mathematics

MDPI

Article

A Fuzzy Grammar for Evaluating Universality and Complexity in Natural Language

Adrià Torrens-Urrutia [1,*,†], María Dolores Jiménez-López [1,†], Antoni Brosa-Rodríguez [1,†] and David Adamczyk [2,†]

[1] Research Group on Mathematical Linguistics (GRLMC), Universitat Rovira i Virgili, 43002 Tarragona, Spain; mariadolores.jimenez@urv.cat (M.D.J.-L.); antoni.brosa@urv.cat (A.B.-R.)
[2] Institute for Research and Applications of Fuzzy Modeling (IRAFM), University of Ostrava, 701 03 Ostrava, Czech Republic; david.adamczyk@osu.cz
* Correspondence: adria.torrens@urv.cat
† These authors contributed equally to this work.

Abstract: The paper focuses on linguistic complexity and language universals, which are two important and controversial issues in language research. A Fuzzy Property Grammar for determining the degree of universality and complexity of a natural language is introduced. In this task, the Fuzzy Property Grammar operated only with syntactic constraints. Fuzzy Natural Logic sets the fundamentals to express the notions of universality and complexity as evaluative expressions. The Fuzzy Property Grammar computes the constraints in terms of weights of universality and calculates relative complexity. We present a proof-of-concept in which we have generated a grammar with 42B syntactic constraints. The model classifies constraints in terms of low, medium, and high universality and complexity. Degrees of relative complexity in terms of similarity from a correlation matrix have been obtained. The results show that the architecture of a Universal Fuzzy Property Grammar is flexible, reusable, and re-trainable, and it can easily take into account new sets of languages, perfecting the degree of universality and complexity of the linguistic constraints as well as the degree of complexity between languages.

Keywords: linguistic universals; linguistic complexity; evaluative expressions; fuzzy grammar; linguistic gradience; linguistic constraints

MSC: 03B65

Citation: Torrens-Urrutia, A.; Jiménez-López, M.D.; Brosa-Rodríguez, A.; Adamczyk, D. A Fuzzy Grammar for Evaluating Universality and Complexity in Natural Language. *Mathematics* **2022**, *10*, 2602. https://doi.org/10.3390/math10152602

Academic Editor: Michael Voskoglou

Received: 30 June 2022
Accepted: 19 July 2022
Published: 26 July 2022

1. Introduction

In this paper, we propose a formal grammar for approaching the study of universality and complexity in natural languages. With this model, we afford from a mathematical point of view two key issues in theoretical linguistics. There has been a long tradition of using mathematics as a modeling tool in linguistics [1]. By formalization, we mean "the use of appropriate tools from mathematics and logic to enhance explicitness of theories" [2]. We can claim that "any theoretical framework stands to benefit from having its content formalized" [2], and complexity and language universals are not an exception.

Linguistic complexity and language universals are two important and controversial issues in language research. Complexity in language is considered a multifaceted and multidimensional research area, and for many linguists, it "is one of the currently most hotly debated notions in linguistics" [3]. On the other hand, theoretically, linguistic universals have been the subject of intense controversy throughout the history of linguistics. Their nature and existence have been questioned, and their analysis has been approached from many different perspectives.

Regarding linguistic complexity, it has been defended for a long time. The so-called dogma of equicomplexity defends that linguistic complexity is invariant, that languages are not measurable in terms of complexity and that there is no sense in trying to show that there

are languages more complex than others. Given this dogma, some questions that come up are the following: if the equicomplexity axiom supports the idea that languages can differ in the complexity of subsystems, why is the global complexity of any language always identical? What mechanism slows down complexity in one domain when complexity increases in another domain? What is the factor responsable for the equi-complexity?

There has been a recent change in linguistics with regard to studies on linguistic complexity that is considerable. It has gone from denying the possibility of calculating complexity—a position advocated by most linguists during the 20th century—to a great interest in studies on linguistic complexity since 2001 [4]. During the 20th century, the dogma of equicomplexity prevailed. Faced with this position, at the beginning of the 21st century, a large group of researchers argued that it is difficult to accept that all languages are equal in their total complexity and that the complexity in one area of the language is compensated by simplicity in another. Therefore, equicomplexity is questioned, and there are monographs, articles and conferences that, in one way or another, are concerned with measuring the complexity of languages.

In fact, the number of papers published in recent years on complexity both in the field of theoretical and applied linguistics [3,5–13] highlights the interest in finding a method to calculate linguistic complexity and in trying to answer the question of whether all languages are equal in terms of complexity or, if on the contrary, they differ in their levels of complexity.

Despite the interest in studies on linguistic complexity in recent years and although, in general, it seems clear that languages exhibit different levels of complexity, it is not easy to calculate exactly these differences. Part of this difficulty may be due to the different ways of understanding the concept of complexity in the study of natural languages.

Different types of complexity can be distinguished. Pallotti [14] classifies the different meanings of the term into three types:

- *Structural complexity*, if we calculate complexity in terms of a formal property of texts related to the number of rules or patterns.
- *Cognitive complexity* is the type of complexity that calculates the cost of processing.
- *Complexity of development*. In this case, we are talking about the order in which linguistic structures emerge and are mastered in second (and possibly first) language acquisition.

These types of complexity identified by Palloti are captured by the two main types of complexity in the literature: *absolute complexity*, an objective property of the system measured in terms of the number of parts of the system, the number of interrelationships among parts, or the length of a phenomenon description [15]; and *relative complexity*, which considers language users and is related to the difficulty or cost of processing, learning or acquisition. Other common dichotomies in the literature are those that distinguish *global complexity* from *local complexity* [16] or those that establish a difference between *system complexity* and *structural complexity* [15].

To measure complexity, studies in the field propose ad hoc measures of complexity that depend on the specific interests of the analysis carried out. The proposed measures are very varied, and the formalisms used can be grouped into two types: (1) measures of absolute complexity (number of categories, number of rules, ambiguity, redundancy, etc. [16]); and (2) measures of relative complexity that face the problem of determining what type of task (learning, acquisition, processing) and what type of agent (speaker, listener, child, adult) to consider. Second language learning complexity (L2) in adults [17,18] or processing complexity [19] are examples of measures that have been proposed in terms of difficulty/cost. In many cases, other disciplines have been turned to in search of tools to calculate the complexity of languages. Information theory, with formalisms such as Shannon entropy or Kolmogorov complexity [15,16], and complex systems theory [20] are some examples of areas that have provided measures for a quantitative evaluation of linguistic complexity.

Most of the studies carried out on the complexity of natural languages adopt an absolute perspective of the concept, and there are a few that address the complexity

from the user's point of view. This situation may be due to the fact that in general, it is considered that the analysis of absolute complexity presents fewer problems than that of relative complexity, since its study does not depend on any particular group of language users [16]. Relative complexity approaches compel researchers to face many problems:

- What do we mean by complex? More difficult, more costly, more problematic, more challenging?
- Different ways of using language (speaking, hearing, language acquisition, second language learning) may differ in classifying something as difficult or easy.
- When we determine that a phenomenon is complex, we must indicate for whom it is complex, since some phenomena can be very complex for one group and instead facilitate the linguistic task for other groups.
- An approach to the concept of complexity based on use requires focusing on a specific user of the language and determining the "ideal" user. How do we decide which is the main use and user of the language?

Although, as we have said, most of the works carried out adopt an absolute perspective of the concept, many specialists are interested in analyzing the relative complexity. From a relative point of view, there are three different questions that could be answered:

- From the point of view of second language learning, is it more difficult for an adult to learn some languages than to learn others?
- If we consider the processing of a language, is it more difficult to speak some languages than others?
- Focusing on the language acquisition process, does it take longer to acquire some languages than others?

Therefore, one of the possible perspectives in studies on relative complexity is one that understands complexity in terms of "learning difficulty". A relative perspective is adopted here that forces us to take the language user into account: the adult learning a language. Trudgill [17], for example, argues that "linguistic complexity equates with difficulty of learning for adults" and Kusters [18] defines complexity as "the amount of effort an outsider has to make to become acquainted with the language in question [...]. An outsider is someone who learns the language in question at a later age, and is not a native speaker". The problem that we find in these studies on relative complexity is the large number of definitions and different measures used that make the results obtained often inconsistent and not comparable. On the other hand, most complexity studies that focus on the learning process pay attention almost exclusively to the target language and the success rate of learners. In general, they do not consider the weight that the learner's mother tongue has in calculating the complexity of L2. They thus consider a kind of "ideal learner" as the basis of their analyses and focus on the complexity of the different subdomains of the target language.

In the model that we present here, we consider that in order to calculate the relative complexity of languages in terms of L2 learning, it is necessary to consider the mother tongue of the learners when calculating the relative complexity of the target language, since it seems clear that the mother tongue can facilitate or complicate the process of learning the target language and, therefore, can condition the assessment of linguistic complexity.

Regarding language universals, we can define a universal of language as a grammatical characteristic present in all or most human languages [21]. Although linguists have always been interested in discovering characteristics shared by languages, it was not until Greenberg's contribution [22], with universals based on a representative set of 30 languages, that the research topic gained popularity and depth. A decade later, despite the interest aroused by Greenberg's findings, the impossibility of improving on these results caused the study of universals to lose interest and usefulness. This object of study became relevant again a decade later, thanks to the innovations in sampling proposed by linguistic typology and authors such as Comrie [23] or Dryer [24,25]. However, this expansion of data and

sampling techniques aggravates the congenital problem of linguistic universals: more and more exceptions to them appear and, therefore, the term is less reliable or representative.

