


Article

Productivity and Predictability for Measuring Morphological Complexity

Ximena Gutierrez-Vasques ^{1,*} and Victor Mijangos ^{2,†}¹ Language and Space Lab, URPP Language and Space, University of Zurich, 8006 Zurich, Switzerland² Institute of Philological Research, National Autonomous University of Mexico, 04510 Mexico City, Mexico; vmijangosc@ciencias.unam.mx

* Correspondence: ximena.gutierrezv@spur.uzh.ch

† These authors contributed equally to this work.

Received: 31 October 2019; Accepted: 23 December 2019; Published: 30 December 2019



Abstract: We propose a quantitative approach for quantifying morphological complexity of a language based on text. Several corpus-based methods have focused on measuring the different word forms that a language can produce. We take into account not only the productivity of morphological processes but also the predictability of those morphological processes. We use a language model that predicts the probability of sub-word sequences within a word; we calculate the entropy rate of this model and use it as a measure of predictability of the internal structure of words. Our results show that it is important to integrate these two dimensions when measuring morphological complexity, since languages can be complex under one measure but simpler under another one. We calculated the complexity measures in two different parallel corpora for a typologically diverse set of languages. Our approach is corpus-based and it does not require the use of linguistic annotated data.

Keywords: language complexity; morphology; TTR; language model; entropy rate

1. Introduction

Languages of the world differ from each other in unpredictable ways [1,2]. Language complexity focuses on determine how these variations occurs in terms of complexity (size of grammar elements, internal structure of the grammar).

Conceptualizing and quantifying linguistic complexity is not an easy task, many quantitative and qualitative dimensions must be taken into account [3]. In general terms, the complexity of a system could be related to the number and variety of elements, but also to the elaborateness of their interrelational structure [4,5].

In recent years, morphological complexity has attracted the attention of the research community [1,6]. Morphology deals with the internal structure of words [7]. Several corpus-based methods are successful in capturing the number and variety of the morphological elements of a language by measuring the distribution of words over a corpus. However, they may not capture other complexity dimensions such as the predictability of the internal structure of words. There can be cases where a language is considered complex because it has a rich morphological productivity, i.e., great number of morphs can be encoded into a single word. However, the combinatorial structure of these morphs in the word formation process can have less uncertainty than other languages, i.e., more predictable.

We would like to quantify the morphological complexity by measuring the type and token distributions over a corpus, but also by taking into account the predictability of the sub-word sequences within a word [8].

We assume that the predictability of the internal structure of words reflects the difficulty of producing novel words given a set of lexical items (stems, suffixes or morphs). We take as our method

the statistical language models used in natural language processing (NLP), which are a useful tool for estimating a probability distribution over sequences of words within a language. However, we adapt this notion to the sub-word level. Information theory-based measures (entropy) can be used to estimate the predictiveness of these models.

Previous Work

Despite the different approaches and definitions of linguistic complexity, there are some main distinctions between the absolute and the relative complexity [3]. The former is defined in terms of the number of parts of a linguistic system; and the latter (more subjective) is related to the cost and difficulty faced by language users. Another important distinction includes global complexity that characterizes entire languages, e.g., as easy or difficult to learn. In contrast, particular complexity focuses only in a specific language level, e.g., phonological, morphological, syntactic.

In the case of morphology, languages of the world have different word production processes. Therefore, the amount of semantic and grammatical information encoded at the word level, may vary significantly from language to language. In this sense, it is important to quantify the morphological richness of languages and how it varies depending on their linguistic typology. Ackerman and Malouf [9] highlight two different dimensions that must be taken into account: the enumerative (e-complexity) that focuses on delimiting the inventories of language elements (number of morphosyntactic categories in a language and how they are encoded in a word); and the integrative complexity (i-complexity) that focuses on examining the systematic organization underlying the surface patterns of a language (difficulty of the paradigmatic system).

Coterell et al. [10] investigate a trade-off between the e-complexity and i-complexity of morphological systems. The authors propose a measure based on the size of a paradigm but also on how hard is to jointly predict all the word forms in a paradigm from the lemma. They conclude that “a morphological system can mark a large number of morphosyntactic distinctions [...] or it may have a high-level of unpredictability (irregularity); or neither. However, it cannot do both”.

Moreover, Bentz et al. [11] distinguishes between paradigm-based approaches that use typological linguistic databases for quantifying the number of paradigmatic distinctions of languages as an indicator of complexity; and corpus-based approaches that estimate the morphological complexity directly from the production of morphological instances over a corpus.

Corpus-based approaches represent a relatively easy and reproducible way to quantify complexity without the strict need for linguistic annotated data. Several corpus-based methods share the underlying intuition that morphological complexity depends on the morphological system of a language, such as its inflectional and derivational processes; therefore, a very productive system will produce a lot of different word forms. This morphological richness can be captured using information theory measures [12,13] or type-token relationships [14], just to mention a few.

It is important to mention that enumerative complexity has been approached using a paradigm-based or a corpus-based perspective. However, the methods that target the integrative complexity seem to be more paradigm-based oriented (which can restrict the number of languages covered). With that in mind, the measures that we present in this work are corpus-based and they do not require access to external linguistic databases.

2. Methodology

In this work, we quantify morphological complexity by combining two different measures over parallel corpora: (a) the type-token relationship (TTR); and (b) the entropy rate of a sub-word language model as a measure of predictability. In this sense, our approach could be catalogued as a corpus-based method for measuring absolute complexity of a specific language level (morphology).

2.1. The Corpora

Parallel corpora are a valuable resource for many NLP tasks and for linguistics studies. Translation documents preserve the same meaning and functions, to a certain extent, across languages. This allows analysis/comparison of the morphological and typological features of languages.

We used two different parallel corpora that are available for a wide set of languages. On one hand, we used a portion of the Parallel Bible Corpus [15]; in particular, we used a subset of 1150 parallel verses that overlapped across 47 languages (the selection of languages and pre-processing of this dataset was part of the Interactive Workshop on Measuring Language Complexity (IWMLC 2019) http://www.christianbentz.de/MLC2019_index.html). These languages are part of the WALS 100-language sample, a selection of languages that are typologically diverse [16] (<https://wals.info/languoid/samples/100>).

