results are the same [72].

In our research, the totality and detailed list of the analyzed languages can be checked in [50]. 46% of the languages are from the Indo-European family, a common bias in typology studies. However, all the different macro-areas of the world (except Australia) are present: Papunesia, Eurasia, North America, South America and Africa. We also find some dead languages (Latin, Sanskrit, Ancient Greek, ...) and several isolated languages (Basque, Japanese,...). Another of the most interesting aspects of the set of languages analyzed is the presence of unusual varieties such as Creoles, code-switching languages or sign languages.

To guarantee the neutrality of the selection and the falsifiability of the data used, we have analyzed the first 1,000 sentences of each of the 241 corpora.

3.2. Methods

We perform three main tasks:

1) Computing theoretical universality and complexity of Greenberg's universals.
2) Computing relative universality and complexity of Greenberg's universals.
3) Computing theoretical, relative universality and complexity of natural languages.

We explain each of these tasks in the following subsections:

3.2.1. Theoretical universality and complexity of Greenberg's universals

The weight of theoretical universality (GU_T) of Greenberg's universals is computed by checking all the languages in which the universal is satisfied (All_{L_s}), divided by the set of languages to which the universal applies ($All_{L_{app}}$) (Eq (3.9)):

$$GU_T = \frac{All_{L_s}}{All_{L_{app}}}. \tag{3.9}$$

When we refer to the fact that a universal does not apply to a language, this can be due to multiple reasons. In short, this means that in an analyzed language L, the elements cited in the universal are not present and, therefore, it is not testable. If it does apply, on the other hand, these elements are present and we check whether Greenberg's proposal is satisfied or violated. The theoretical complexity of a Greenberg's universal (GC_T) is computed as a negation of the weight of theoretical universality (GU_T) in Eq (3.10):

$$GC_T = -GU_T + 1. \tag{3.10}$$

Therefore, we establish a co-relation in which the more universal a language is, the less complex it is. Thus, the language sharing more rules with all the other languages is theoretically less complex than the language that shares less of the universals concerning the rest of the set of languages.

To estimate the degree of universality and complexity, we can apply fuzzy/linguistic IF-THEN rules. Using them, we can replace evaluation using numbers with words.

In this case, the degree of universality and complexity can be estimated using fuzzy/linguistic IF-THEN rules as follows:

- IF *a Greenberg's universal* is *highly satisfied* THEN *the degree of universality* is *high.*
- IF *a Greenberg's universal* is *quite satisfied* THEN *the degree of universality* is *medium.*
- IF *a Greenberg's universal* is *barely satisfied* THEN *the degree of universality is* is *low.*

Similarly, we can express:

- IF *the degree of universality* is *high* THEN *the degree of complexity* is *low.*
- IF *the degree of universality* is *medium* THEN *the degree of complexity* is *medium.*
- IF *the degree of universality* is *low* THEN *the degree of complexity* is *high.*

3.2.2. Relative universality and complexity of Greenberg's universals

We compute the weight of relative universality and complexity of Greenberg's universals, considering how many languages each universal is satisfied, violated or does not apply.

The value of a satisfied relative universality of a Greenberg's universal (GU_R) is computed by checking how each universal behaves in all our sets of languages. As a result, each universal has a relative weight for each state: A weight of satisfaction, violation and non-applicability.

The value of relative universality satisfaction of a universal (GU_{RS}) is computed by considering all the languages in which the universal is satisfied (All_{L_s}), divided by our full set of languages (All_L) (Eq (3.11)):

$$GU_{RS} = \frac{All_{L_s}}{All_L}. \tag{3.11}$$

The value of relative universality violation of a universal (GU_{RV}) is computed by considering all the languages in which the universal is violated (All_{L_v}), divided by our full set of languages (All_L) (Eq (3.12)):

$$GU_{RV} = \frac{All_{L_v}}{All_L}. \tag{3.12}$$

The value of relative universality non-applicability of a universal (GU_{Rnapp}) is computed by considering all the languages in which the universal is non-applicable (All_{L_napp}), divided by our full set of languages (All_L) (Eq (3.13)):

$$GU_{Rnapp} = \frac{All_{L_napp}}{All_L}. \tag{3.13}$$

The value of relative complexity of a Greenberg's universal (GC_R) is computed as a negation of the weight of relative universality for three of the behaviors of a Greenberg's universal GU_{RS} as in (3.14), GU_{RV} as in (3.15) and GU_{Rnapp} as in (3.16).

$$GC_R = -GU_{RS} + 1 \tag{3.14}$$

$$GC_R = -GU_{RV} + 1 \tag{3.15}$$

$$GC_R = -GU_{Rnapp} + 1. \tag{3.16}$$

Therefore, we establish again a correlation between linguistic universality and complexity. The more universal a language is, the less complex it is since it shares more rules with all the other languages. We can express our results computing with words as with those fuzzy evaluative expressions mentioned above.

3.2.3. Theoretical universality and complexity of languages

We have based this step on the results from calculating the theoretical and relative universality and complexity of Greenberg's universals. Therefore:

- If the language applies and satisfies a Greenberg's universal, the language adds a value of one to its weight.
- On the contrary, if a language does not satisfy nor apply the universal, the universal does not add any value to its weight. Therefore, such universal weight is zero with respect to the language.

Consequently, we compute the theoretical value of universality of a language (LU_T) by taking into account all the satisfied universals (All_{GU_S}) in it, divided by all the Greenberg's universals of our set (All_{GU}).

$$LU_T = \frac{All_{GU_S}}{All_{GU}}. \tag{3.17}$$

On the other hand, we compute the value of complexity of a language (LC_T) again as the negation of its universality, as in Eq (3.18.):

$$LC_T = -LU_T + 1. \tag{3.18}$$

Table 1 is an example of a calculation of theoretical universality and complexity.

Table 1. Example of computing theoretical universality and complexity of languages.

Language	u30	u31	u32	u34	u36	u40	u42	u43	Theoretical Universality	Theoretical Complexity	Fuzzy Evaluative Expressions
Slovenian	1	1	0	1	1	1	1	1	0,875	0,125	High-Low
Wolof	1	0	1	0	1	0	1	0	0,5	0,5	Medium
Guarani	0	0	1	0	0	0	0	0	0,125	0,875	Low-High

3.2.4. Relative universality and complexity of languages

We have taken the value of relative universality (GU_R) that corresponds to each language per each universal.

Following the example of Table 1, U30 has the relative values if satisfied weights 0.56, if violated 0.07 and if non-applicable 0.37. Naija computes U30 as satisfied; therefore, it has a value of 0.56. In Japanese, U30 is non-applicable, adding a value of 0.37 for U30. In Arabic, U30 is violated, so it adds 0.07. The same applies to the rest of the universals. Therefore, the final value of relative universality of a language (RU_L) is computed as the addition of all the relative values of Greenberg's universals in the language ($All_{GU_R}\ in\ _L$), divided by all the set of Greenberg's universals (All_{GU}), as shown in Eq (3.19):

$$RU_L = \frac{All_{GU_R}\ in\ _L}{All_{GU}}. \tag{3.19}$$

On the other hand, the relative complexity of a language is computed as a negation of the relative universality of that same language, as in Eq (3.20):

$$RC_L = -RU_L + 1. \tag{3.20}$$

To compute the values of theoretical and relative universality and complexity of language families and word order dominance groups, we have applied the calculations by grouping all the values of each language in their families, or word order dominance groups, and dividing it by the total amount of objects.