In recent years, although the problem presented above has not been solved, the boost in Natural Language Processing has rescued linguistic universals from oblivion. There is a clear symbiosis between the two fields, since NLP offers many tools, resources and techniques that improve the study of universals and, above all, make it more efficient [26,27]. In turn, a true understanding of the features shared by all languages implies that recent advances in NLP, applicable only in English and a few other languages, can be more easily extended to low-resource languages.

Language universals have been investigated from two different perspectives in linguistics: on the one hand, the typological, functional or Greenbergian approach; and on the other hand, the formal or Chomskyan approach [28]. From the typological point of view, taking into account the limited data available, the universals are derived inductively from a cross-linguistic sample of grammatical structures [29]. In contrast, in the formal approach, universals are derived deductively, taking into account assumptions about innate linguistic capacity and using grammatical patterns in languages (Universal Grammar) [30].

Linguistic universals have been classified taking into account the modality and domain [21]. If we consider the *modality*, we can distinguish the following types of universals:

- *Absolute universals*. These universals are those that do not present exception and that, therefore, are fulfilled in all members of the universe. Absolute universals defend the hypothesis that a grammatical property be present in a language.
- *Probabilistic or statistical universals*. These types of universals are valid for most languages, but not for all. Probabilistic universals defend the hypothesis that a grammatical property can be present in languages with a certain degree of probability.

To the typology proposed by Moravcsik [21], we can add another common concept in the literature on universals: the concept of *rara* or *rarissima* [31,32]. In this case, we are talking about a linguistic feature that is completely opposite to the one that is considered universal. We are referring to those characteristics that are not common in languages.

Taking into account the *domain*, linguistic universals can be divided into two main types:

- *Unrestricted universals*. These type of universals may be stated for the whole universe of languages. These universals are applicable to any human language.
- *Restricted, implicational or typogical universals*. These universals affect only a part of the world's languages: those that share a given characteristic previously (if x, then y).

The above four types of universals can be schematized as follows [21]:

- Unrestricted and absolute: In all languages, Y.
- Unrestricted and probabilistic: In most languages, Y.
- Restricted and absolute: In all languages, if there is X, there is also Y.
- Restricted and probabilistic: In most languages, if there is X, there is also Y.

As Moravcsik [21] states, taking into account that it is not possible to analyze every natural language, all language universals are nothing more than mere hypotheses. As a consequence, "The empirical basis of universals research can only be (a sample of) a subset of the domain for which universals could maximally claim validity, and have traditionally been claiming validity: that of humanly possible languages. Therefore, the only viable domain for universals research, then, is all-languages-present-and-past-as-known-to-us-now" [33].

What we have said reveals both the significance of complexity and universals in language studies and the difficulties to deal with these notions. In this paper, we aim to contribute to the field by proposing a fuzzy grammar for determining the degree of universality and complexity of a natural language. By considering the degree of universality, the model calculates the relative complexity of a language. In fact, in our proposal, an inversely proportional relation between universality and complexity is established: the more universal a language is, the less complex it is. With our model, we can calculate the

degree of complexity by checking the number of universal rules this language contains. The idea at the base of our model is that those languages that have high universality values will be more similar to each other, and therefore, their level of relative complexity will be lower. On the contrary, those languages with low levels of universality will have a high number of specific rules, and this will increase their level of relative complexity.

The paper is organized as follows. In Section 2, we present the models of Fuzzy Universal Property Grammar and Fuzzy Natural Logic as a strategy to define linguistic universality and language complexity as vague concepts. In Section 3, material and methods are described. In Section 4, we provide a description of the experimental results. Finally, in Section 6, we discuss the results and highlight future research directions.

2. Related Work

Both Fuzzy Property Grammars (FPGr) and Fuzzy Natural Logic (FNL) can provide strategies to approach linguistic universals and the measurement of language complexity.

Regarding linguistic universality, we consider that the concept of universal can be better defined considering a continuous scale than adopting a discrete perspective. We disregard the fact that only those linguistic rules shared by all known languages—roughly 7000—can be regarded as linguistic universals. On the other hand, we define "universality" as a continuum using the FNL's gradient features of the theory of evaluative expressions. As a result, we shall continue to respect the two extreme points that already exist: 0 (non-universal) and 1 (full-universal). However, we create a spectrum in which we shall fit those linguistic rules known as "*quasi-universals*" between these two positions. On the other hand, FPGr and FNL make it possible to devise universal models, that is language-independent models, that can be applied to all natural languages, and they use a fuzzy-gradient technique to describe linguistic universals in fuzzy terms [0, 1] rather than labeling universals with a confusing nomenclature. The times that a fuzzy universal is fulfilled or violated in a fuzzy grammar determines the fuzzy degree membership of a linguistic rule. To create a Fuzzy Universal Grammar, we will use the model of FPGr. Finally, FPGr is a cheap and reusable architecture to define linguistic phenomena and their variation and makes it possible to advance in the systematization of variation in languages.

Regarding the complexity of language, it can be captured in quantitative terms (absolute complexity), such as the more rules a grammar has, the more complex it is. Therefore, if we have a system that provides all the rules of a language, we could capture its degree of complexity under the architecture of FNL of the theory of evaluative expressions. It is also possible to measure complexity between languages with a FPGr. The more rules are shared between two languages, the less complexity will be found between those two languages. The fewer rules those languages share, the more complex they will be in relation to each other. However, this approach will have a higher cost, since it will demand checking how many rules of a targeted language are shared with respect to all the other languages. That is why we have disregarded such an approach, and we have implemented a Fuzzy Universal Property Grammar that will consider all the possible combinations of rules. Therefore, for every single set of languages, we will only need to check coincidences in our Fuzzy Universal Property Grammar. Thus, a value in terms of degree [0–1] will arise from the number of coincidences defining universality: as in what is the membership degree of a set of rules of a language with respect to a Fuzzy Universal Property Grammar. In this way, the notion of universality can help measure the relative complexity of a language, assuming that those languages that have a lot of specific rules are meant to be more complex. That is, the more universal a language is, the less complex it is; the less universal a language is, the more complex it is.

In the following, we present the models of Fuzzy Universal Property Grammar and Fuzzy Natural Logic as a strategy to define linguistic universality and language complexity as vague concepts.

2.1. Fuzzy Property Grammars for Linguistic Universality

Fuzzy Property Grammars (FPGr) [34–36] combine the formalism of Fuzzy Natural Logic [37–43] and linguistic constraints typically used in linguistics. The higher-order fuzzy logic as a formalism describes the grammar at a higher level (abstractly), enabling a mathematical formalization of the degrees of grammaticality. In comparison, linguistic constraints allow us to describe vague phenomena on a local-sentence level, characterizing the objects (constraints) as prototypical and borderline ones. Therefore, both assets are necessary to build an FPGr. There are three key concepts of an FPGr: linguistic constraint, universe of the linguistic domains and fuzzy grammar, and linguistic construction.

2.1.1. Linguistic Constraint

A linguistic **constraint** is a **relation that puts together two or more linguistic elements** such as *linguistic categories* or *parts-of-speech*. Formally, a linguistic constraint is an n-tuple $\langle A_1, ..., A_n \rangle$ where A_i are linguistic categories. We usually have $n = 2$. For example, the following linguistic categories can be distinguished for this work:

1. DET (determiner);
2. ADJ (adjectve);
3. $NOUN$ (noun);
4. $PROPN$ (proper noun);
5. $VERB$ (verb);
6. ADV (adverb);
7. $CONJ$ (conjunction);
8. $SCONJ$ (subordinate conjunction);
9. ADP (preposition).

There are four types of constraints in the Fuzzy Property Grammars (FPGr):

1. **General or universal constraints** that are valid for a universal grammar. They are built from all the possible combinations between linguistic objects and constraints.
2. **Specific constraints** that are applicable to a specific grammar.
3. **Prototypical constraints** that definitely belong to a specific grammar, i.e., their degree of membership is 1.
4. **Borderline constraints** that belong to a specific language with some degree only (we usually measure it by a number from (0, 1)).

The constraints from FPGr that we will work with to describe linguistic universality and complexity are the following (the A and B are understood as linguistic categories):

- *Linearity* of precedence order between two elements: A precedes B, in symbols $A \prec B$. For example, $DET \prec NOUN$ in "*The girl*".
- *Co-occurrence* between two elements: A requires B, in symbols $A \Rightarrow B$. For example, $ADJ \Rightarrow NOUN$ in "*the red car*".
- *Exclusion* between two elements: A and B never appear in co-occurrence in the specified construction, in symbols $A \otimes B$. That is, only A or only B occurs. For example, $PRON \otimes NOUN$ in "*He runs*".
- *Uniqueness* means that neither a category nor a group of categories (constituents) can appear more than once in a given construction. For example, there is only one $PRON$ in "*She eats pizza*".
- *Dependency*. An element A has a dependency$_i$ on an element B in symbols $A \rightsquigarrow_i B$. Typical dependencies (but not exclusively) for $_i$ are *subj* (subject), *mod* (modifier), *obj* (object), *spec* (specifier), *verb* (verb), and *conj* (conjunction).