On the other hand, we used the JW300 parallel corpus that compiles magazine articles for many languages [17] (these articles were originally obtained from the Jehovah’s Witnesses website <https://www.jw.org>). In this case, we extracted a subset of 68 parallel magazine articles that overlapped across 133 languages. Table 1 summarizes information about the corpora.

Table 1. General information about the parallel corpora.

Corpus	Languages Covered	Total Tokens	Avg. Tokens Per Language
Bibles	47	1.1 M	24.8 K
JW300	133	22.4 M	168.9 K

We ran the experiments in both corpora independently. The intersection of languages covered by the two parallel corpora is 25. This shared set of languages was useful to compare the complexity rankings obtained with our measures, i.e., test if our complexity measures are consistent across different corpora.

It is important to mention that no sentence alignment was applied to the corpora. The Bibles corpus was already aligned at the verse level while the JW300 corpus was only aligned at the document level. However, for the aim of our experiments, alignment annotation (at the sentence or verse level) was not required.

2.2. Type-Token Relationship (TTR)

The type-token relationship (TTR) has proven to be a simple, yet effective, way to quantify the morphological complexity of a language using relatively small corpora [14]. It has also shown a high correlation with other types of complexity measures such as paradigm-based approaches that are based on typological information databases [11].

Morphologically rich languages will produce many different word forms (types) in a text, this is captured by measures such as TTR. From a linguistic perspective, Joan Bybee [18] affirms that “the token frequency of certain items in constructions [i.e., words] as well as the range of types [...] determines representation of the construction as well as its productivity”.

TTR can be influenced by the size of a text (Heaps’ law) or even by the domain of a corpus [19,20]. Some alternatives to make TTR more comparable include normalizing the text size or using logarithm, however, Covington and McFall [19] argue that these strategies are not fully successful, and they propose the moving-Average Type-Token Ratio. On the other hand, using parallel corpora has shown to be a simple way to make TTR more comparable across languages [21,22]. In principle, translations preserve the same meaning in two languages, therefore, there is no need for the texts to have the exact same length in tokens.

We calculated the TTR for a corpus by simply using Equation (1). Where *#types* are the different word types in the corpus (vocabulary size), and *#tokens* is the total number of word tokens in the

corpus. Values closer to 1 would represent greater complexity. This simple way of measuring TTR, without any normalization, has been used in similar works [11,22,23].

$$\text{TTR} = \frac{\#types}{\#tokens} \quad (1)$$

We use this measure as an easy way to approach the e-complexity dimension; i.e., different morphosyntactic distinctions, and their productivity, could be reflected in the type and token distribution over a corpus.

2.3. Entropy Rate of a Sub-Word Language Model

Entropy as a measure of unpredictability represents a useful tool to quantify different linguistic phenomena, in particular, the complexity of morphological systems [9,12,24].

Our method aims to reflect the predictability of the internal structure of words in a language. We conjecture that morphological processes that are irregular/suppletive, unproductive, etc., will increase the entropy of a model that predicts the probability of sequences of morphs/sub-word units within a word.

To do this, we estimate a stochastic matrix P , where each cell contains the transition probability between two sub-word units in that language (see example Table 2). These probabilities are estimated using the corpus and a neural language model that we will describe below.

Table 2. Toy example of a stochastic matrix using the trigrams contained in the word ‘cats’. The symbols #, \$ indicate beginning/end of a word.

	#ca	cat	ats	ts\$
#ca	0.01	0.06	0.07	0.33
cat	0.9	0.04	0.05	0.22
ats	0.06	0.78	0.05	0.23
ts\$	0.03	0.12	0.83	0.22

We calculate the stochastic matrix P as follows (2):

$$P = p_{ij} = p(w_j|w_i) \quad (2)$$

where w_i and w_j are sub-word units. We used a neural probabilistic language model to estimate a probability function.

2.3.1. Sub-Word Units

Regarding to sub-word units, one initial thought would be to use character sequences that correspond to the linguistic notion of morphemes/morphs. However, it could be difficult to perform morphological segmentation to all the languages in the corpora. There are unsupervised morphological segmentation approaches, e.g., Morfessor [25], BPE encoding [26], but they still require parameter tuning to control over-segmentation/under-segmentation (making these approaches not completely language independent).

Instead of this, we focused on fixed-length sequences of characters (n-grams), which is more easily applicable to all the languages in the corpora. This decision is also driven by the evidence that trigrams encode morphological properties of the word [27]. Moreover, in some tasks such as language modeling, the use of character trigrams seems to lead to better word vector representations than unsupervised morphological segmentations [28].

Therefore, we trained the language models using character trigrams. We also took into account unigrams (characters) sequences, since there are languages with syllabic writing systems in the datasets and in these cases a single character can encode a whole syllable.

2.3.2. Neural Language Model

Our model was estimated using a feedforward neural network; this network gets trained with pairs of consecutive n-grams that appear in the same word. Once the network is trained we can retrieve from the output layer the probability p_{ij} for any pair of n-grams. This architecture is based on [29]; however, we used character n-grams instead of words. The network comprises the following layers: (1) an input layer of one-hot vectors representing the n-grams; (2) an embedding layer; (3) a hyperbolic tangent hidden layer; (4) and finally, an output layer that contains the conditional probabilities obtained by a SoftMax function defined by Equation (3).

$$p_{ij} = \frac{e^{a_{ij}}}{\sum_k e^{a_{ik}}} \quad (3)$$

The factor a_{ij} in Equation (3) is the j th output of the network when the n-gram w_i is the input. The architecture of the network is presented in Figure 1.

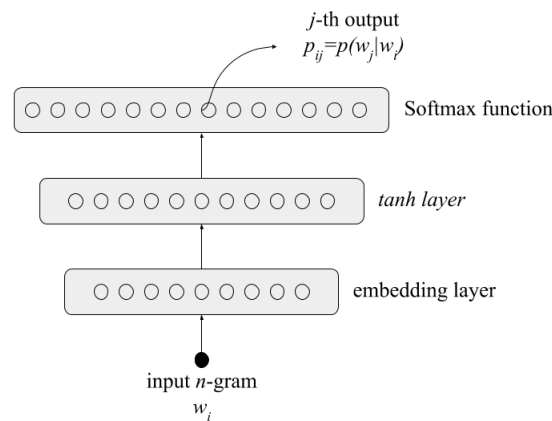


Figure 1. Neural probabilistic language model architecture, w_i, w_j are n-grams.