4. Results

4.1. Theoretical universality and complexity of Greenberg's universals

Table 2 shows the theoretical universality and complexity values for each of Greenberg's universals. We confirm that universals U31, U32, U34, U36 and U43 are *highly* satisfied. Only U42 displays a tendency toward showing a *medium* value of satisfaction, but still belonging to the set of *high universal*. Differently, U30 and U40 do belong to the set of a *medium* universal. Therefore, U40 and U30 could be questioned as a linguistic universal. The rest of the universals fall under the spectrum of 0.7 and one; thus they are on the set of *high* universal and, consequently, are definitely a universal rule from the theoretical point of view.

Table 2. Degree of theoretical universality and complexity of Greenberg's universals.

Degree of theoretical universality and complexity of Greenberg's universals						
Greenberg's Universals	Set of Languages	Languages in which applies	Satisfied	Violated	Theoretical Universality	Theoretical Complexity
u30	143	85	37	48	0,43	0,57
u31	143	94	94	0	1	0
u32	143	119	108	11	0,91	0,09
u34	143	10	10	0	1	0
u36	143	76	76	0	1	0
u40	143	38	16	22	0,42	0,58
u42	143	92	64	28	0,7	0,3
u43	143	63	60	3	0,95	0,05

Figure 3 shows the same information as Table 2, but in the shape of a fuzzy evaluative expression graph. Y-axis is the degree of membership of a Greenberg's universal according to how many applicable languages a universal has been satisfied. X-axis displays the conversion of the applicable language for each universal to 100. The left set defines the spectrum of low universality, the medium set represents medium universality and the right set represents high universality.

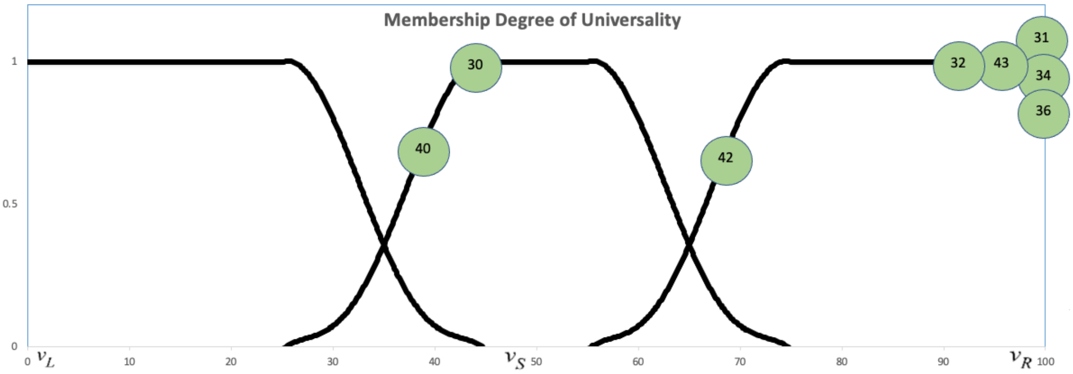

Figure 3. Degree of theoretical universality and complexity of Greenberg's universals in the form of an evaluative expression graph.

4.2. *Relative universality and complexity of Greenberg's universals*

Table 3 presents the relative weight of satisfaction, violation and non-applicability of each of Greenberg's universals. We mark the highest value on green and on red, the lowest. Therefore, we observed that most languages satisfy Greenberg's universals. On the other hand, those universals that are non-applicable in many languages have the highest weight on the label of non-applicable. Therefore, no universal with a higher weight of violation exists, showing that Greenberg wasn't entirely wrong in any of those universals.

Table 3. Degree of relative universality and complexity of Greenberg's universals.

Degree of relative universality and complexity of Greenberg's universals										
Geenberg's Universal	Set of languages	YES	NO	NOTAPP	Relative Universality			Relative Complexity		
					Weight of Satisfaction	Weight of Violation	Weight of Non-applicable	Weight of Satisfaction	Weight of Violation	Weight of Non-applicable
u30	126	37	48	41	0,29	0,38	0,37	0,71	0,62	0,66
u31	143	94	0	49	0,66	0,00	0,34	0,34	1,00	0,66
u32	143	108	11	24	0,76	0,08	0,17	0,24	0,92	0,83
u34	123	10	0	113	0,08	0,00	0,92	0,92	1,00	0,08
u36	143	76	0	67	0,53	0,00	0,47	0,47	1,00	0,53
u40	143	16	22	105	0,11	0,15	0,73	0,89	0,85	0,27
u42	143	64	28	51	0,45	0,20	0,36	0,55	0,80	0,64
u43	143	60	3	80	0,42	0,02	0,56	0,58	0,98	0,44

We propose two interpretations for those universals that display non-applicability as their highest weight regarding both Tables 2 and 3. First, we consider those universals that fall on non-applicability over satisfaction, such as U34, U40 and U43, as a mistake from Greenberg by proposing universals that are too specific or, on the other hand, taking into account the general trend of Table 2 and considering that, in case those universals would ever be applicable in all languages, we predict that the tendency would be to fall on a hard satisfaction weight.

On the other hand, only U30 falls on the weight of violation over satisfaction and non-applicability. More than half of the languages where this universal applies in our set of languages do not necessarily have tense-mode categories when they bear either person-number or gender. Therefore, Greenberg created a universal that could be questioned.

4.3. *Theoretical universality and complexity of languages*

Table 4 displays a classification of our set of 143 languages classified by their theoretical degree of universality and complexity. It is described with a fuzzy evaluative expression below in the last row. For example, Slovenian and Turkish German fall on the spectrum of being highly universal and not so complex from an absolute theoretical point of view. Bengali and Guarani fall on the spectrum of having a low value of universality and, therefore, being highly complex. This result expresses how many of Greenberg's universals are satisfied or not in each language. Thus, universality and complexity are always interpreted strictly from Greenberg's perspective in this results.

Table 4. Degree of theoretical universality and complexity of languages.