2.1.2. Definition of a Fuzzy Property Grammar

Definition 1. *A Fuzzy Property Grammar (FPGr) is a couple*

$$FPGr = \langle U, FGr \rangle \tag{1}$$

where U is a universe

$$U = Ph_\rho \times Mr_\mu \times X_\chi \times S_\delta \times L_\theta \times Pr_\zeta \times Ps_\kappa. \quad (2)$$

The subscripts $\rho, \ldots, \kappa$ denote types, and the sets in Equation (2) are sets of the following constraints:

- $Ph_\rho = \{ph_\rho \mid ph_\rho$ *is a phonological constraint*$\}$ *is the set of constraints that can be determined in phonology.*
- $Mr_\mu = \{mr_\mu \mid mr_\mu$ *is a morphological constraint*$\}$ *is the set of constraints that can be determined in morphology.*
- $X_\chi = \{x_\chi \mid x_\chi$ *is a syntactic constraint*$\}$ *is the set of constraints that characterize syntax.*
- $S_\delta = \{s_\delta \mid s_\delta$ *is a semantic constraint*$\}$ *is the set of constraints that characterize semantic phenomena.*
- $L_\theta = \{l_\theta \mid l_\theta$ *is a lexical constraint*$\}$ *is the set of constraints that occur on a lexical level.*
- $Pr_\zeta = \{pr_\zeta \mid pr_\zeta$ *is a pragmatic constraint*$\}$ *is the set of constraints that characterize pragmatics.*
- $Ps_\kappa = \{ps_\kappa \mid ps_\kappa$ *is a prosodic constraint*$\}$ *is the set of constraints that can be determined in prosody.*

The second component is a function:

$$FGr : U \to [0, 1] \quad (3)$$

which can be obtained as a composition of functions $F_\rho : Ph_\rho \to [0, 1], \ldots, F_\kappa : Ps_\kappa \to [0, 1]$. Each of the latter functions characterizes the degree in which the corresponding element x belongs to each of the above linguistic domains (with respect to a specific grammar).

Technically speaking, *FGr* in Equation (3) is a fuzzy set with the membership function computed as follows:

$$FGr(\langle x_\rho, x_\mu, \ldots, x_\kappa \rangle) = \min\{F_\rho(x_\rho), F_\mu, \ldots, F_\kappa(x_\kappa)\} \quad (4)$$

where $\langle x_\rho, x_\mu, \ldots, x_\kappa \rangle \in U$.

Let us now consider a set of constraints from an external linguistic input $D = \{d \mid d$ is a dialect constraint$\}$. Each $d \in D$ can be seen as an n-tuple $d = \langle d_\rho, d_\mu, \ldots, d_\kappa \rangle$. Then, the membership degree $FGr(d) \in [0, 1]$ is a degree of grammaticality of the given utterance that can be said in arbitrary dialect (of the given grammar).

2.2. Fuzzy Property Grammars for Linguistic Universality

To take into account linguistic universality, we have to point out the following considerations to the previous definitions.

We constraint the universe of our FPGr to only the syntactic domain X_χ. At this point, it is only possible to generate all the possible constraints for the syntactic domain. However, we assume that this formulation is a proof of concept for future work on the rest of the domains.

Therefore, *U-FPGr* will be understood exactly as shown in Equation (2).

Definition 2. *A Universal Fuzzy Property Grammar (U-FPGr) is a couple*

$$U\text{-}FPGr = \langle U, FGr \rangle \quad (5)$$

However, its (linguistic) universe in written language stands for a simplified version of an *FPGr* because only the syntactical domains (x) are relevant for the proof of concept that we are presenting in this work: $< x >$, the others are neglected: $FPGr =< \overline{U}, \overline{FGr} >$.

Definition 3. *U-FPGr is*

$$\overline{U} =< x > \tag{6}$$

In this case, $\overline{U}$ is generated as a Cartesian product of all the possible constraints.

$$U = Pos_\alpha \times Dep_\beta \times X_\chi \times Pos_\alpha \times Dep_\beta \times Pos_\alpha \times Dep_\beta. \tag{7}$$

The subscripts $\alpha, \ldots, \gamma$ denote types, and the sets in Equation (7) are sets of the following constraints:

- $Pos_\alpha = \{pos_\alpha \mid pos_\alpha$ is a linguistic category in terms of part-of-speech$\}$ is the set of linguistic categories that can be determined in all languages.
- $Dep_\beta = \{dep_\beta \mid dep_\beta$ is a linguistic dependency$\}$ is the set of dependencies that can be determined in all languages.
- $X_\chi = \{x_\chi \mid x_\chi$ is a syntactic constraint$\}$ is the set of constraints that characterize syntax.

From the linguistic point of view, each combination of $U = Pos_\alpha \times Dep_\beta$ is interpreted as a linguistic element such as a noun with subject dependency $NOUN_{[nsubj]}$, a determiner with a determiner dependency $DET_{[det]}$, or a verb as the root of the sentence $VERB_{[root]}$. Therefore, by repeating this three times, we assume that all the rules follow a linguistic constituent, such as a linguistic element (category and dependency) in relation in terms of syntactic linguistic constraints with another element (category and dependency) and a third element (category and dependency). Because of the fact that some constraints do not need this third element, we will include in our universe the possibility of having a rule without the third element.

Any language that will be computed in terms of linguistic universality will need to follow this formalism to describe its universe. The targeted language will be our linguistic input $L = \{l \mid l$ is a language constraint$\}$. Each $l \in L$ can be seen as an n-tuple $l = \langle l_\alpha, l_\beta, \ldots, l_\chi \rangle$. Then, the membership degree $FGr(l) \in [0, 1]$ is a degree of universality given a language as a set. As seen, this is just an adaptation of how FPGr treats grammaticality. Therefore, the universality of a targeted language is computed in terms of being grammaticality understood as the membership degree of a targeted language set with respect to *U-FPGr* . Therefore, our gradient model suggests the convenience of a terminological change. We consider that it is not necessary to define our proposal as a "search for universals" task. However, on the contrary, what we intend is to search for or define a "spectrum of the universal", or what is the same, any linguistic rule that can fit to a membership degree of universality in terms of [0, 1].

Additionally, we have implemented an $IF - THEN$ rule to assign a weight value to each rule of the *U-FPGr* .

- If a rule in L coincided with a rule in the *U-FPGr* , then add +1;
- The more weight a rule has, the more universal it is in a representative set;
- The less weight a rule has, the less universal it is in a representative set.

This is quite a natural way of representing universality, since our knowledge of the universals is dependent on the system of language that we know. A rule that might be considered as a universal can become a *quasi-universal* in the moment that new languages are discovered, and such languages do not consider such a rule. Therefore, we are always computing universality in terms of a finite representative set out of the infinite sets of languages. In this case, *U-FPGr* is flexible and re-usable, since it can update the weight of universality according any new language inserted as a linguistic input.

2.3. Fuzzy Natural Logic Computing Universals and Linguistic Complexity with Words

In order to better grasp gradient terminology as it relates to linguistic universals and complexity, we propose to compute the continuum with natural language words. For this, the concepts of *universality* and *complexity* are assumed.

Fuzzy natural logic 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 [37,38,40–44] such as *length*, *age*, *depth*, *thickness*, *beauty*, and *kindness*, among others. In this case, we will take into account "*universality*" and "*complexity*" as evaluative expressions.

FNL assumes that the simple evaluative linguistic expression has the general form:

$$\langle intensifier \rangle \langle TE\text{-}head \rangle \tag{8}$$

$\langle TE\text{-}head \rangle$ can be grouped together to form a *fundamental evaluative trichotomy* consisting of two antonyms and a middle term, for example $\langle good, normal, bad \rangle$. For our work, we will take into account the trichotomy of $\langle low, medium, high \rangle$. In this sense, as proposed in [45], the membership scale of universality in linguistic rules recognize:

- *High Satisfied Universal*. Linguistic rules that trigger a *high* truth value of satisfaction in *U-FPGr*, therefore, they are found satisfied in quasi-all languages. This fuzzy set includes those rules known as *Full Universals*, absolute rules, which are located in (almost) all languages.
- *Medium Satisfied Universal*. Linguistic rules that trigger a *medium* truth value of satisfaction in *U-FPGr*; therefore, they are found satisfied in the overall average of languages.
- *Low Satisfied Universal*. Linguistic rules that trigger a *low* truth value of satisfaction in *U-FPGr*; therefore, they are found satisfied in almost none of the languages.

The value of complexity is obtained from $IF - THEN$ rules such as:

Definition 4. *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 [45]:

- *Low Complexity*. Linguistic rules that have a *high* truth value in terms of weight in *U-FPGr* . They are found satisfied in quasi-all languages. This fuzzy set includes rules known as *full universals*, absolute rules, which are located in (almost) all languages.
- *Medium Complexity*. Linguistic rules that have a *medium* truth value in terms of weight in *U-FPGr*: rules found in the overall average of languages.
- *High Complexity*. Lnguistic rules that have a *low* truth value in terms of weight in *U-FPGr*: rules satisfied in almost none of the languages.

A *possible world* is defined as a specific context in which a linguistic expression is used. In 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 be simply the membership degree [0–1], while our extension will be dependent on the number of languages we are taking into account in a representative set for evaluating universality and complexity.

Figure 1 represents how Fuzzy Natural Logic accounts for the fuzzy-gradient notion of universality in fuzzy sets. The fuzzy limits between sets must be established. In terms of mathematical fairness rather than from a cognitive perspective, the possible world of 7000 languages has been divided into three parts for each fuzzy set. Therefore,

roughly, each set is computed by 2333 language grammars. The proposed cut-off could be changed. However, we consider that there would not be a big change between the perceived perspective of the fuzzy transitions and the three-cut part criteria. We claim that the concept of universality would be better captured with a trichotomical expression of $\langle small - medium - big \rangle$ in terms of $\langle low - medium - high \rangle$. This new way of accounting for universals may have advantages over the classical nomenclature found in the literature [29,46,47] (*universal trend*, *statistical universal*, *rara*, *rarisima*, *typological generalization*, etc.).