Once the neural network is trained, we can build the stochastic matrix P using the probabilities obtained for all the pairs of n-grams. We determine the entropy rate of the matrix (P) by using Equation (4) [30]:

$$H(P) = - \sum_{i=1}^N \mu_i \sum_{j=1}^N p_{ij} \log_N p_{ij} \quad (4)$$

where p_{ij} are the entries of the matrix P , N is the size of the n-grams vocabulary, and μ represents the stationary distribution. This stationary distribution can be obtained using Equation (5), for each $i = 1, \dots, N$:

$$\mu_i = \frac{1}{N} \sum_{k=1}^N p_{ik} \quad (5)$$

This equation defines a uniform distribution (we selected a uniform distribution since we observed that the stationary distribution, commonly defined by $P\mu = \mu$, was uniform for several small test corpora. Due to the neural probabilistic function, we can guarantee that the matrix P is irreducible; we assume that the irreducibility of the matrices is what determines the uniform stationary distribution. See [31]). To normalize the entropy, we use the logarithm base N . Thus, $H(P)$ can take values from 0 to 1. A value close to 1 would represent higher uncertainty in the sequence of n-grams within the words in a certain language, i.e., less predictability in the word formation processes.

The overall procedure can be summarized in the following steps: (the code is available at <http://github.com/elotlmx/complexity-model>)

1. For a given corpus, divide every word into its character n-grams. A vocabulary of size N (the number of n-grams) is obtained.

2. Calculate the probability of transitions between n-grams, $p_{ij} = p(w_j|w_i)$. This is done using the neural network described before.
3. A stochastic matrix $P = p_{ij}$ is obtained.
4. Calculate the entropy rate of the stochastic matrix $H(P)$.

3. Results

We applied the measures to each language contained in the JW300 and Bibles corpora. We use the notations H_1 , H_3 for the entropy rate calculated with unigrams and trigrams respectively; TTR is the type-token relationship.

To combine the different complexity dimensions, we ranked the languages according to each measure, then we averaged the obtained ranks for each language (since we ranked the languages from the most complex to the less complex, we used the inverse of the average in order to be consistent with the complexity measures (0 for least complex, 1 for the most complex)). The notation for these combined rankings are the following: TTR+ H_1 (TTR rank averaged with H_1 rank); TTR+ H_2 (TTR rank averaged with H_2 rank); TTR+ H_1 + H_3 (TTR rank averaged with H_1 and H_3 ranks). In all the cases the scales go from 0 to 1 (0 for the least complex and 1 for the most complex).

Tables 3 and 4 contain the measures described above for each corpus. These tables only show the set of 25 languages that are shared between the two corpora. In Figures 2 and 3 we plot these different complexities, and their combinations. The complete list of languages and results are included in Appendices A and B.

Table 3. Complexity measures on the Bibles corpus (H_1 : unigrams entropy; H_3 : trigrams entropy; TTR: Type-token relationship); bold numbers indicate the highest and the lowest values for each measure, the rank is in brackets.

Language	H_1	H_3	TTR	TTR+ H_1	TTR+ H_3	TTR+ H_1 + H_3
Arabic	0.726 (3)	0.748 (4)	0.31 (3)	0.333 (2)	0.333 (3)	0.333 (2)
Burmese	0.74 (2)	0.823 (2)	0.791 (1)	0.667 (1)	0.667 (1)	0.6 (1)
Eastern Oromo	0.652 (10)	0.573 (22)	0.196 (9)	0.105 (7)	0.065 (18)	0.073 (12)
English	0.703 (5)	0.667 (10)	0.082 (19)	0.083 (11)	0.069 (16)	0.088 (10)
Fijian	0.569 (19)	0.519 (24)	0.048 (24)	0.047 (21)	0.042 (24)	0.045 (24)
Finnish	0.696 (6)	0.59 (20)	0.266 (5)	0.182 (5)	0.08 (9)	0.097 (8)
French	0.607 (17)	0.609 (18)	0.139 (12)	0.069 (16)	0.067 (17)	0.064 (18)
Georgian	0.632 (12)	0.67 (9)	0.238 (6)	0.105 (7)	0.133 (5)	0.107 (5)
German	0.588 (18)	0.664 (12)	0.136 (13)	0.065 (17)	0.08 (9)	0.07 (13)
Hausa	0.61 (16)	0.614 (17)	0.098 (18)	0.059 (19)	0.057 (21)	0.059 (21)
Hindi	0.54 (22)	0.729 (6)	0.057 (22)	0.045 (23)	0.071 (13)	0.06 (20)
Indonesian	0.662 (9)	0.599 (19)	0.115 (17)	0.077 (12)	0.056 (22)	0.067 (16)
Korean	0.394 (25)	0.861 (1)	0.348 (2)	0.074 (14)	0.667 (1)	0.107 (5)
Modern Greek	0.683 (7)	0.655 (14)	0.181 (10)	0.118 (6)	0.083 (8)	0.097 (8)
Malagasy (Plateau)	0.568 (20)	0.519 (24)	0.14 (11)	0.065 (17)	0.056 (22)	0.054 (23)
Russian	0.751 (1)	0.732 (5)	0.225 (8)	0.222 (4)	0.154 (4)	0.214 (3)
Sango	0.538 (23)	0.56 (23)	0.025 (25)	0.042 (25)	0.042 (24)	0.042 (25)
Spanish	0.647 (11)	0.656 (13)	0.133 (15)	0.077 (12)	0.071 (13)	0.077 (11)
Swahili	0.613 (14)	0.576 (21)	0.233 (7)	0.091 (9)	0.071 (13)	0.07 (13)
Tagalog	0.632 (12)	0.629 (16)	0.121 (16)	0.071 (15)	0.063 (19)	0.068 (15)
Thai	0.554 (21)	0.752 (3)	0.055 (23)	0.045 (23)	0.074 (11)	0.063 (19)
Turkish	0.705 (4)	0.63 (15)	0.297 (4)	0.25 (3)	0.105 (6)	0.13 (4)
Vietnamese	0.406 (24)	0.684 (8)	0.066 (20)	0.045 (23)	0.071 (13)	0.058 (22)
Western Farsi	0.67 (8)	0.705 (7)	0.135 (14)	0.091 (9)	0.095 (7)	0.103 (7)
Yoruba	0.613 (14)	0.666 (11)	0.064 (21)	0.057 (20)	0.062 (20)	0.065 (17)

Table 4. Complexity measures on the JW300 corpus (H_1 : unigrams entropy; H_3 : trigrams entropy; TTR: Type-token relationship); bold numbers indicate the highest and the lowest values for each measure, the rank is in brackets.