	Degree of theoretical universality and complexity per languages						
Universality-Complexity Value	*0,875-0,125*	*0,75-0,25*	*0,625-0,375*	*0,5-0,5*	*0,375-0,625*	*0,25-0,75*	*0,125-0,875*
Languages	Slovenian	Turkish_German, English, Danish, Welsh, Umbrian, Old_Church_Slavonic, Irish, Upper_Sorbian, Ukrainian, Ancienthebrew, Latvian	Naija, Manx, Kazakh, Hittite, North_Sami, Erzya, Moksha, German, Lowsaxon, Scottish_Gaelic, Faroese, Norwegian, Greek, Latin, Italian, Czech, Breton, Belarussian, French, Kurmanji, Spanish, Russian, Icelandic, Croatian, Serbian, Pomak, Gothic, Ligurian, Lithuanian, Ancientgreek, Slovak, Arabic, Bulgarian	Uyghur, Estonian, Karelian, Western_Sierra_Puebla_Nahuatl, Xibe, Turkish, Komizyrian, Amharic, Buryat, Persian, Coptic, Wolof, Swedish, Gheg, Yakut, Galician, Sinhala, Tamil, Marathi, Hindi, Sanskrit, Hebrew, Albanian, Catalan, Polish, Armenian, Portuguese, Western_Armenian, Oldeastslavic, Romanian	Tatar, Finnish, Karo, Livvi, Yoruba, Zaar, Komipermayah, Skolt_Sami, Korean, Mbyaguarani, Bambara, Apurina, Cebuano, Malayalam, Nheengatu, Beja, Indonesian, Kiche, Afrikaans, Hungarian, Basque, Assyrian, Urdu, Bhojpuri	Tupinamba, Sud_Chinesepud, Bengali, Sud_Chinesegsd, Sud_Chinesegsdsimp, Tagalog, Warlpiri, Khunsari, Nayini, Thai, Cantonese(Hk), Sud_Chinesecfl, Sud_Chinesehk, Sud_Chinesepatentchar, Sud_Chukchihse, Frisiandutch, Hindi_English, Kangri, Maltese, South_Levantine_Arabic, Swedish_Sign_Language, Swiss_German, Telugu, Vietnamese, Abaza, Yupik, Old_French, Dutch, Javanese, Akkadian, Madi	Teko, Akuntsu, Guarani, Kaapor, Soi, Xavante, Makurap, Munduruku, Japanese, Neapolitan, Old_Turkish, Guajajara, Sud_Classical_Chinesekyoto
Evaluative Expression	*High Universality, Low Complexity*		*Medium Universality and Complexity*		*Low Universality, High Complexity*		

4.3.1. Degree of theoretical universality and complexity of language families

Figure 4 shows a classification of theoretical universality (y-axis) and complexity (x-axis) per language family. By far, Creole and Indoeuropean language families are the most universal and less complex ones. Creole is usually hardly influenced by Indo-European languages; therefore, it is logical that they fall on a similar spectrum. The less universal and more complex are Sino-Tibetan, Tupian and Japanese.

There are two possible interpretations of these graphs: Indo-European languages are the largest group and, therefore, the one with the most possibilities of being the universal one, or Greenberg's universals are biased toward Indo-European languages, as well as our data.

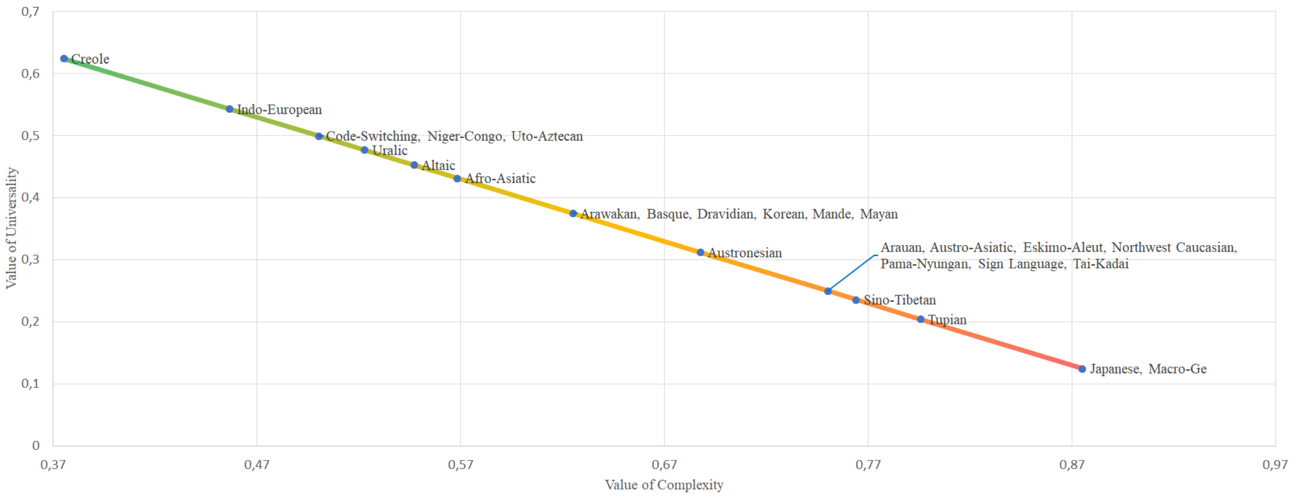

Figure 4. Degree of theoretical universality and complexity of language families.

4.3.2. Degree of theoretical universality and complexity of basic word order dominance of languages

Figure 5 shows a classification of theoretical universality (y-axis) and complexity (x-axis) per basic word order dominance. The groups are distributed throughout the whole gradience. OVS (Object-Verb-Subject) and VSO (Verb-Subject-Object) satisfy almost entirely all Greenberg's universals, and NDO (Non dominant order) and SOV (Subject-Object-Verb), the group of languages that meet Greenberg's universals, satisfy less.

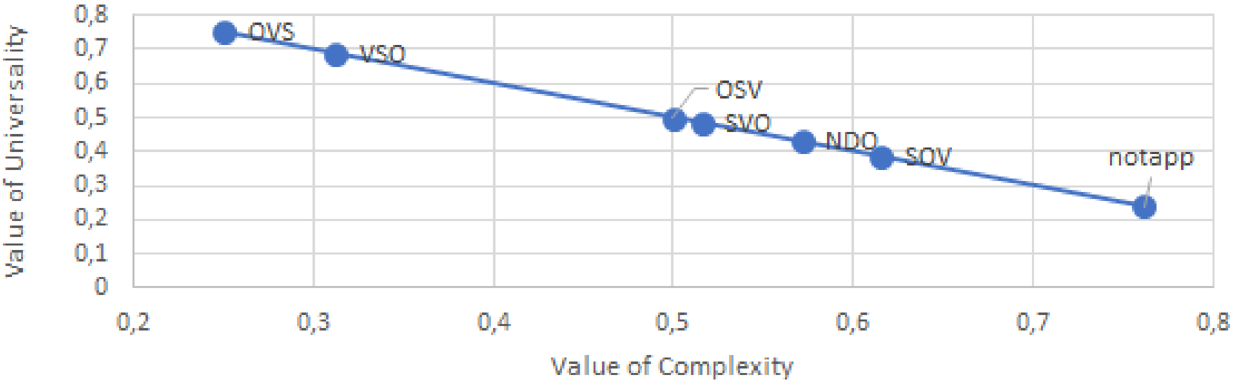

Figure 5. Degree of theoretical universality and complexity of basic word order dominance.

4.4. Relative universality and complexity of languages

Figure 6 displays a radar chart with a degree of relative universality and complexity per language with colors of evaluative expressions: Green-high/low, orange-medium and red-low/high. The angular axis displays the value of complexity, while the radial axis displays the value of universality. Most languages have a medium weight of relative complexity and universality, while almost none have a high complexity value (in red). This distribution coincides with Table 4. However, the languages are very differently distributed, such as what happens with Slovenian, which was the most universal from a theoretical point of view. It appears as not a universal one from the viewpoint of relative universality. However, in Table 4, we are evaluating the theoretical universality of language according to only Greenberg's universals, and in Figure 6, we are evaluating universality and complexity concerning the relative weights, meaning Slovenian was very solid on a more discrete counting (only one and zero). At the same time, it is not that similar to the other languages on a more fuzzy counting, considering different weights concerning the behavior of the rest of the languages.