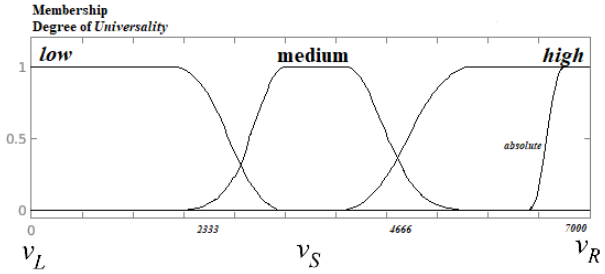

Figure 1. Linguistic Universality as a Evaluative Expression.

The advantages of the proposed model can be summarized as follows:

- The model presents a consistent classification without contradictions in terms of degree for the concepts of universality and complexity.
- The model provides a characterization of the vagueness of linguistic universality. This characterization fits the data surveys (atlas) that quatitatively collect linguistic universals terms such as WALS [48] and the Universals Archive [49].

The proposed model aims to collect the work already done in linguistics and present a universal characterization for the description of fuzzy linguistic universals and linguistic complexity.

3. Materials and Methods

3.1. A Fuzzy Universal Grammar with a Representative Set

From the 7000 languages in the world (an oscillating and debatable number), there are still a large number of them without adequate documentation. Therefore, when one wants to predict possible trends in the set of human languages as a whole, one has to investigate a selection of languages, hoping that the results will be extensible to the rest. This extension of languages is what we will consider in FNL our extension regarding the possible world of our evaluative expression of the notions of linguistic *"universality"* and *"complexity"*. To this end, creating a representative and balanced set is essential. However, this task is by no means easy, as there are many other limitations [29,50,51].

For this reason, linguistic typology has classically proposed different ways of configuring a set that is as varied and independent as possible in order to be as close as possible to the reality of the 7000 languages. The selection of this independence between languages can be based on different criteria: typological, genetic, areal or a combination of them. However, it is still very difficult to find perfect samples due to what is known as bibliographical bias: the data available to us are very limited.

The representative set is build under the data from linguistic corpus. Such data allow us to create sets of languages. Working with a linguistic corpus helps us to obtain a deeper and more quantitative knowledge of cross-linguistic tendencies [52,53]. The problem with this methodology is that the available data are still very limited given its novelty and level of depth, especially in comparison with other resources based on manual notes such as the World Atlas of Linguistic Structures (WALS) [48,50]. Therefore, in order to reduce as much as possible the bibliographic bias of the languages present in Universal Dependencies [54], we have opted for a typological balance.

To create our set, we have taken into consideration three basic typological requirements that influence many other grammatical aspects in languages and their behavior [55]:

(1) Difference in the order of the subject–verb relation.
(2) Difference in the order of the object–verb relation.
(3) Difference in the order of the noun–adjective relation.

We have managed to find a good balance on points (2) and (3); however, this has not been the case with the first aspect, since it is very uncommon for the verb to precede the subject as an unmarked order. Therefore, in the representative set, its presence is also lower. We respect the proportion seen in WALS of a one-tenth part. However, it should be noted that the ascription to a particular typological order is a convenient discrete simplification [56,57].

Subsequently, we have also tried to consider the following aspects of languages to set a useful representative set:

- Languages from different genus, representatives of the main families.
- Languages from different macro-areas.
- Languages with non-dominant order in different features.
- Isolated languages.
- Agglutinating languages.
- Languages with a greater and lesser degree of ascription to other characteristics such as, for example, the use of cases.
- Corpora with enough tokens and whose source of origin does not have any type of bias, as can be the case of FAQs corpus, for example.

After setting all these requirements, primary and secondary, we decided to use the data from the Universal Dependencies corpora [58]. This data source is chosen, firstly, because it annotates a lot of different languages by part-of-speech, constituents, and dependencies, and, secondly, because it is the only formalism in which MarsaGram [59] can be applied to automatically induce sets of syntactic constraints which can be used to match coincidences between them and our *U-FPGr* . After looking at the possibilities offered by Universal Dependencies, the set established consists of the following languages:

(1) Arabic (ar.);
(2) German (de.);
(3) Basque (eu.);
(4) Spanish (es.);
(5) Estonian (et.);
(6) Indonesian (id.);
(7) Korean (ko.);
(8) Turkish (tr.);
(9) Yoruba (yo.).

Our extension will have a value of 9, and the sets of $\langle low, medium, high \rangle$ will range as follows in Figure 2:

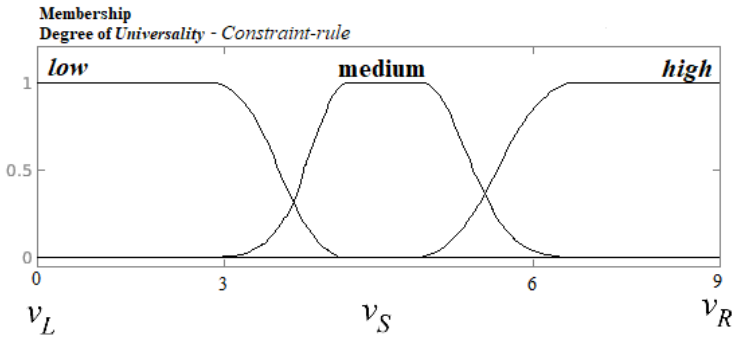

Figure 2. Linguistic Universality of the representative set as a Evaluative Expression.

1. IF a rule has between 0 and 3 coincidences, THEN the rule is a *low* universal and it is *high* in complexity.
2. IF a rule has between 4 and 6 coincidences, THEN the rule is a *medium* universal and it is *medium* in complexity.
3. IF a rule has between 7 and 9 coincidences, THEN the rule is a *high* universal and it is *low* in complexity.

Additionally, we have implemented an $IF - THEN$ rule to assign a weight value to each rule of the *U-FPGr*.

- If a rule in a set of languages coincides with a rule in the *U-FPGr*, then add +1.
- The more weight a rule has, the more universal it is in a representative set.
- The less weight a rule has, the less universal it is in a representative set.

As we have mentioned, we are aware that we cannot completely avoid bibliographical bias and, surely, the presence of a language representative of an unrepresented area or another language whose verb precedes the subject should be added. However, the model proposed here allows us to enrich the set once Universal Dependencies has such data in the future.

3.2. Application of the Tasks to Computationally Build a Universal Fuzzy Property Grammar

We have downloaded the Universal Dependency corpora for each of our sets of representative languages [58], and we have applied Marsagram [59]. Universal Dependency provides us with the constraint of dependency between constituents, and Marsagram automatically induces the constraints of *linearity*, "*co-occurrence*", "*exclusion*", and "*uniqueness*" over a Universal Dependency corpus.

Marsagram will provide us already with quantitative data; however, it is impossible to know which rules are coincident in a *U-FPGr*. Therefore, the interpretations that we obtain are more related with the notion of complexity rather than the notion of universality.

Marsagram presents data and rules in the following way:

Figure 3 is an extract for the marsagram from Arabic language. The rule means that verb as root excludes adjective as advcl next to ADJ as c-sub. Because of the fact that we are not interested in the other number, we will clean the data, erasing such noise. Therefore, to satisfy the coincidences, we will only keep the elements in #headproperty, symbol1, symbol2.

```
#headpropertysymbol-1symbol-2rulesoccurrencesfrequencebraking-rulesbreaking-occurrencesbreaking-frequencew0w1
VERB-rootexcludeADJ-advclADJ-csubj240.0006000.00001.00000.0006
```

Figure 3. Rule of the corpus or Arabic in Marsagram.

3.2.1. Building the Universal Fuzzy Property Grammar

To build the Universal Fuzzy Property Grammar (*U-FPGr*), we applied Equation (7). Table 1 is a representation of such a thing. To clarify, we take into account all the categories or part-of-speech (POS) for all languages according to the tagging in the universal dependencies. We then have 17 elements in POS. We consider the 64 dependencies that are present in the whole system of the universal dependencies. We consider the remaining four constraints. We combine this with two contextual linguistic elements, so, again, POS-dep is repeated twice. We obtain therefore, 4,242,536,496 rules. We repeat the same process again but considering the possibility that the rule only needs one contextual element, so POS-dep-properties-pos-dep. We obtain from that 4,734,976 rules. After summing up the both output of rules in terms of linguistic constraints, we obtain a *U-FPGr* with 4,247,273,472 rules belonging to the syntactic domain.

Table 1. Representation of the elements involved in the production of a *U-FPGr* for syntactic constraints.

POS (17)	DEP (64)	Properties (4)	Pos-dep (1088)	4,242,538,496
ADJ	acl	exclude	empty (1088)	4,734,976
ADP	acl:relcl	precede		4,247,273,472
ADV	advcl	unicity		
AUX	advmod	require		
CCONJ	advmod:emph			
DET	advmod:lmod			
INTJ	amod			
NOUN	appos			
NUM	aux			
PART	aux:pass			
PRON	case			
PROPN	cc			
PUNCT	cc:preconj			
SCONJ	ccomp			
SYM	clf			
VERB	compound			
X	..+48			

The technical summary is the following:

(1) Read the excel file with all the grammar rules.
(2) Make all possible combinations as strings. For example, one combination can look like: "$rule_1\ rule_2\ rule_3$".
(3) Insert all possible combinations into the MongoDB database.
The MongoDB database is not necessary, but it solves many problems instead of storing universal grammars in a pure text file. The text file may be huge and have an impact on time complexity for searching for grammar rules.