Language	H_1	H_3	TTR	TTR+ H_1	TTR+ H_3	TTR+ H_1 + H_3
Arabic	0.586 (8)	0.826 (2)	0.171 (4)	0.166 (4)	0.333 (1)	0.214 (1)
Burmese	0.514 (19)	0.75 (5)	0.016 (22)	0.048 (23)	0.074 (12)	0.065 (17)
Eastern Oromo	0.552 (14)	0.568 (23)	0.111 (6)	0.1 (10)	0.068 (16)	0.069 (15)
English	0.682 (2)	0.712 (12)	0.053 (16)	0.111 (9)	0.071 (14)	0.1 (7)
Fijian	0.517 (18)	0.66 (17)	0.022 (21)	0.051 (21)	0.052 (23)	0.053 (21)
Finnish	0.563 (10)	0.628 (20)	0.184 (1)	0.181 (2)	0.095 (6)	0.096 (9)
French	0.522 (17)	0.673 (16)	0.072 (11)	0.071 (14)	0.074 (12)	0.068 (16)
Georgian	0.563 (10)	0.728 (9)	0.175 (2)	0.153 (6)	0.181 (2)	0.136 (3)
German	0.636 (3)	0.686 (14)	0.084 (9)	0.166 (4)	0.086 (9)	0.115 (5)
Hausa	0.527 (16)	0.619 (22)	0.035 (18)	0.058 (17)	0.05 (25)	0.053 (21)
Hindi	0.591 (6)	0.783 (3)	0.023 (19)	0.076 (12)	0.086 (9)	0.103 (6)
Indonesian	0.556 (12)	0.624 (21)	0.051 (17)	0.068 (15)	0.052 (23)	0.06 (19)
Korean	0.349 (24)	0.907 (1)	0.057 (14)	0.052 (20)	0.133 (4)	0.076 (14)
Modern Greek	0.594 (5)	0.753 (4)	0.09 (8)	0.153 (6)	0.166 (3)	0.176 (2)
Malagasy (Plateau)	0.499 (22)	0.537 (25)	0.062 (12)	0.058 (17)	0.054 (22)	0.05 (24)
Russian	0.5 (21)	0.722 (11)	0.137 (5)	0.076 (12)	0.125 (5)	0.081 (12)
Sango	0.385 (23)	0.724 (10)	0.01 (25)	0.041 (24)	0.057 (20)	0.051 (23)
Spanish	0.59 (7)	0.65 (18)	0.079 (10)	0.117 (8)	0.071 (14)	0.085 (10)
Swahili	0.598 (4)	0.565 (24)	0.098 (7)	0.181 (2)	0.064 (18)	0.085 (10)
Tagalog	0.514 (19)	0.676 (15)	0.054 (15)	0.057 (19)	0.066 (17)	0.06 (19)
Thai	0.552 (14)	0.74 (7)	0.013 (24)	0.051 (21)	0.064 (18)	0.065 (17)
Turkish	0.684 (1)	0.65 (18)	0.175 (2)	0.5 (1)	0.09 (8)	0.13 (4)
Vietnamese	0.344 (25)	0.692 (13)	0.014 (23)	0.041 (24)	0.055 (21)	0.049 (25)
Western Farsi	0.569 (9)	0.738 (8)	0.061 (13)	0.09 (11)	0.095 (6)	0.1 (7)
Yoruba	0.553 (13)	0.748 (6)	0.023 (19)	0.062 (16)	0.08 (11)	0.078 (13)

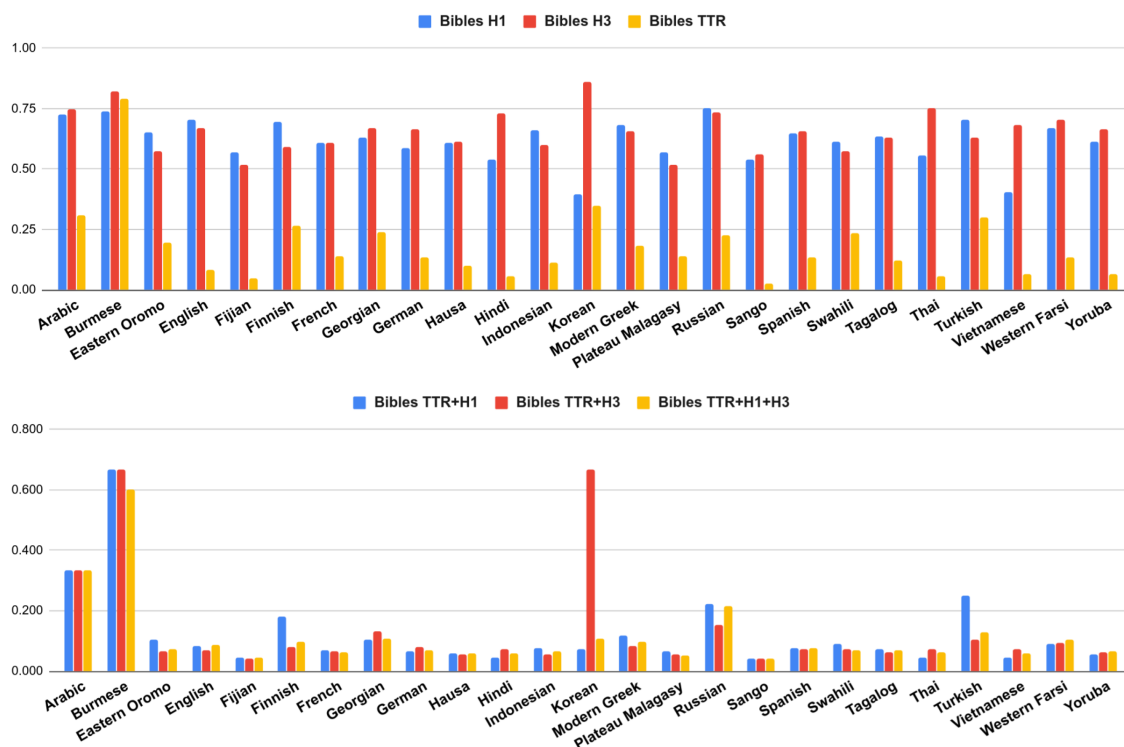


Figure 2. Different complexity measures (above) and their combinations (below) from Bibles corpus.