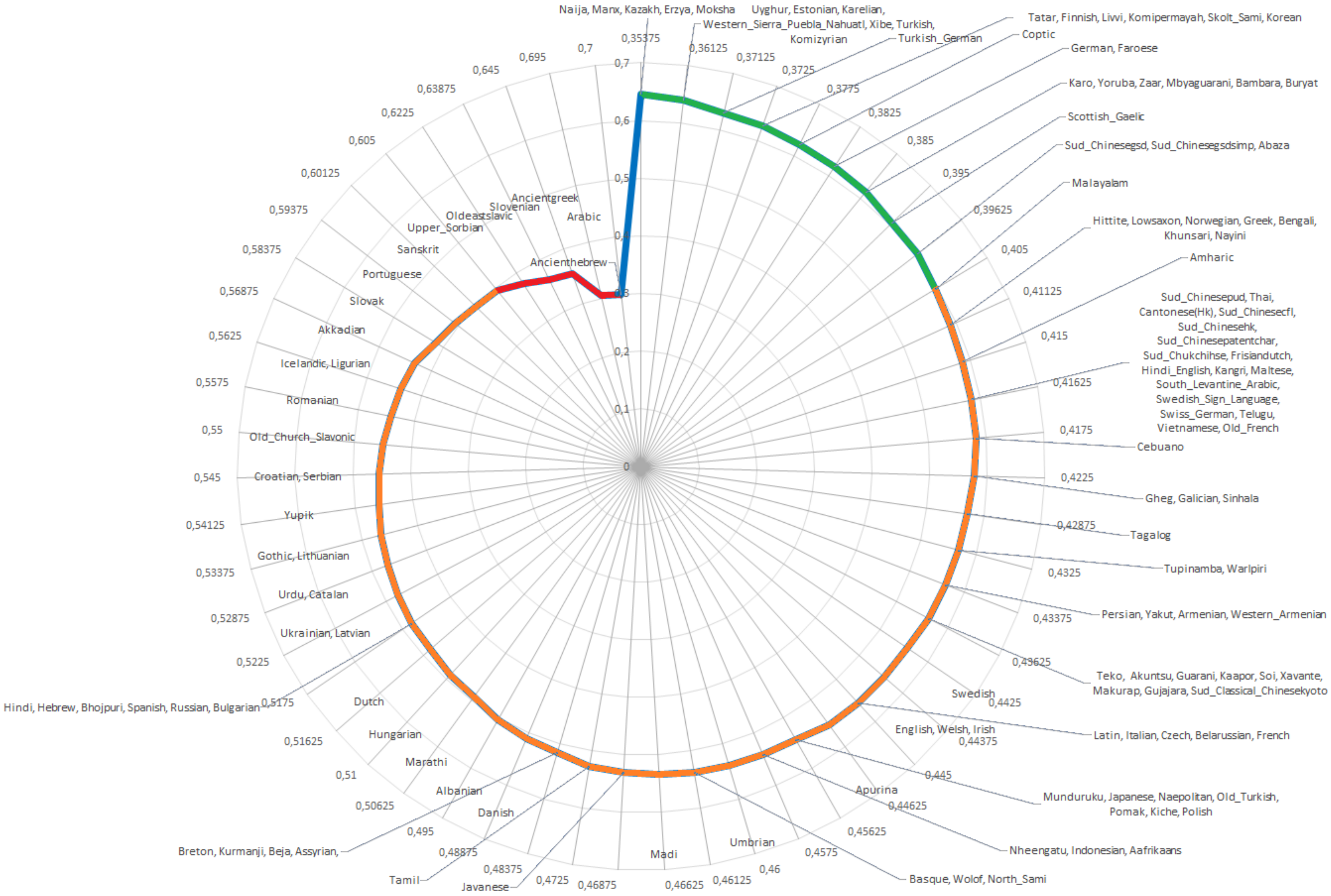

Figure 6. Degree of relative universality and complexity of languages.

4.4.1. Degree of relative universality and complexity of language families

Figure 7 presents a classification of relative universality (y-axis) and complexity (x-axis) per language family. No language falls after a value of the medium. The lowest value is for Eskimo-Aleut, with a relative universality of 0.54. Therefore, all the language families have a high or medium value on relative universality and complexity. In contrast with Figure 4, Creole and Uto-Azteca languages are the most universal. On the other hand, Indo-European languages display a different behavior by being the third less universal family group.

4.4.2. Degree of relative universality and complexity of basic word order dominance of languages

Figure 8 shows a classification of relative universality (y-axis) and complexity (x-axis) per basic word order dominance. The groups are distributed throughout the whole gradience. OSV Greenberg's universals the most. NDO and VSO are the languages that meet Greenberg's universals the least. These results disagree with the data from evaluating theoretical universality and complexity in Figure 5.

A possible interpretation of this data is that basic word order doesn't provide much relevant information regarding language universality and complexity and, therefore, to have a proper gradient classification, we need to dig in with bigger groups, such as groups of languages or families.

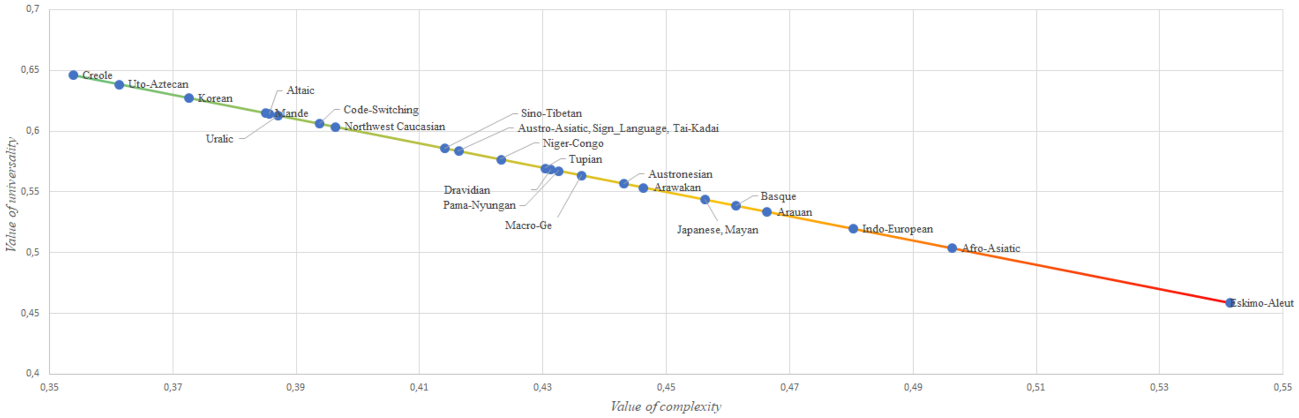

Figure 7. Degree of relative universality and complexity of language families.

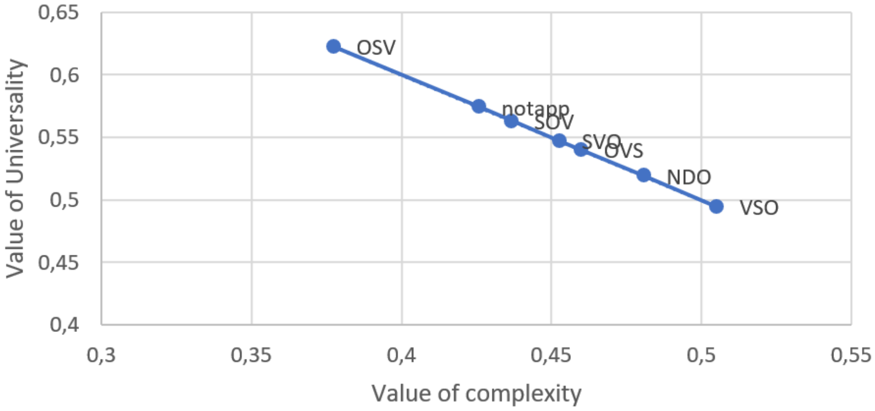

Figure 8. Degree of relative universality and complexity of basic word order dominance.