3.2.2. Preparing Languages for the Universal Fuzzy Property Grammar

As mentioned in Figure 3, the data of each language set had to be cleaned and prepared before checking coincidences. Therefore, we have followed these steps:

(1) We have a set of possible grammars for n languages stored in CSV/TSV files.
(2) For one language:
    (a) Load all possible files with language grammar rules.
    (b) Preprocessing (for example: remove empty spaces, replace *, . . .).

Finally, we have applied the weights to each rule, so we can measure its universality and complexity:

- For evaluating universality in one rule:
    (1) Send query to the MongoDB database for search rule in Universal Grammar.
    (2) If we found a rule in Universal Grammar, we insert a new row to the Pandas Dataframe, where the Universal Grammar column will have a current rule, and the column for the current language will have 1 and other languages will have 0.
    (3) If we not found a rule in Universal Grammar, we insert a Universal Grammar rule and 0 for all language columns.
    (4) Then, we can compute the total numbers for one rule, and we can put them into the Total column.

Table 2 is a visualization of the output, in this case, of the verb with the dependency of conjunction excluding two elements next to each other. We can observe how the final weights vary from one rule to the other. It is also clear how robust and flexible the system is. We only would need to add another column for any other language set that we would like to include to make our final weight of universality more representative.

Table 2. Example of output of rules with universality weight.

VERB-conj exclude CCONJ-cc VERB-ccomp, 0, 1, 1, 1, 1, 1, 1, 0, 7, 0.7777777777777778
VERB-conj exclude CCONJ-cc PART-advmod, 0, 1, 0, 0, 1, 1, 0, 0, 0, 3, 0.3333333333333333
VERB-conj exclude CCONJ-cc ADJ-amod, 0, 0, 1, 0, 0, 1, 1, 1, 0, 4, 0.4444444444444444
VERB-conj exclude CCONJ-cc NOUN-nsubj, 0, 1, 0, 0, 0, 1, 0, 0, 0, 2, 0.2222222222222222
VERB-conj exclude CCONJ-cc PROPN-obj, 0, 1, 1, 1, 0, 1, 1, 0, 0, 5, 0.5555555555555556
VERB-conj exclude CCONJ-cc PROPN-obl, 0, 1, 1, 1, 1, 1, 0, 1, 0, 6, 0.6666666666666666
VERB-conj exclude CCONJ-cc VERB-xcomp, 0, 1, 1, 1, 1, 1, 0, 0, 0, 5, 0.5555555555555556
VERB-conj exclude NOUN-obj NOUN-obl, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0.1111111111111111
VERB-conj exclude NOUN-obj PROPN-obj, 0, 1, 1, 1, 0, 1, 1, 0, 0, 5, 0.5555555555555556
VERB-conj exclude NOUN-obj PROPN-obl, 0, 1, 1, 1, 1, 1, 0, 1, 0, 6, 0.6666666666666666
VERB-conj exclude NOUN-obj VERB-xcomp, 0, 1, 1, 1, 1, 1, 0, 0, 0, 5, 0.5555555555555556
VERB-conj exclude NOUN-obj ADJ-amod, 0, 0, 1, 0, 0, 1, 1, 1, 0, 4, 0.4444444444444444
VERB-conj exclude NOUN-obj VERB-ccomp, 0, 1, 1, 1, 1, 1, 1, 1, 1, 8, 0.8888888888888888
VERB-conj exclude NOUN-obj NOUN-nsubj:pass, 0, 1, 1, 0, 0, 1, 0, 0, 0, 3, 0.3333333333333333
VERB-conj exclude NOUN-obj NOUN-compound, 0, 0, 0, 0, 1, 1, 1, 1, 0, 4, 0.4444444444444444
VERB-conj exclude NOUN-obj ADP-case, 0, 0, 1, 0, 0, 1, 1, 1, 0, 4, 0.4444444444444444
VERB-conj exclude NOUN-obj PRON-obj, 0, 1, 1, 1, 1, 1, 1, 1, 1, 8, 0.8888888888888888
VERB-conj exclude NOUN-obj VERB-acl, 0, 1, 0, 0, 0, 1, 1, 0, 0, 3, 0.3333333333333333
VERB-conj exclude NOUN-obj PRON-obl, 0, 1, 1, 1, 0, 1, 0, 1, 1, 6, 0.6666666666666666
VERB-conj exclude NOUN-obj VERB-fixed, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0.1111111111111111
VERB-conj exclude NOUN-obj NUM-nummod, 0, 0, 0, 0, 0, 1, 1, 0, 0, 2, 0.2222222222222222
VERB-conj exclude NOUN-obj PROPN-nsubj, 0, 1, 1, 1, 1, 1, 1, 0, 0, 6, 0.6666666666666666
VERB-conj exclude NOUN-obj PART-advmod, 0, 1, 0, 0, 1, 1, 0, 0, 0, 3, 0.3333333333333333
VERB-conj exclude NOUN-obj ADJ-xcomp, 0, 1, 1, 1, 0, 1, 0, 0, 0, 4, 0.4444444444444444
VERB-conj exclude NOUN-obj PRON-nsubi, 0, 1, 1, 1, 1, 1, 1, 1, 0, 7, 0.7777777777777778
VERB-conj exclude NOUN-obj VERB-advcl, 0, 1,1, 1, 1, 1, 0, 1, 0, 6, 0.6666666666666666

To evaluate complexity, it is necessary to just negate the value (apply -1), since universality and complexity work as opposites in terms of $\langle low, medium, high \rangle$. Therefore, a rule is that if its universality weights 0.7, its complexity would be 0.3; if its universality is 0.4, its complexity would be 0.6, and so on.

4. Results

Our results can be found in four outputs: the quantitative data of Marsagram for linguistic complexity, the quantitative data regarding number of rules by weights, distribution of the weighted rules per language, and coincidences between languages.

4.1. Quantitative Data of Marsagram

Table 3 and Figure 4 show the results of the data induction from the application of Marsagram on the representative set of the universal dependency corpora. We see a high disparity between sets, among which German is the set with the most structures (32 K) and constraints (54 K), and Yoruba is the set with the fewest structures (243) and constraints (1647). We appreciate how the induction creates constraints per structure, providing a bias in evaluating complexity. That is because German is the language that has more structures. Therefore, it is the language that will have more properties.

Table 3. Quantitative data of Marsagram.

Language	Trees	Structures	Constraints	Order	s/Constraints
German	150,921	32,242	54,410	sv-ndo	1.6875504
Korean	27,363	8853	48,097	sv-ov	5.43284762
Turkish	18,687	5275	30,937	sv-ov	5.86483412
Spanish	16,013	8114	28,808	ndo-vo	3.5504067
Estonian	30,972	9468	28,570	sv-vo	3.01753274
Arabic	19,738	11,226	21,062	vs-vo	1.8761803
Euskera	8993	3283	14,703	sv-ov	4.47852574
Indonesian	5593	3143	12,530	sv-vo	3.98663697
Yoruba	318	243	1647		6.77777778

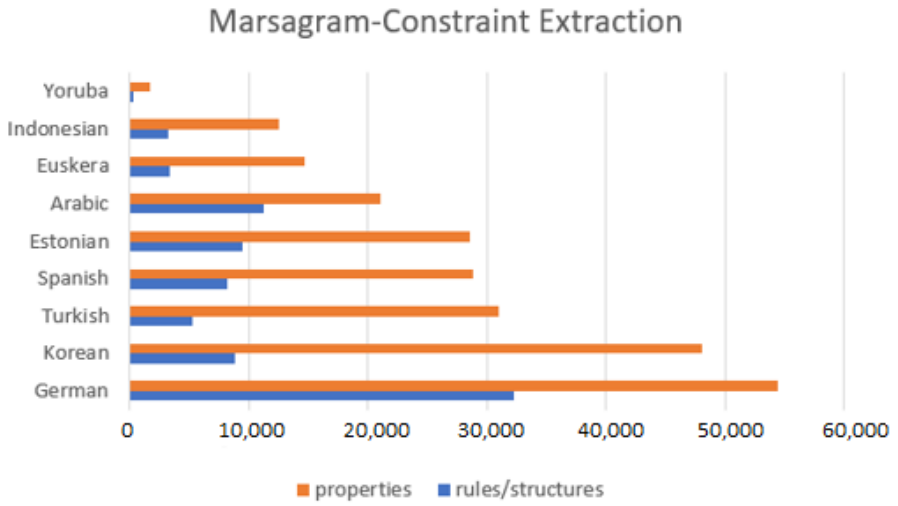

Figure 4. Quantitative data of Marsagram.

On the other hand, even though Estonian and Arabic have a similar amount of structures (9468 and 11,226) and constraints (28,570 and 21,062), they have a considerable difference in the number of trees as the input data (30,972, and 19,738). Additionally, they display fewer constraints than languages with fewer structures in the corpora, such as Turkish (5275 structures, 30,937 constraints) or Spanish (8114 structures, 28,808 constraints). Because of this data, we imply that it will be enough for future tasks to experiment with a corpus with around 20,000 dependency trees as the input data to generate structures and constraints of a language set.