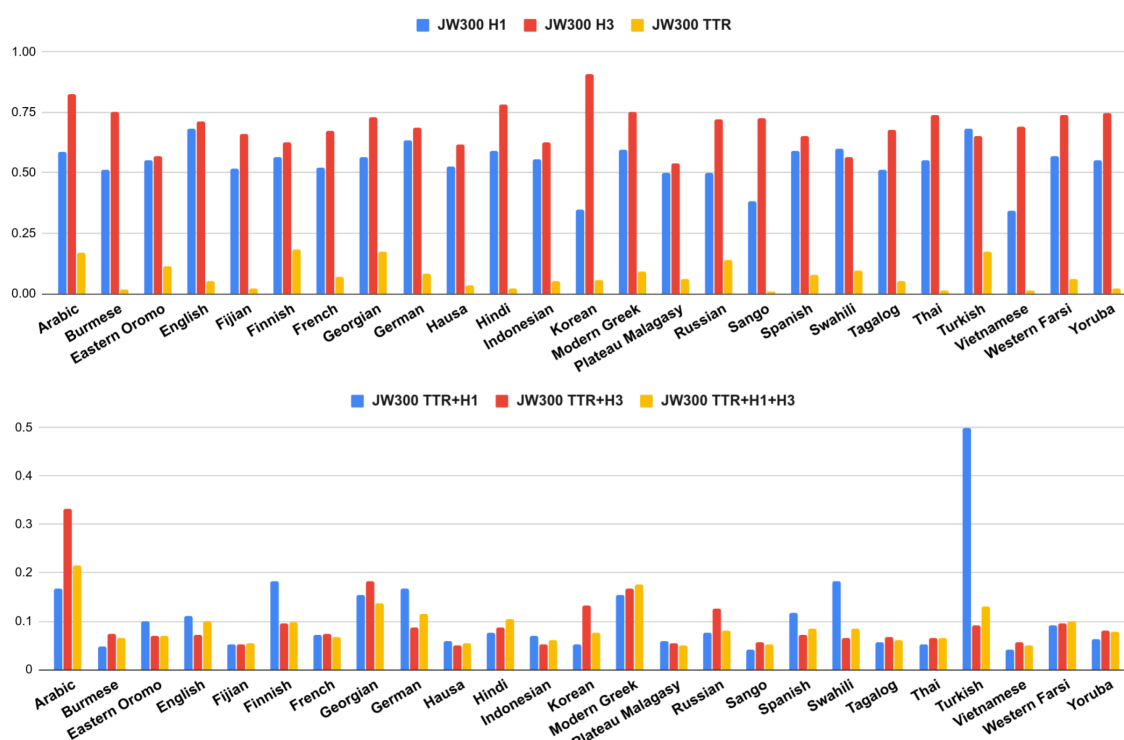


Figure 3. Different complexity measures (**above**) and their combinations (**below**) from JW300 corpus.

We can see that languages can be complex under one measure but simpler under another one. For instance, in Figures 2 and 3, we can easily notice that Korean is the most complex language if we only take into account the entropy rate using trigrams (H_3). However, this entropy dramatically drops using unigrams (H_1); therefore, when we combine the different measures, Korean is not the most complex language anymore.

There are cases such as English where its TTR is one of the lowest. This is expected since English is a language with poor inflectional morphology. However, its entropy is high. This suggests that a language such as English, usually not considered morphologically complex, may have many irregular forms that are not so easy to predict for our model.

We can also find the opposite case, where a language has a high TTR but low entropy, suggesting that it may produce many different word forms, but the inner structure of the words was “easy” to predict. This trend can be observed in languages such as Finnish (high TTR, low H_3), Korean (high TTR, low H_1) or Swahili (high TTR, low H_3).

The fact that a language has a low value of TTR does not necessarily imply that its entropy rate should be high (or vice versa). For instance, languages such as Vietnamese or Malagasy (Plateau), have some of the lowest values of entropy (H_1 , H_3); however, their TTR values are not among the highest in the shared subset. In this sense, these languages seem to have low complexity in both dimensions.

Burmese language constitutes a peculiar case, it behaves differently among the two corpora. Burmese complexity seems very high in all dimensions (TTR and entropies) just in the Bibles corpora. We conjecture that TTR is oddly high due to tokenization issues [32]: this is a language without explicit word boundary delimiters, if the words are not well segmented then the text will have many different long words without repetitions (high TTR). The tokenization pre-processing of the Bibles was based only on whitespaces and punctuation marks, while the JW300 had a more sophisticated tokenization. In the latter, Burmese obtained a very low TTR and H_1 entropy.

Cases with high complexity in both dimensions were less common. Arabic was perhaps the language that tends to be highly complex under both criteria (TTR and entropy) and this behavior

remained the same for the two corpora. We conjecture that this is related to the root-and-pattern morphology of the language, i.e., these types of patterns were difficult to predict for our sequential character n-grams language model. We will discuss more about this in Section 4.

3.1. Correlation across Corpora

Since our set of measures was applied to two different parallel corpora, we wanted to check if the complexities measures were, more or less, independent from the type of corpora used, i.e., languages should get similar complexity ranks in the two corpora.

We used Spearman's correlation [33] for the subset of shared languages across corpora. Table 5 shows the correlation coefficient for each complexity measure between the two corpora. Burmese language was excluded from the correlations due to the tokenization problems.