5. Discussion

The contributions of this study extend across various domains. Noteworthy are its contributions to the field of linguistic complexity, which constitutes the central focus of this paper, as well as its relevance to the realm of linguistic universals and the mathematical theory of evaluative expressions.

In linguistic complexity, we present a mathematical-formal model and employ computational tools to compute relative complexity across real data corpora. This approach enables us to avoid the challenges encountered by studies that determine relative complexity through psycholinguistic experiments (subject to the influence of extralinguistic factors or individual variability) or those grappling with calculating absolute complexity based on grammars (which lacks grounding in authentic data). In fact, as highlighted by Kortmann [24], one issue in the literature on linguistic complexity is that often these works rely on rather unsystematic and intuition-based evidence. When grounded in actual data, they tend to be confined to reference grammars in conjunction with certain typological sampling techniques. By working with real data on 143 languages, we aim to provide a possible solution to the limitations encountered in much of the literature on linguistic complexity.

Another significant contribution of this research lies in our approach to complexity assessment. Unlike a predominant focus on either absolute or relative complexity in much of the existing literature, our work bridges the gap between these two forms of complexity and is presented as disconnected entities. Specifically, we establish a connection between absolute and relative measures of complexity. In this study, based on calculating the degree of universality of a language, we determine its complexity for acquisition by adults who already possess proficiency in a first language. Consequently, we employ absolute measures that allow for assessing complexity in relative terms. This perspective aligns with viewpoints like that of Sinnemäki [75], who highlights the need for further research to explore the interplay between complexity and difficulty through psycholinguistic experiments, asserting that a comprehensive understanding of intricate phenomena necessitates multifaceted investigation.

We establish a correlation between the concepts of linguistic complexity and linguistic universals. Despite the extensive body of work conducted in recent years within the realm of typological studies on complexity [4, 16, 29–33, 75, 76], these concepts have rarely been jointly analyzed in the manner how they are interconnected within this work. On the one hand, we understand the concept of complexity in terms of the difficulty of learning one language from another (second language acquisition); on the other hand, we interpret universals as structures/categories present in all languages. From this standpoint, we establish an inversely proportional relationship between the two concepts: The greater the degree of shared characteristics between two languages, the less challenging it will be to learn one from the other. In essence, the higher the universality of a language, the lower its complexity level when learned as a second language.

The relationship between complexity and universality aligns with the concept of the connection between rarity and complexity [75]. Scholars like Newmeyer [77] and Harris [78] have linked cross-linguistic rarity to linguistic complexity. Miestamo [29] suggests that while a direct correlation between rarity and absolute complexity might not always exist, some level of association between rarity and difficulty, namely, relative complexity, can be anticipated. Hawkins [33] sheds light on this relationship by noting that structures that are easy and efficient in performance tend to grammaticalize more frequently in languages, while those that are complex and inefficient tend to grammaticalize less often. Additionally, Sinnemäki [75] points out that low probability has been tied to complexity [79], and typological rarities (opposite of universals) may consequently demonstrate higher grammatical complexity [78].

The relationship between the level of universality and the degree of language complexity is established in a previous work [80]. Although Greenberg's universals were not the focus of that paper, the same philosophical approach was employed. It is worth noting the advantages of the method presented here compared to the previous work. The prior study examined only nine languages, whereas this work analyzes 143 languages. In our last approach, it was necessary to generate a grammar for each of the analyzed languages, in addition to a universal grammar with 42 billion syntactic constraints. A universality weight was assigned to each rule using each of the generated grammars, and constraints were classified as having low, medium or high universality. This method enabled the determination of language complexity: The most complex languages were those with fewer universal constraints. This analysis required computing a correlation between every pair of languages to determine the complexity levels regarding shared syntactic constraints.

The analysis presented in this current work employs a more abstract and generic concept, that of linguistic universals. Through this concept, we are able to examine more specific characteristics, thereby enabling the grouping of languages by type (three distinct types, varying according to the universal). We work with fewer rules and, therefore, cannot construct a complete grammar, as was done in [80]. However, given our detailed understanding of each language's behavior concerning each universal (due to the smaller scope, it is more manageable), we can gain better insights into the formed language groups. This approach, in turn, allows us to calculate linguistic complexity more effectively.

We introduce a fuzzy approach to both the complexity and universality concepts. This innovative framework enhances their description and classification, providing transparency and coherence with its non-discrete (fuzzy) nature. Concerning complexity, establishing a fuzzy definition and presenting a formal model for calculating its levels is a challenging endeavor. Regarding universals, this fuzzy approach effectively addresses classical terminological challenges in linguistic typology. While authors like Tomlin [81] and Dryer [82] advocate for universals with exceptions and present compelling reasons to engage with them, they often fall short of offering a system capable of classifying and comprehending them as non-discrete entities.

In the context of universals, with a specific focus on Greenberg's universals, we present a formalized approach. Our proposal differs from typological studies, where it is often difficult to encounter formal models and where the prevailing norm uses nonformal and occasionally ambiguous formulations to describe these linguistic regularities [14, 83].

Another contribution of this study is the validation of the universals formulated by Greenberg. We assess the validity of these universals using a quantitative, objective and verifiable methodology. The universals under investigation in this article have yet to undergo an in-depth analysis of the existing literature. While there have been isolated analyses of certain universals [84], a systematic analysis like the one presented in this study, grounded in real-text data, has yet to be previously conducted.

While computational validations of Greenberg's universals are becoming increasingly common, the analysis presented in this paper offers a distinct set of characteristics. First and foremost, computational analyses often tend to be isolated and focus on specific universals, in contrast to our approach in this article. Moreover, the universals typically scrutinized are often associated with Word Order [73, 74], whereas this study delves into a different category of Greenberg's universals: The morphological ones. Furthermore, most approaches to universals still do not employ quantitative methodologies based on occurrences within a corpus of real texts; instead, they often lean toward grammar-based analyses [83].

Finally, we provide fine-grained results for various linguistic types within the different universals, without this approach being incompatible with the categorical approach to validate/refute Greenberg's universals (which we can also and do employ). This more fine-grained analysis enhances precision, showcasing non-prototypical or less canonical cases. This level of granularity is less common in more traditional approaches, such as [14, 85].

Regarding the limitations of our results, the fact of dealing only with 8/45 universals and questioning how well-distributed the universals are across languages (i.e., Indo-European vs. other languages) could seem detrimental to the validity of our study. With this respect, analyzing eight out of 45 universals is significant since there are no experiments or other experiments considering these universals across an extended number of languages, such as our corpus of 146 selected languages (up to our knowledge). Moreover, studies are usually on only two or three universals. Providing research results concerning eight universals in a paper is not usual, either, cf. [48, 73, 74, 86, 87].