However, we can apply an operation to compute the average of constraints per structure (s/Constraints column). Therefore, we will determine that languages with more constraints per structure are the most complex. According to this computation, we find that Turkish, Korean, and Euskera are the most complex language sets according to our operation, with 5.86, 5.43, and 4.47 constraints per structure, respectively. On the other hand, we disregard German and Yoruba for this first evaluation of linguistic complexity with Marsagram because of the over-generation of structures and constraints in the language set of German and the lack of data for the language set of Yoruba.

To sum up, the inductive data extracted from Marsagram can be helpful for a first glance to take into account complexity. However, it tells us nothing regarding universality, and it seems we cannot extract feasible results from asymmetric data.

4.2. Quantitative Data Regarding Number of Rules by Weights

Figure 5 shows the distribution of the number of constraints per weight. We observe that it is less frequent to find high universal constraints, which are present in all the

languages. On the contrary, there seems to be a peak in constraints of weights 1 and 2. Such a thing means that two languages bear a lot of specific constraints.

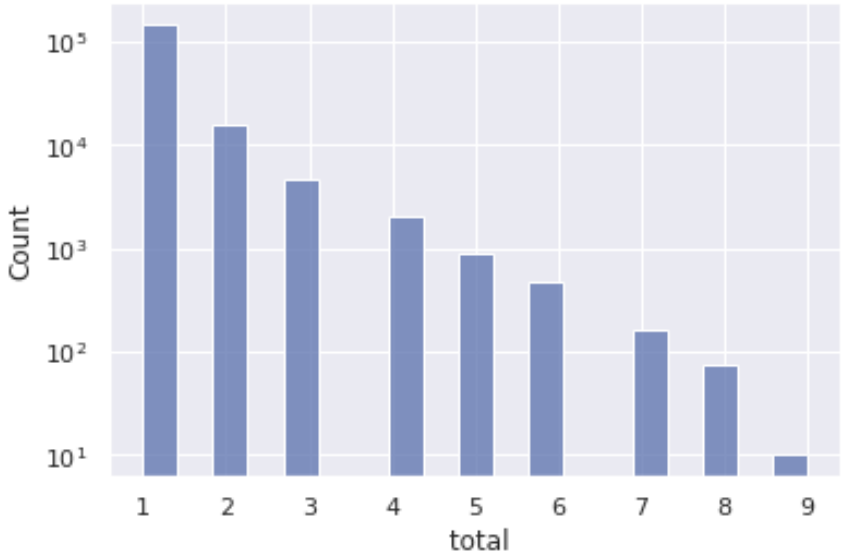

Figure 5. Quantitative data regarding number of rules by weights.

4.3. Distribution of the Weighted Rules per Set of Language

Figure 6 clarifies the data in Figure 5. The plot tends to converge in 9, since a weight of 9 means that all the sets have the constraints that weight 9. This plot displays high membership in the less universal rules. We now acknowledge that Korean, German, Turkish, and Euskera have most of the rules with weight 1 or 2. That means, in principle, that they will be the most complex languages and the ones that bear less universal constraints. It stands out how Korean has a membership degree of 1.0 for the rules of weight 1. This means that Korean has *all* the specific constraints found in the *U-FPGr*.

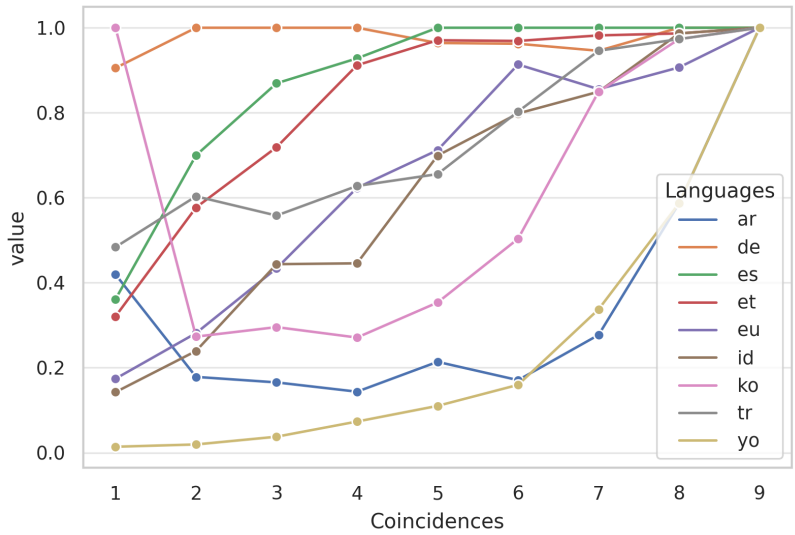

Figure 6. Distribution of the weighted rules per set of language.

4.4. Coincidences between Languages

Figure 7 displays the total number of rules found as coincident in our *U-FPGr*, and, simultaneously, the number of coincident rules between sets. For example, Spanish (es), and German (de) share 6968 rules. The matrix diagonal displays the total number of rules per set. For example, German has a total of (49,895), while Estonian has (22,477) rules. This matrix

demonstrates how our system cleaned the data from Marsagram from Figure 4, where according to the induction, German had 54,510 constraints, and Estonian had 28,570. That means that Marsagram was over-generating, as we were assuming previously. This result demonstrates that our architecture is good for cleaning data from inductive algorithms that might over-generate outputs.

Figure 8 is a failed output because the data are not symmetric, so we cannot evaluate the similarity of the sets: that is, how many constraints or rules they share in terms of degree [0–1]. Additionally, this output can be read as percentages. For example, it is better not to convert these outputs into percentages. For example, AR-DE 0.020263 or AR-ES 0.0290902 correspond to 2.0263% and 2.90902%, respectively. Because the values are too small, we do not think this provides much information. It makes sense that the coincidences are so low because every set represents different languages. If the number of coincidences was higher, those two sets with high coincidences could be considered part of the same language. Normalizing the matrix was a bad option since some sets had much lower data, and it was not possible to add data. Therefore, our solution was to compute a correlation between every pair of languages to obtain a correlation matrix such as Figure 9. To make it readable, we applied a heat scale of color.

Figure 9 is the correlation matrix of our set of languages with respect to each other. This matrix gives us information about how complex the sets between each other are in terms of sharing syntactic constraints:

- Red for low quantity of shared rules;
- Yellow for quite average quantity of shared rules;
- Green for quite high quantity of shared rules;
- Blue for high quantity of shared rules.

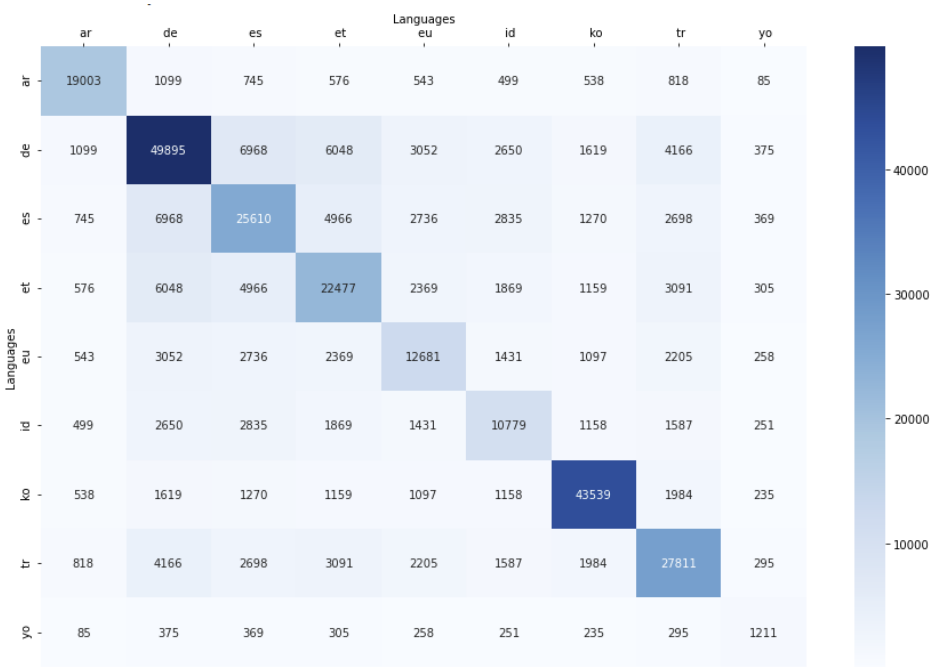

Figure 7. Coincidences between languages: number of rules.

Figure 8. Coincidences between languages: degree of similarity.

Figure 9. Coincidences between languages in a correlation matrix.

Therefore, we can represent gradually linguistic complexity in terms of evaluative expressions.

In this matrix, the closer to 1, the more similar. For example, Korean and German (ko-de) have a value of −0.3, while Korean and Spanish (ko-es) have a value of −0.19. The second relation expresses more similarity/sharing of rules than the first one since Therefore, Ko-es are more similar than ko-de and less complex because they share more rules that have a closer value to 1.

5. Discussion

With our research, we have corroborated the hypothesis that it is possible to build a system that characterizes both universality and relative complexity. The model proposed has been tested only with a proof-of-concept. However, we see clear potential despite the fact of its incompleteness. Some of the theoretical criticism that this model might receive include the following.

The first criticism can be that it only takes into account syntactic constraints. Therefore, it does not really measure a language's linguistic universality or complexity.