Table 5. Correlation of complexities between the JW300 and Bibles corpora (H_1 : unigrams entropy; H_3 : trigrams entropy; TTR: Type-token relationship).

	H_1	H_3	TTR	TTR+ H_1	TTR+ H_3	TTR+ H_1 + H_3
Correlation	0.520	0.782	0.890	0.776	0.858	0.765

Although the Bibles and the JW300 corpora belong to the same domain (religion), they greatly differ in size and in the topics covered (they are also parallel at different levels). Despite this, all the measures were positively correlated. The weaker correlation was obtained with H_1 , while complexity measures such as TTR or TTR+ H_3 were strongly correlated across corpora.

The fact that the complexity measures are correlated among the two corpora suggest that they are not very dependent of the corpus size, topics and other types of variations.

3.2. Correlation between Complexity Measures

In addition to the correlation across different corpora, we were interested in how the different complexity measures correlate between them (in the same corpus). Tables 6 and 7 show the Spearman's correlation between measures in each corpus.

Table 6. Spearman's correlations between measures in the corpus JW300 (all languages considered) (H_1 : unigrams entropy; H_3 : trigrams entropy; TTR: Type-token relationship).

	H_1	H_3	TTR	TTR+ H_1	TTR+ H_3	TTR+ H_1 + H_3
H_1	1.0	0.271	0.423	0.839	0.471	0.788
H_3	-	1.0	0.112	0.238	0.746	0.64
TTR	-	-	1.0	0.843	0.732	0.709
TTR+ H_1	-	-	-	1.0	0.72	0.892
TTR+ H_3	-	-	-	-	1.0	0.909
TTR+ H_1 + H_3	-	-	-	-	-	1.0

Table 7. Spearman's correlations between measures in the Bibles corpus (all languages considered) (H_1 : unigrams entropy; H_3 : trigrams entropy; TTR: Type-token relationship).

	H_1	H_3	TTR	TTR+ H_1	TTR+ H_3	TTR+ H_1 + H_3
H_1	1.0	0.276	0.384	0.828	0.464	0.810
H_3	-	1.0	0.006	0.152	0.693	0.585
TTR	-	-	1.0	0.815	0.654	0.637
TTR+ H_1	-	-	-	1.0	0.668	0.866
TTR+ H_3	-	-	-	-	1.0	0.862
TTR+ H_1 + H_3	-	-	-	-	-	1.0

In both corpora, the entropy-based measures (specially H_3) were poorly correlated (or not correlated) with the type-token relationship TTR. If these two types of measures are capturing, in fact, two different dimensions of the morphological complexity then it should be expected that they are not correlated.

The combined measures ($TTR+H_1$, $TTR+H_3$ and $TTR+H_1+H_3$) tend to be strongly correlated between them. It seems that all of them can combine, to some extent, the two dimensions of complexity (productivity and predictability).

Surprisingly, the entropy-based measures (H_1 and H_3) are weakly correlated between them, despite both trying to capture predictability. We conjecture that this could be related to the fact that for some languages, is more suitable to apply a trigram model and for some others the unigram model. For instance, in the case of Korean, one character is equivalent to a whole syllable (syllabic writing system). When we took combinations of three characters (trigrams) the model became very complex (high H_3), this does not necessarily reflect the real complexity. On the other hand, languages such as Turkish, Finnish or Yaqui (see Appendix B) obtained a very high value of H_1 (difficult to predict using only unigrams, very long words), but if we use the trigrams the entropy H_3 decrease, trigram models may be more appropriate for these type of languages.

3.3. Correlation with Paradigm-Based Approaches

Finally, we compared our corpus-based morphological complexity measures against two paradigm-based measures. First, we used the C_{WALS} measure proposed by [11], it is based on 28 morphological features/chapters extracted from the linguistic database WALS [16]. This measure maps each morphological feature to a numerical value, the complexity of a language is the average of the values of the morphological features.

The measure C_{WALS} was originally applied to 34 typologically diverse languages. However, we only took 19 languages (the shared set of languages with our Bibles corpus). We calculated the correlation between our complexity measures and C_{WALS} (Table 8).

In addition, we included the morphological counting complexity (MCC) as implemented by [34]. Their metric counts the number of inflectional categories for each language, the categories are obtained from the annotated lexicon UniMorph [35].

This measure was originally applied to 21 languages (mainly Indo-European), we calculated the correlation between MCC and our complexity measures using the JW300 corpus (which contained all of those 21 languages) Table 8.

Appendices C and D contain the list of languages used for each measure and the complexities.

Table 8. Spearman's correlation between C_{WALS} , MCC and our complexity measures (H_1 : unigrams entropy; H_3 : trigrams entropy; TTR: Type-token relationship).

	H_1	H_3	TTR	$TTR+H_1$	$TTR+H_3$	$TTR+H_1+H_3$
C_{WALS}	0.322	−0.392	0.882	0.730	0.395	0.406
MCC	0.064	0.024	0.851	0.442	0.585	0.366

C_{WALS} and TTR are strongly correlated, this was already pointed out by [11]. However, our entropy-based measures are weakly correlated with C_{WALS} , it seems that they are capturing different things. MCC metric shows a similar behavior, it is highly correlated with TTR but not with H_1 (unigrams entropy) or H_3 (trigrams entropy).

It has been suggested that databases such as WALS, which provide paradigmatic distinctions of languages, reflect mainly the e-complexity dimension [2]. This could explain the high correlation between C_{WALS} , MCC, and measures such as TTR. However, the i-complexity may be better captured by other types of approaches, e.g., the entropy rate measure that we have proposed.

The weak correlation between our entropy-based measures and C_{WALS} (even negative correlation in the case of H_3) could be a hint of the possible trade-off between the i-complexity and e-complexity. However, further investigation is needed.

4. Discussion

Our corpus-based measures tried to capture different dimensions that play a role in the morphological complexity of a language. H_1 and H_3 are focused on the predictability of the internal structure of words, while TTR is focused on how many different word forms can a language produce. Our results show that these two approaches poorly correlate, especially H_3 and TTR (0.112 for JW300 and 0.006 for the Bibles), which give us a lead that these quantitative measures are capturing different aspects of the morphological complexity.