On the other hand, the distribution of universals depends on the premise used:

- If the universals are analyzed to find the weight of universals, i.e., which universals are most respected by each of the languages, all universals are highly respected except for U30, U40 and U42.
- On the other hand, if the languages are analyzed individually to find the universality weight of each language by taking into account how many of Greenberg's universals are respected in each of the languages individually, then we find that practically no language has a high level of universality. Additionally, we conclude that the universality weight of Greenberg's rules is based on a homogeneous distribution across the languages rather than on a large group of languages satisfying the universals.
- In our corpus, there are more Indo-European languages than families from other languages since UD are biased toward them. However, we have provided a very varied distribution, which includes 25 different language families. By providing a distribution of universality considering language families as a criterion, we show that Indo-European languages have a lower degree of universality than most other language families. Indo-European languages are very diverse and they display very different features amongst them. Contrarily, other smaller language families, such as Creoles or Korean, gather more universals since smaller language families tend to be more homogeneous.

Otherwise, to clarify the issue regarding the distribution of languages, we can refer to it as a "bibliographical bias". In other words, our study is influenced by the available sources. However, we must emphasize that our options are limited in this regard. It might seem that the number of Indo-European languages is much lower than 46% in WALS (World Atlas of Language Structures)-type databases. Still, it's crucial to consider that while there are entries from various languages, many of these entries contain only limited or nonspecific information. As a result, we need more substantive data regarding the aspects covered by these universals. Given these constraints, our approach is based on the available data within the UD framework. As UD continues to expand, we propose this study method, which can be further refined by incorporating data from other languages in the future.

6. Conclusions

This paper contributed to the studies on linguistic complexity by demonstrating differences in complexity among languages and by providing an objective and valid method to calculate the relative complexity of natural languages, particularly in the context of second language learning (L2) in adults.

To calculate linguistic complexity, the paper introduced the use of computational tools and mathematical models that can provide objective and reliable measures for quantitatively evaluating linguistic complexity.

Overall, the work presented in this paper seeks to advance the understanding of linguistic complexity and offer valuable insights into the nature of language complexity. The use of computational and mathematical tools can contribute to challenge the long-held assumption of equicomplexity among languages and may have significant implications for various areas, including theoretical linguistics, comparative linguistics, language acquisition, second language teaching and language technologies. It suggests that acknowledging differences in linguistic complexity can lead to improved language teaching methods and better-designed language technologies.

Furthermore, this study conducted a cross-validation on Greenberg's universals by utilizing data from 143 languages and formalizing them in a structured manner. Our approach was grounded in the analysis of several languages and relied on a source that is distinct from Greenberg's, involving real texts and actual utterances. We extracted simplified frequencies within a particular grammatical category using authentic texts produced by speakers.

Our objective was to verify the consistency of Greenberg's universals by examining a significantly larger dataset, encompassing 113 additional languages compared to Greenberg's original work. Additionally, by using texts generated by native speakers, we aimed to align our findings with Greenberg's, further validating his proposed universals. Our approach and results serve to reinforce the credibility of Greenberg's proposals, given the alignment of results across this expanded and diversified dataset.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This paper was supported by the project PID2020-120158GB-I00 funded by MCIN/AEI/10.13039/501100011033.

Conflict of interest

The authors have no conflicts of interest.

References

1. G. Deutscher, *Through the language glass: Why the world looks different in other languages*, New York: Metropolitan Books, 2010.

2. P. W. Culicover, *Grammar and complexity: Language at the intersection of competence and performance*, Oxford: Oxford University Press, 2013.

3. T. Givón, M. Shibatani, editors, *Syntactic complexity: Diachrony, acquisition, neuro-cognition, evolution*, Amsterdam: John Benjamins, 2009. https://doi.org/10.1075/tsl.85

4. G. Sampson, D. Gil, P. Trudgill, *Language complexity as an evolving variable*, Oxford: Oxford University Press, 2009.

5. Y. M. Oh, F. Pellegrino, Towards robust complexity indices in linguistic typology: A corpus-based assessment, *Stud. Lang.*, 2022, 18–31. https://doi.org/10.1075/sl.22034.oh

6. D. Gil, *How complex are isolating languages*? In: M. Miestamo, K. Sinnemäki, F. Karlsson, editors, Language Complexity: Typology, Contact, Change, Amsterdam: John Benjamins, 2008, 109–131. https://doi.org/10.1075/slcs.94.08gil

7. S. Leufkens, Measuring redundancy: The relation between concord and complexity, *Linguist. Vanguard*, **9** (2023), 95–106. https://doi.org/10.1515/lingvan-2020-0143

8. J. E. Joseph, Why does language complexity resist measurement? *Front. Commun.*, **6** (2021). https://doi.org/10.3389/fcomm.2021.624855

9. I. Korzen, Are some languages more complex than others? On text complexity and how to measure it, *Globe J. Lang. Cult. Commun.*, **12** (2021), 18–31. https://doi.org/10.5278/ojs.globe.v12i.6665

10. J. Nichols, *Linguistic complexity: A comprehensive definition and survey*, In: G. Sampson, D. Gil, P. Trudgill, editors, Language Complexity as an Evolving Variable, Oxford: Oxford University Press, 2009, 110–125.

11. G. Deutscher, *Overall complexity: A wild goose chase*? In: G. Sampson, D. Gil, P. Trudgill, editors, Language Complexity as an Evolving Variable, Oxford: Oxford University Press, 2009, 243–251.

12. Ç. Çöltekin, T. Rama, What do complexity measures measure? Correlating and validating corpus-based measures of morphological complexity, *Linguist. Vanguard*, **9** (2023), 27–43. https://doi.org/10.1515/lingvan-2021-0007

13. E. A. Moravcsik, *Explaining language universals*, In: J. J. Song, editor, The Oxford Handbook of Linguistic Typology, Oxford: Oxford University Press, 2010, 69–89. https://doi.org/10.1093/oxfordhb/9780199281251.013.0005

14. J. H. Greenberg, *Universals of language*, Cambridge, MA: MIT Press, 1963.

15. G. Palloti, A simple view of linguistic complexity, *Second Lang. Res.*, **31** (2015), 117–134. https://doi.org/10.1177/0267658314536435

16. J. McWhorter, The world's simplest grammars are creole grammars, *Linguist. Typol.*, **6** (2001), 125–166. https://doi.org/10.1515/lity.2001.001

17. C. Bentz, X. Gutierrez-Vasques, O. Sozinova, T. Samardžić, Complexity trade-offs and equi-complexity in natural languages: A meta-analysis, *Linguist. Vanguard*, **9** (2023), 9–25. https://doi.org/10.1515/lingvan-2021-0054

18. O. Shcherbakova, V. Gast, D. Blasi, H. Skirgard, R. Gray, S. Greenhil, A quantitative global test of the complexity trade-off hypothesis: The case of nominal and verbal grammatical marking, *Linguist. Vanguard*, **9** (2023), 155–167. https://doi.org/10.1515/lingvan-2021-0011

19. R. Baechler, G. Seiler, *Complexity, isolation, and variation*, Berlin: De Gruyter, 2016. https://doi.org/10.1515/9783110348965

20. B. Baerman, D. Brown, G. G. Corbett, *Understanding and measuring morphological complexity*, Oxford: Oxford University Press, 2015. https://doi.org/10.1093/acprof:oso/9780198723769.001.0001

21. G. Coloma, *La Complejidad de los Idiomas*, Berlin: Peter Lang, 2017. https://doi.org/10.3726/b10613

22. C. C. Jiménez, *Complejidad lingüística: Orígenes y revisión crítica del concepto de lengua compleja*, Berlin: Peter Lang, 2018. https://doi.org/10.3726/b14515

23. E. Di Domenico, *Syntactic complexity from a language acquisition perspective*, Newcastle upon Tyne: Cambridge Scholars Publishing, 2017.