However, the *U-FPGr* is a model based on the FPGr. FPGr models vagueness from any linguistic concept in terms of degrees that can be described with linguistic constraints. FPGr is compatible with the theory of evaluative expressions of FNL. Consequently, any vague linguistic concept modeled with FPGr can also be modeled with an evaluative expression with the formalism of FNL. The main issue for such a thing is that, in fact, it is always necessary that constraints characterize the concept that will be defined. Therefore, one of the improvements of both FPGr and *U-FPGr* is that it needs a tradition that will describe each linguistic domain in terms of constraints. Such a thing is not that difficult in other domains, such as phonetics and phonology. To evaluate the universality of phones and phonemes, it would only be necessary to apply the same architecture to all phonetic and phonological tables of all the languages and/or dialects worldwide. The most coincident ones will be the most universal ones and the least complex ones. Therefore, it is not the architecture of either FPGr or *U-FPGr* which fails to represent universality and complexity. It is the limitations of the lack of tradition in defining linguistic domains in terms of constraints.

A second criticism is that it only considers a representative set of nine languages. Therefore, the results are not a reliable definition of what syntactic constraints are universal or complex.

We claim that as a proof-of-concept, our task succeeds in showing how the *U-FPGr* can characterize constraints and its complexity in terms of weights of universality, such as in Figure 5. From which image, it is easy to identify the sets of $\langle low, medium, high \rangle$ with 1–3, 4–6, and 7–9, respectively. Additionally, our task reveals that *high* universals are rare, which is something that goes in the same line as the linguistic literature. At the same time, it reveals that few languages, such as Korean or Turkish, trigger a lot of specific constraints. It should be very interesting and necessary that we would widen the data for further experiments, so we could corroborate if such languages keep their extreme specificity. It would also be exciting to include more language sets. However, as we have presented, the inclusion of more data per set or more sets of languages do not change the basic architecture of the *U-FPGr*. Our main issue is to provide a system that can, in fact, represent "*universality*", and "*complexity*" as a continuum. We are looking forward to testing our model with more data and seeing if the architecture is still robust.

A third criticism is that the constraints are built upon linguistic corpus, which has induced constraints from real text sources such as Wikipedia, reviews, and newspapers. Therefore, Marsagram induced constraints of non-grammatical sentences. In such a sense, it does not represent the real complexity of the standard variant of the languages in the representative set.

There is no such a problem with respect to the induction of non-grammatical constraints regarding the standard variety of a language. FPGr is fuzzy because it takes into account both grammatical and borderline constraints. Suppose the algorithm induces constraints that are not canonical within a language or dialectal constraints; that is more than welcome. We acknowledge a language definition similar to the definition of phoneme and sound. The phoneme /s/ is the abstract representation (in a non-academic way, the summary) of several sounds, such as (s) or (z). The distinction between (s) and (z) is, in particular, that one is voiceless and the other is voiced. However, most speakers would recognize them under the same perception. This phenomenon is represented abstractly as /s/. Following this same reasoning, for FPGr, the language is an abstract representation of all the possible performances of such language. Therefore, the English language /English/ can

be represented in different dialects or sociolects such as (geordie), (scouse), (apallachian), and so on. However, they are all "summarized" in the abstract representation of what English is. Therefore, if the specific constraints of these specific grammars of English are induced by an algorithm or included ad hoc by a linguist, it is more than welcome because if that constraint exists in English, it has to be included in the set of English language.

A fourth criticism is that the values in the correlation matrix display, in general, low values; therefore, it is not a valid representation of the degree of complexity.

We believe that the correlation matrix displays low values because, in fact, they are different languages. This output reinforces the idea that the architecture of *U-FPGr* is robust. Otherwise, if the values are too similar, we might even have to say that those sets of languages are alike and, probably, they are dialects of each other. This question brings up an interesting matter for our future work: testing *U-FPGr* with sets of dialects, or very close languages, such as the romance languages. If the values displayed in the correlation matrix are very close to value 1, it will reinforce the idea of the reliability of this architecture.

Finally, a criticism can be that the values of the relative complexity in Figure 9 are not meant to be necessarily equal between two sets since, even though two languages share a similar amount of rules, the rules that are not shared could be potentially more complex, affecting their degree of similarity. Therefore, it fails to represent possible hypothetical cases such as, for example, that for the majority of the speakers of German, it is easier to learn English than for English speakers to learn German. In this sense, the correlation of complexity should be represented asymmetrically.

However, our model does not reflect complexity in terms of difficulty between native speakers of a language. It would be very interesting to do so. Still, it would be necessary to include many more constraints in different domains to see more accurately how the constraints interact between domains and between domains with respect to other languages. It would probably be best to include the formalism of the agent-based models.

6. Conclusions and Future Work

We believe that the work presented here is a satisfactory proof-of-concept, which opens a new research line in pursuing the evaluation and definition of language universals and complexity. There is no such thing in linguistics as a "traditional method" or "fixed way" of computing such concepts of linguistic universality and complexity. Particularly, there is no proposed method that pursues to evaluate the concepts of universality and complexity as vague terms that can be defined in terms of degree. We believe that by considering such terms as gradient and fuzzy, we will be in a better position to describe multiple natural languages in linguistics, considering their idiosyncrasies and complexities. The best framework to do so is the theory of evaluative expressions of Fuzzy Natural Logic, which sets the basis to compute vague concepts with natural language and trichotomous expressions, together with Fuzzy Property Grammars, which provide the linguistic constraints to be evaluated. Furthermore, we believe this work opens a research line to make more appealing the fact of defining languages in terms of constraints to provide explicative or white-box methodologies for the characterization of languages and their features.

Regarding the future work of this research, it is necessary to test the model with data sets of dialects, and close related languages, such as romance languages. Therefore, we can test the model's outputs with data sets that, a priori, should display a larger number of coincidences and try to establish a fuzzy–numerical boundary between language and dialect. That is, to obtain a fuzzy–value which characterizes when a dialect starts to be considered a different language in terms of membership degrees. On the other hand, testing the model with larger and symmetric data sets would be necessary to reassure its robustness. Another test that has to be run in the future is to compute linguistic complexity taking into account other linguistic domains, such as computing similarity between lexicons and phonemic charts of different languages and dialects by incorporating an Optimality Theory approach. Similarly, it would be necessary to work on comparing language sets within the constraints of the morphological domain.

Author Contributions: Conceptualization, A.T.-U., M.D.J.-L. and A.B.-R.; Formal analysis, A.T.-U. and M.D.J.-L.; Software, D.A.; Writing—original draft, A.T.-U., M.D.J.-L. and A.B.-R.; Writing—review and editing, A.T.-U., M.D.J.-L. and A.B.-R. All authors have read and agreed to the published version of the manuscript.

Funding: This paper has been supported by the project CZ.02.2.69/0.0/0.0/18_053/0017856 "Strengthening scientific capacities OU II" and by Grant PID2020-120158GB-I00 funded by Ministerio de Ciencia e Innovación. Agencia Estatal de Investigación MCIN/AEI/10.13039/501100011033.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: https://universaldependencies.org/ (accessed on 12 December 2021), Marsagram corpus can be requested to adria.torrens@urv.cat.

Acknowledgments: We also want to give special thanks to Vilém Novák, Grégoire Montcheuil and Jan Hůla for their collaboration and support during this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nefdt, R.M. The Foundations of Linguistics: Mathematics, Models, and Structures. Ph.D. Thesis, University of St Andrews, St Andrews, Scotland, 2016.
2. Pullum, G.K. The central question in comparative syntactic metatheory. *Mind Lang.* **2013**, *28*, 492–521. [CrossRef]
3. Kortmann, B.; Szmrecsanyi, B. *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*; Walter de Gruyter & Co.: Berlin, Germany, 2012.
4. McWhorter, J.H. The Worlds Simplest Grammars Are Creole Grammars. *Linguist. Typology* **2001**, *5*, 125–166. [CrossRef]
5. Baechler, R.; Seiler, G. *Complexity, Isolation, and Variation*; de Gruyter, Walter GmbH & Co.: Berlin, Germany, 2016; Volume 57.
6. Baerman, M.; Brown, D.; Corbett, G.G. *Understanding and Measuring Morphological Complexity*; Oxford University Press: New York, NY, USA, 2015.
7. Coloma, G. *La Complejidad de Los Idiomas*; Peter Lang Limited, International Academic Publishers: Bern, Switzerland, 2017.
8. Conti Jiménez, C. *Complejidad lingüística: Orígenes y Revisión Crítica del Concepto de Lengua Compleja*; Peter Lang Limited, International Academic Publishers: Bern, Switzerland, 2018.
9. Di Domenico, E. *Syntactic Complexity From a Language Acquisition Perspective*; Cambridge Scholars Publishing: Newcastle upon Tyne, UK, 2017.
10. La Mantia, F.; Licata, I.; Perconti, P. *Language in Complexity: The Emerging Meaning*; Springer: Berlin/Heidelberg, Germany, 2016.
11. McWhorter, J.H. *Linguistic Simplicity and Complexity: Why Do Languages Undress?* Walter de Gruyter: Berlin, Germany, 2011; Volume 1.
12. Newmeyer, F.J.; Preston, L.B. *Measuring Grammatical Complexity*; Oxford University Press: New York, NY, USA, 2014.
13. Ortega, L.; Han, Z. *Complexity Theory and Language Development: In Celebration of Diane Larsen-Freeman*; John Benjamins Publishing Company: Amsterdam, The Netherlands, 2017; Volume 48.
14. Pallotti, G. A simple view of linguistic complexity. *Second. Lang. Res.* **2015**, *31*, 117–134. [CrossRef]
15. Dahl, Ö. *The Growth and Maintenance of Linguistic Complexity*; John Benjamins Publishing Company: Amsterdam, The Netherlands, 2004; Volume 10.
16. Miestamo, M.; Sinnemäki, K.; Karlsson, F. (Eds.) Grammatical Complexity in a Cross-Linguistic Perspective. In *Language Complexity: Typology, Contact, Change*; John Benjamins Publishing Company: Amsterdam, The Netherlands, 2008; pp. 22–42.
17. Trudgill, P. Contact and simplification: Historical baggage and directionality in linguistic change. *Linguist. Typology* **2001**, *5*, 371–374.
18. Küsters, W. *Linguistic Complexity*; LOT: Utrecht, The Netherlands, 2003.
19. Hawkins, J.A. An efficiency theory of complexity and related phenomena. In *Language Complexity as an Evolving Variable*; Sampson, S., Gil, D., Trudgill, P., Eds.; Oxford University Press: New York, NY, USA, 2009; pp. 252–268.
20. Andrason, A. Language complexity: An insight from complexsystem theory. *Int. J. Lang. Linguist.* **2014**, *2*, 74–89.
21. Moravcsik, A. Explaining language universals. In *The Oxford Handbook of Language Typology*; Song, J., Ed.; Oxford University Press: Oxford, UK, 2010; pp. 69–89.
22. Greenberg, J. *Universals of Language*; The MIT Press: Cambridge, MA, USA, 1963.
23. Comrie, B. *Language Universals and Linguistic Typology*; The University of Chicago Press: Great Britain, UK, 1989.
24. Dryer, M. Large linguistic areas and language sampling. *Stud. Lang.* **1989**, *13*, 257–292. [CrossRef]
25. Dryer, M. The Greenbergian Word Order Correlations. *Language* **1992**, *68*, 81–138. [CrossRef]