This is interesting since, in fields such as NLP, languages are usually considered complex when their morphology allows them to encode many morphological elements within a word (producing many different word forms in a text). However, a language that is complex in this dimension can also be quite regular (low entropy) in its morphological processes, e.g., a predictable/regular process can be applied to a large number of roots, producing many different types; this is a common phenomenon in natural languages [36].

We can also think in the opposite case, a language with poor inflectional morphology may have low TTR; however, it may have suppletive/irregular patterns that will not be fully reflected in TTR but they will increase the entropy of a model that tries to predict these word forms.

The aim of calculating the entropy rate of our language models was to reflect the predictability of the internal structure of words (how predictable sequences of n-grams are in a given language). We think this notion is closer to the concept of morphological integrative complexity (i-complexity); however, there are probably many other additional criteria that play a role in this type of complexity. In any case, it is not common to find works that try to conceptualize this complexity dimension based only on raw corpora, our work could be an initial step towards that direction.

Measures such as H_3 , TTR (and all the combined versions) were consistent across the two parallel corpora. This is important since these corpora had different sizes and characteristics (texts from the JW300 corpus were significantly bigger than the Bibles one). These corpus-based measures may not necessarily require big amounts of text to grasp some typological differences and quantify the morphological complexity across languages.

The fact that measures such as C_{WALS} highly correlated with TTR but negative correlated with H_3 , suggests that C_{WALS} and TTR are capturing the same type of complexity, closer to the e-complexity criteria. This type of complexity may be easier to capture by several methods, contrary to the i-complexity dimension, which is related to the predictability of forms, among other morphological phenomena.

Adding typological information of the languages could help to improve the complexity analysis. As a preliminary analysis, in Appendix E we classified a subset of languages as concatenative vs isolating morphology using WALS. As expected, there is a negative (weak) correlation between the TTR and H_3 . However, this sign of possible trade-off is more evident in isolating languages compared to the ones that are classified as concatenative. This may be related to the fact that languages with isolating tendency do not produce many different word forms (low TTR); however, their derivative processes were difficult to predict for our sub-word language model (high entropy). More languages and exhaustive linguistic analysis are required.

One general advantage of our proposed measures for approaching morphological complexity is that they do not require linguistic annotated data such as morphological paradigms or grammars. The only requirement is to use parallel corpora, even if the texts are not fully parallel at the sentence level.

There are some drawbacks that are worth to discuss. We think that our approach of entropy rate of a sub-word language model may be especially suitable for concatenative morphology. For instance,

languages with root-and-pattern morphology may not be sequentially predictable, making the entropy of our models go higher (Arabic is an example); however, these patterns may be predictable using a different type of model.

Furthermore, morphological phenomena such as stem reduplication may seem quite intuitive from a language user perspective; however, if the stem is not frequent in the corpus, it could be difficult for our language model to capture these patterns. In general, derivational processes could be less predictable by our model than the inflectional ones (more frequent and systematic).

On the other hand, these measures are dealing with written language, therefore, they can be influenced by factors such as the orthography, the writing systems, etc. The corpus-based measures that we used, especially TTR, are sensitive to tokenization and word boundaries.

The lack of a “gold-standard” makes it difficult to assess the dimensions of morphological complexity that we are successfully capturing. The type-token relationship of a language seems to agree more easily with other complexity measures (Section 3.3). On the other hand, our entropy rate is based on sub-word units, this measure did not correlate with the type-token relationship, nor with the degree of paradigmatic distinctions obtained from certain linguistic databases. We also tested an additional characteristic, the average word length per language (see Appendix F), and this does not strongly correlate either with H_3 or H_1 .

Perhaps the question of whether this latter measure can be classified as i-complexity remains open. However, we think our entropy-based measure is reflecting to some extent the difficulty of predicting a word form in a language, since the entropy rate would increase with phenomena like: (a) unproductive processes; (b) allomorphy; (c) complex system of inflectional classes; and (d) suppletive patterns [37], just to mention a few.

Both approaches, TTR and the entropy rate of a sub-word language model, are valid and complementary, we used a very simple way to combine them (average of the ranks). In the future, a finer methodology can be used to integrate these two corpus-based quantitative approximations.

5. Future Work

In this section, we discuss some of the limitations that could be addressed as future work. The use of parallel corpora offers many advantages for comparing characteristics across languages. However, it is very difficult to find parallel corpora that cover a great amount of languages and that is freely available. Usually, the only available resources belong to specific domains, moreover, the parallel texts tend to be translations from one single language, e.g., English. It would be interesting to explore how these conditions affect the measurement of morphological complexity.

The character n-grams that we used for training the language models could be easily replaced by other types of sub-word units in our system. A promising direction could be testing different morphological segmentation models. Nevertheless, character trigrams seem to be a good initial point, at least for many languages, since these units may be capturing syllable information and this is related to morphological complexity [38,39].

Our way to control the influence of a language script system in the complexity measures was to consider two different character n-gram sizes. We noticed that trigrams (H_3) could be more suitable for languages with Latin script, while unigrams (H_1) may be better for other script systems (like Korean or Japanese). Automatic transliteration and other types of text pre-processing could be beneficial for this task.

There are still many open questions, as a future work we would like to make a more fine-grained typological analysis of the languages and complexity trends that resulted from these measures. Another promising research direction would be to quantify other processes that also play a role in the morphological complexity. For example, adding a tone in tonal languages is considered to add morphological complexity [3].

6. Conclusions

In this work we tried to capture two dimensions of morphological complexity. Languages that have a high TTR have the potential of encoding many different functions at the word level, therefore, they produce many different word forms. On the other hand, we proposed that the entropy rate of a sub-word language model could reflect how uncertain are the sequences of morphological elements within a word, languages with high entropy may have many irregular phenomena that are harder to predict than other languages. We were particularly interested in this latter dimension, since there are less quantitative methods, based on raw corpora, for measuring it.

The measures were consistent across two different parallel corpora. Moreover, the correlation between the different complexity measures suggest that our entropy rate approach is capturing a different complexity dimension than measures such as TTR or C_{WALS} .

Deeper linguistic analysis is needed; however, corpus-based quantitative measures can complement and deepen the study of morphological complexity.