24. B. Kortmann, B. Szmrecsanyi, *Linguistic complexity: Second language acquisiton, indigenization, contact*, Berlin: Mouton de Gruyter, 2012. https://doi.org/ 10.1515/9783110229226

25. F. L. Mantia, I. Licata, P. Perconti, *Language in complexity: The emerging meaning*, Berlin: Springer, 2017. https://doi.org/10.1007/978-3-319-29483-4

26. J. McWhorter, *Linguistic simplicity and complexity: Why do languages undress*? Berlin: Mouton de Gruyter, 2012. https://doi.org/10.1515/9781934078402

27. F. J. Newmeyer, L. B. Preston, *Measuring grammatical complexity*, Oxford: Oxford Univesity Press, 2014. https://doi.org/10.1093/acprof:oso/9780199685301.001.0001

28. L. Ortega, Z. H. Han, *Complexity theory and language development*, Amsterdam: John Benjamins, 2017. https://doi.org/10.1075/lllt.48

29. M. Miestamo, *Grammatical complexity in a cross-linguistic perspective*, In: M. Miestamo, K. Sinnemäki, F. Karlsson, editors, Language Complexity: Typology, Contact, Change, Amsterdam: John Benjamins, 2008, 23–42. https://doi.org/10.1075/slcs.94.04mie

30. Ö. Dahl, *The growth and maintenance of linguistic complexity*, Amsterdam: John Benjamins, 2004. https://doi.org/10.1075/slcs.71

31. W. Kusters, *Linguistic complexity: The influence of social change on verbal inflection*, Utrecht: LOT, 2003.

32. P. Trudgill, Contact and simplification: Historical baggage and directionality in linguistic change, *Linguist. Typol.*, **5** (2001), 371–374.

33. J. A. Hawkins, *An efficiency theory of complexity and related phenomena*, In: G. Sampson, D. Gil, P. Trudgill, editors, Language Complexity Evolving Variation, Oxford: Oxford University Press, 2009, 252–268.

34. K. Ehret, *Kolmogorov complexity as a universal measure of language complexity*, In: Proceedings of the First Shared Task on Measuring Language Complexity, 2018, 8–14.

35. A. Andrason, Language complexity: An insight from complex-system theory, *Int. J. Lang. Linguist.*, **2** (2014), 74–89. https://doi.org/10.11648/J.IJLL.20140202.15

36. P. Blache, *A computational model for linguistic complexity*, In: G. Bel-Enguix, V. Dahl, M. D. Jiménez-López, editors, Biology, Computation and Linguistics, New Interdisciplinary Paradigms, Amsterdam: IOS Press, 2011, 155–167. https://doi.org/10.3233/978-1-60750-762-8-155

37. B. Bulté, A. Housen, *Defining and operationalising L2 complexity*, In: A. Housen, F. Kuiken, I. Vedder, editors, Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA, Amsterdam: John Benjamins, 2012, 21–46. https://doi.org/10.1075/lllt.32.02bul

38. F. Kuiken, Linguistic complexity in second language acquisition, *Linguist. Vanguard*, **9** (2023), 83–93. https://doi.org/10.1515/lingvan-2021-0112

39. M. Mohammadi, Complexity of language and SLA, *J. Soc. Sci. Human. Res.*, **8** (2020), 13–17. https://doi.org/10.24200/jsshr.vol8iss03pp13-17

40. A. Housen, B. De Clercq, F. Kuiken, I. Vedder, Multiple approaches to complexity in second language research, *Second Lang. Res.*, **35** (2019), 3–21. https://doi.org/10.1177/0267658318809765

41. A. Housen, H. Simoens, Cognitive perspectives on difficulty and complexity in L2 acquisition, *Stud. Second Lang. Acq.*, **38** (2016), 163–175. https://doi.org/10.1017/S0272263116000176

42. A. Housen, F. Kuiken, I. Vedder, *Complexity, accuracy and fluency*, In: A. Housen, F. Kuiken, I. Vedder, editors, Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA, Amsterdam: John Benjamins, 2012, 1–20. https://doi.org/10.1075/lllt.32.01hou

43. P. Ramat, *The (early) history of linguistic typology*, In: J. J. Song, editor, The Oxford Handbook of Linguistic Typology, Oxford: Oxford University Press, 2010, 9–24. https://doi.org/10.1093/oxfordhb/9780199281251.013.0002

44. C. Mauri, *Obiettivi, metodi e strumenti della tipologia*, In: N. Grandi, C. Mauri, editors, La tipologia linguistica: unità e diversità nelle lingue del mondo, Roma: Carocci Editore, 2022, 23–54.

45. H. O'Horan, Y. Berzak, I. Vulić, R. Reichart, A. Korhonen, *Survey on the use of typological information in natural language processing*, In: Y. Matsumoto, R. Prasad, editors, Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka: COLING, 2016, 1297–1308.

46. E. M. Ponti, H. O'Horan, Y. Berzak, I Vulić, R. Reichart, T. Poibeau, et al., Modeling language variation and universals: A survey on typological linguistics for natural language processing, *Comput. Linguist.*, **45** (2019), 1–156. https://doi.org/10.1162/coli_a_00357

47. N. Levshina, Corpus-based typology: Applications, challenges and some solutions, *Linguist. Typol.*, **26** (2022), 129–160. https://doi.org/10.1515/lingty-2020-0118

48. K. Gerdes, S. Kahane, X. Chen, Typometrics: From implicational to quantitative universals in word order typology, *Glossa*, **6** (2021), 1–6. https://doi.org/10.5334/gjgl.764

49. B. Bickel, *Absolute and statistical universals*, In: P. Colm Hogan, editor, The Cambridge Encyclopedia of the Language Sciences, Cambridge: Cambridge University Press, 2010, 77–79.

50. J. Nivre, M. C. Marneffe, F. Ginter, J. Hajič, C. D. Manning, S. Pyysalo, et al., *Universal dependencies*, 2023. Available from: `https://universaldependencies.org/`.

51. S. Petrov, D. Das, R. McDonald, *A universal part-of-speech tagset*, In: N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis, editors, Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul: European Language Resources Association, 2012, 2089–2096.

52. M. de Marneffe, T. Dozat, N. Silveira, K. Haverinen, F. Ginter, J. Nivre, et al., *Universal Stanford dependencies: A cross-linguistic typology*, In: N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis, editors, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC), Reykjavik: European Language Resources Association (ELRA), 2014, 4585–4592.