26. O'Horan, H.; Berzak, Y.; Vulic, I.; Reichart, R.; Korhonen, A. Survey on the Use of Typological Information in Natural Language Processing. In Proceedings of the COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, 11–16 December 2016; Scalise, S., Magni, E., Bisetto, A., Eds.; The COLING 2016 Organizing Committee: Osaka, Japan, 2016; pp. 1297–1308..
27. Ponti, E.M.; O'Horan, H.; Berzak, Y.; Vulic, I.; Reichart, R.; Poibeau, T.; Shutova, E.; Korhonen, A. Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing. *Comput. Linguist.* **2019**, *45*, 559–601. [CrossRef]
28. Lahiri, A.; Plank, F. Methods for Finding Language Universals in Syntax. In *Universals of Language Today*; Scalise, S., Magni, E., Bisetto, A., Eds.; Spinger: Berlin/Heidelberg, Germany, 2009; pp. 145–164.
29. Bakker, D. Language Sampling. In *The Oxford Handbook of Linguistic Typology*; Song, J.J., Ed.; Oxford University Press: Oxford, UK, 2010; pp. 1–26.
30. Chomsky, N. *Aspects of the Theory of Syntax*; The MIT Press: Cambridge, MA, USA, 1965.
31. Wohlgemuth, J.; Cysouw, M. *Rara and Rarissima*; De Gruyter Mouton: Berlin, Germany, 2010.
32. Wohlgemuth, J.; Cysouw, M. *Rethinking Universals: How Rarities Affect Linguistic Theory*; De Gruyter Mouton: Berlin, Germany, 2010.
33. Lahiri, A.; Plank, F. What Linguistic Universals Can be True of. In *Universals of Language Today*; Scalise, S., Magni, E., Bisetto, A., Eds.; Spinger: Berlin/Heidelberg, Germany, 2009; pp. 31–58.
34. Torrens Urrutia, A.; JiménezLópez, M.D.; Blache, P. Fuzziness and variability in natural language processing. In Proceedings of the 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Naples, Italy, 9–12 July 2017; pp. 1–6.
35. Torrens Urrutia, A. An approach to measuring complexity within the boundaries of a natural language fuzzy grammar. In Proceedings of the International Symposium on Distributed Computing and Artificial Intelligence, Toledo, Spain, 20–22 June 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 222–230.
36. Torrens Urrutia, A. A Formal Characterization of Fuzzy Degrees of Grammaticality for Natural Language. Ph.D. Thesis, Universitat Rovira i Virgili, Tarragona, Spain, 2019.
37. Novák, V. The Concept of Linguistic Variable Revisited. In *Recent Developments in Fuzzy Logic and Fuzzy Sets*; Sugeno, M., Kacprzyk, J., Shabazova, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2020; pp. 105–118.
38. Novák, V. Fuzzy Logic in Natural Language Processing. In Proceedings of the 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Naples, Italy, 9–12 July 2017.
39. Novák, V.; Perfilieva, I.; Dvořák, A. *Insight into Fuzzy Modeling*; Wiley & Sons: Hoboken, NJ, USA, 2016.
40. Novák, V. Mathematical Fuzzy Logic: From Vagueness to Commonsese Reasoning. In *Retorische Wissenschaft: Rede und Argumentation in Theorie und Praxis*; Kreuzbauer, G., Gratzl, N., Hielb, E., Eds.; LIT-Verlag: Wien, Austria, 2008; pp. 191–223.
41. Novák, V. What is Fuzzy Natural Logic. In *Integrated Uncertainty in Knowledge Modelling and Decision Making*; Huynh, V., Inuiguchi, M., Denoeux, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2015; pp. 15–18.
42. Novák, V. Fuzzy Natural Logic: Towards Mathematical Logic of Human Reasoning. In *Fuzzy Logic: Towards the Future*; Seising, R., Trillas, E., Kacprzyk, J., Eds.; Springer: Berlin/Heidelberg, Germany, 2015; pp. 137–165.
43. Novák, V. Evaluative linguistic expressions vs. fuzzy categories? *Fuzzy Sets Syst.* **2015**, *281*, 81–87. [CrossRef]
44. Novák, V. Mining information from time series in the form of sentences of natural language. *Int. J. Approx. Reason.* **2016**, *78*, 192–209. [CrossRef]
45. Urrutia, A.T.; López, M.D.J.; Brosa-Rodríguez, A. A Fuzzy Approach to Language Universals for NLP. In Proceedings of the 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Luxembourg, 11–14 July 2021; pp. 1–6.
46. Pagel, M. The History, Rate and Pattern of World Linguistic Evolution. In *The Evolutionary Emergence of Language*; Knight, M.S.K., Hurford, J., Eds.; Cambridge University Press: Cambridge, MA, USA, 2009; pp. 391–416.
47. Newymeyer, F. *Possible and Probable Languages: A Generative Perspective on Linguistic Typology*; Oxford University Press: Oxford, UK, 2007.
48. Dryer, M.; Haspelmath, M. The World Atlas of Language Structures Online. 2013. Available online: http://wals.info (accessed on 25 January 2022).
49. Plank, F.; Filimonova, E. The Universals Archive: A Brief Introduction for Prospective Users. *STUF Lang. Typol. Universals* **2000**, *53*, 109–123. [CrossRef]
50. Guzmán Naranjo, M.; Becker, L. Quantitative word order typology with UD. In Proceedings of the 17th International Workshop on Treebanks and Linguistic, Oslo, Norway, 13–14 December 2018; Linkoping University Electronic Press: Linkoping, Sweden, 2018; pp. 91–104.
51. Choi, H.; Guillaume, B.; Fort, K. Corpus-based language universals analysis using Universal Dependencies. In Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021), Sofia, Bulgaria, December 2021; pp. 33–44.
52. Schnell, S.; Schiborr, N.N. Crosslinguistic Corpus Studies in Linguistic Typology. *Annu. Rev. Linguist.* **2022**, *8*, 171–191. [CrossRef]
53. Levshina, N. Corpus-based typology: Applications, challenges and some solutions. *Linguist. Typology* **2021**, *26*, 129–160.
54. Nivre, J.; de Marneffe, M.; Ginter, F.; Goldberg, Y.; Hajic, J.; Manning, C.D.; McDonald, R.; Petrov, S.; Pyysalo, S.; Silveira, N.; et al. Universal dependencies v1: A multilingual treebank collection. In Proceedings of the Language Resources and Evaluation Conference, Portoroz, Slovenia, 23–28 May 2016; pp. 1659–1666.
55. Siewierska, A. *Word Order Rules*; Croom Helm: New York, NY, USA, 1988.

56. Gerdes, K.; Kahane, S.; Chen, X. Typometrics: From Implicational to Quantitative Universals in Word Order Typology. *Glossa J. Gen. Linguist.* **2021**, *6*, 17.
57. Levshina, N.; Namboodiripad, S.; Allassonnière-Tang, M.; Kramer, M.A.; Talamo, L.; Verkerk, A.; Wilmoth, S.; Garrido Rodriguez, G.; Gupton, T.; Kidd, E.; et al. Why we need a gradient approach to word order. *PsyArXiv* **2021**, 1–53. *preprint*.
58. Universal Dependency Corpora. Available online: https://universaldependencies.org/ (accessed on 12 December 2021).
59. Blache, P.; Rauzy, S.; Montcheuil, G. MarsaGram: An excursion in the forests of parsing trees. In Proceedings of the Language Resources and Evaluation Conference, Portoroz, Slovenia, 23–28 May 2016; p. 7.