53. R. Futrell, R. P. Levy, E. Gibson, Dependency locality as an explanatory principle for word order, *Language*, **96** (2020), 371–412. https://doi.org/ 10.1353/lan.2020.0019

54. B. Guillaume, *Graph matching and graph rewriting: GREW tools for corpus exploration, maintenance and conversion*, In: D. Gkatzia, D. Seddah, editors, Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, 2021, 168–175. https://doi.org/10.18653/v1/2021.eacl-demos.21

55. V. Novák, Mining information from time series in the form of sentences of natural language, *Int. J. Approx. Reason.*, **78** (2016), 192–209. https://doi.org/10.1016/j.ijar.2016.07.006

56. V. Novák, *The concept of linguistic variable revisited*, In: M. Sugeno, J. Kacprzyk, S. Shabazova, editors, Recent Developments in Fuzzy Logic and Fuzzy Sets, Studies in Fuzziness and Soft Computing, Berlin/Heidelberg, Germany: Springer, 2020, 105–118.

57. V. Novák, *Fuzzy logic in natural language processing*, In: Proceedings of the 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Naples, Italy: IEEE, 2017. https://doi.org/10.1109/FUZZ-IEEE.2017.8015405

58. V. Novák, *Mathematical fuzzy logic: From vagueness to commonsese reasoning*, In: G. Kreuzbauer, N. Gratzl, E. Hielb, editors, Retorische Wissenschaft: Rede und Argumentation in Theorie und Praxis, Wien, Austria: LIT-Verlag, 2008, 191–223.

59. V. Novák, *What is fuzzy natural logic*, In: Integrated Uncertainty in Knowledge Modelling and Decision Making, V. Huynh, M. Inuiguchi, T. Denoeux, editors, Berlin/Heidelberg, Germany: Springer, 2015, 15–18.

60. V. Novák, Fuzzy natural logic: *Towards mathematical logic of human reasoning*, In: R. Seising, E. Trillas, J. Kacprzyk, editors, Fuzzy Logic: Towards the Future, Berlin/Heidelberg, Germany: Springer, 2015, 137–165.

61. V. Novák, Evaluative linguistic expressions vs. fuzzy categories? *Fuzzy Set. Syst.*, **281** (2015), 81–87.

62. A. Torrens-Urrutia, V. Novák, M. D. Jiménez-López, Describing linguistic vagueness of evaluative expressions using fuzzy natural logic and linguistic constraints, *Mathematics*, **10** (2022), 2760. https://doi.org/10.3390/math10152760

63. A. Torrens-Urrutia, M. D. Jiménez-López, S. Campillo-Muñoz, Dealing with evaluative expressions and hate speech metaphors with Fuzzy Property Grammar Systems, *Axioms*, **12** (2023), 484. https://doi.org/10.3390/axioms12050484

64. A. Torrens-Urrutia, V. Novák, M. D. Jiménez-López, Fuzzy property grammars for gradience in natural language, *Mathematics*, **11** (2023), 735. https://doi.org/10.3390/math11030735

65. A. Torrens-Urrutia, M. D. Jiménez-López, A. Brosa-Rodríguez, D. Adamczyk, A fuzzy grammar for evaluating universality and complexity in natural language, *Mathematics*, **10** (2023), 602. https://doi.org/10.3390/math10152602

66. A. Torrens-Urrutia, M. D. Jiménez-López, A. Brosa-Rodríguez, A fuzzy approach to language universals for NLP, In: Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Luxembourg: IEEE, 2021, 1–6. https://doi.org/10.1109/FUZZ45933.2021.9494516

67. M. Daniel, *Linguistic typology and the study of language*, In: J. J. Song, editor, The Oxford Handbook of Linguistic Typology, Oxford: Oxford University Press, 2010, 43–68. https://doi.org/10.1093/oxfordhb/9780199281251.013.0004

68. H. Hammarström, Counting languages in dialect continua using the criterion of mutual intelligibility, *J. Quant. Linguist.*, **15** (2008), 34–45. https://doi.org/10.1080/09296170701794278

69. M. Cysouw, Using the world atlas of language structures, *Lang. Typol. Univ.*, **61** (2009), 1–6. https://doi.org/10.1524/stuf.2008.0018

70. D. Bakker, *Language sampling*, In: J. J. Song, editor, The Oxford Handbook of Linguistic Typology, Oxford: Oxford University Press, 2010, 1–26. https://doi.org/10.1093/oxfordhb/9780199281251.013.0007

71. M. Miestamo, D. Bakker, A. Arppe, Sampling for variety, *Linguist. Typol.*, **20** (2016), 233–296. https://doi.org/10.1515/lingty-2016-0006

72. M. G. Naranjo, L. Becker, Statistical bias control in typology, *Linguist. Typol.*, **26** (2022), 605–670. https://doi.org/10.1515/lingty-2021-0002

73. A. Brosa-Rodríguez, M. D. Jiménez-López, *A typometrical study of Greenberg's linguistic universal 1*, In: R. Mehmood, et al., editors, Distributed Computing and Artificial Intelligence, Lecture Notes in Networks and Systems, Berlin: Springer, **741** (2023), 186–196. https://doi.org/10.1007/978-3-031-38318-2_19

74. K. Gerdes, S. Kahane, X. Chen, *Rediscovering Greenberg's word order universals in UD*, In: A. Rademaker, F. Tyers (Editors), editors, Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019), Paris: Association for Computational Linguistics, 2019, 124–131. https://doi.org/10.18653/v1/W19-8015

75. K. Sinnemäki, *Language universals and linguistic complexity: Three case studies in core argument marking*, Unpublished PhD dissertation, Helskinki: University of Helsinki, 2011.

76. M. Miestamo, K. Sinnemäki, F. Karlsson, *Language complexity: Typology, contact, change,* Amsterdam: John Benjamins, 2008.

77. J. F. Newmeyer, *More complicated and hence, rarer: A look at grammatical complexity and crosslinguistic rarity*, In: V. S. Karimi, V. Samiian, W. K. Wilkins, editors, Phrasal and clausal architecture:Syntactic derivation and interpretation, Berlin: Mouton de Gruyter, 2007, 221–242.

78. A. C. Harris, *On the explanation of typologically unusual structures*, In: J. Good, editor, Linguistic universals and language change, Oxford: Oxford University Press, 2008, 54–76.

79. B. Edmonds, *Syntactic measures of complexity*, Unpublished PhD dissertation, Manchester: University of Manchester, 1999.

80. A. Torrens-Urrutia, M. D. Jiménez-López, A. Brosa-Rodríguez, D. Adamczyk, A fuzzy grammar for evaluating universality and complexity in natural language, *Mathematics*, **10** (2022), 2602. https://doi.org/10.3390/math10152602
81. R. Tomlin, *Basic word order: Functional principles*, London: Croom Helm, 1986.
82. M. Dryer, *Why statistical universals are better than absolute universals*, In: Papers from the 33rd Annual Meeting of the Chicago Linguistics Society, 1998, 1–23.
83. M. Dryer, On the order of demonstrative, numeral, adjective, and noun, *Language*, **94** (2018), 798–833. https://doi.org/10.1353/lan.2018.0054
84. W. Croft, *Typology and universals*, Cambridge: Cambridge University Press, 2003.
85. M. Dryer, M. Haspelmath, *The world atlas of language structures online*, WALS Online (v2020.3), Data set, Zenodo, 2023. https://doi.org/10.5281/zenodo.7385533
86. H. S. Choi, B. Guillaume, K. Fort, Corpus-based language universals analysis using universal dependencies, *ACL Anthology*, 2021, 1–15.
87. H. S. Choi, B. Guillaume, K. Fort, Investigating dominant word order on universal dependencies with graph rewriting, *Int. Conf. Recent Adv. Nat. Lang. Proc.*, 2021, 281–290.