Author Contributions: Conceptualization, V.M. and X.G.-V.; Investigation, X.G.-V. and V.M.; Methodology, X.G.-V. and V.M.; Writing—original draft, X.G.-V. and V.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Swiss Government Excellence Scholarship and the Mexican Council of Science and Technology (CONACYT). Fellowships 2019.0022 and 442471

Acknowledgments: We thank the reviewers, and Tanja Samardzic, for their valuable comments.

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

Appendix A. Complexity Measures for JW300 Corpus

Table A1. Complexity measures on the JW300 corpus (for all languages).

Language	H ₁	H ₃	TTR	TTR+H ₁	TTR+H ₃	TTR+H ₁ +H ₃
Afrikaans	0.566 (73)	0.674 (69)	0.047 (82)	0.013 (79)	0.013 (76)	0.013 (79)
Amharic	0.582 (56)	0.875 (4)	0.2 (8)	0.031 (22)	0.167 (1)	0.044 (8)
Arabic	0.586 (53)	0.827 (6)	0.171 (15)	0.029 (25)	0.095 (4)	0.041 (12)
Azerbaijani	0.661 (6)	0.728 (32)	0.151 (21)	0.074 (5)	0.038 (15)	0.051 (5)
Bicol	0.622 (18)	0.69 (57)	0.049 (79)	0.021 (44)	0.015 (63)	0.019 (40)
Cibemba	0.527 (107)	0.581 (115)	0.108 (39)	0.014 (73)	0.013 (79)	0.011 (99)
Bulgarian	0.56 (80)	0.68 (66)	0.091 (45)	0.016 (63)	0.018 (46)	0.016 (62)
Bislama	0.548 (88)	0.662 (75)	0.009 (132)	0.009 (120)	0.01 (117)	0.01 (112)
Bengali	0.546 (90)	0.801 (9)	0.06 (69)	0.013 (83)	0.026 (26)	0.018 (46)
Cebuano	0.543 (93)	0.708 (42)	0.051 (75)	0.012 (87)	0.017 (54)	0.014 (71)
Chuukese	0.579 (58)	0.618 (104)	0.037 (90)	0.014 (75)	0.01 (107)	0.012 (91)
Seychelles Creole	0.593 (46)	0.645 (87)	0.024 (107)	0.013 (77)	0.01 (107)	0.012 (85)
Czech	0.668 (4)	0.777 (13)	0.125 (29)	0.061 (10)	0.048 (9)	0.065 (4)
Danish	0.617 (22)	0.695 (53)	0.063 (65)	0.023 (36)	0.017 (55)	0.021 (34)
German	0.636 (14)	0.686 (62)	0.084 (50)	0.031 (22)	0.018 (47)	0.024 (29)
Ewe	0.488 (124)	0.717 (39)	0.05 (77)	0.01 (109)	0.017 (53)	0.012 (85)
Efik	0.61 (30)	0.657 (80)	0.043 (85)	0.017 (56)	0.012 (94)	0.015 (64)
Modern Greek	0.594 (44)	0.753 (19)	0.09 (47)	0.022 (40)	0.03 (21)	0.027 (22)
English	0.682 (3)	0.713 (41)	0.053 (74)	0.026 (29)	0.017 (51)	0.025 (26)
Spanish	0.59 (48)	0.65 (82)	0.079 (54)	0.02 (51)	0.015 (63)	0.016 (57)
Estonian	0.623 (17)	0.663 (74)	0.155 (19)	0.056 (12)	0.022 (32)	0.027 (22)
Western Farsi	0.569 (71)	0.739 (27)	0.061 (68)	0.014 (71)	0.021 (34)	0.018 (43)
Finnish	0.563 (75)	0.628 (96)	0.184 (9)	0.024 (34)	0.019 (40)	0.017 (52)
Fijian	0.517 (111)	0.66 (77)	0.022 (115)	0.009 (124)	0.01 (105)	0.01 (115)
French	0.522 (110)	0.674 (68)	0.072 (62)	0.012 (90)	0.015 (59)	0.012 (85)
Ga	0.547 (89)	0.664 (73)	0.046 (83)	0.012 (90)	0.013 (83)	0.012 (89)
Kiribati	0.506 (118)	0.592 (113)	0.031 (101)	0.009 (118)	0.009 (125)	0.009 (128)
Gujarati	0.542 (95)	0.835 (5)	0.048 (81)	0.011 (93)	0.023 (28)	0.017 (53)
Gun	0.575 (65)	0.691 (55)	0.024 (108)	0.012 (92)	0.012 (91)	0.013 (81)
Hausa	0.527 (106)	0.619 (102)	0.035 (94)	0.01 (107)	0.01 (109)	0.01 (114)
Hebrew	0.595 (43)	0.763 (17)	0.17 (16)	0.034 (19)	0.061 (6)	0.039 (13)
Hindi	0.591 (47)	0.783 (10)	0.022 (111)	0.013 (81)	0.017 (57)	0.018 (46)

Table A1. Cont.

Language	H ₁	H ₃	TTR	TTR+H ₁	TTR+H ₃	TTR+H ₁ +H ₃
Hiligaynon	0.564 (74)	0.699 (48)	0.045 (84)	0.013 (81)	0.015 (62)	0.015 (70)
Hiri Motu	0.543 (94)	0.604 (111)	0.012 (128)	0.009 (122)	0.008 (131)	0.009 (129)
Croatian	0.63 (16)	0.735 (30)	0.109 (38)	0.037 (15)	0.029 (22)	0.036 (16)
Haitian Creole	0.552 (85)	0.662 (76)	0.022 (114)	0.01 (106)	0.011 (104)	0.011 (105)
Hungarian	0.694 (1)	0.747 (22)	0.172 (14)	0.133 (2)	0.056 (7)	0.081 (3)
Armenian	0.575 (64)	0.736 (29)	0.117 (33)	0.021 (44)	0.032 (19)	0.024 (29)
Indonesian	0.556 (82)	0.624 (97)	0.051 (76)	0.013 (81)	0.012 (98)	0.012 (94)
Igbo	0.576 (60)	0.613 (107)	0.032 (99)	0.013 (83)	0.01 (115)	0.011 (101)
Iloko	0.611 (29)	0.64 (89)	0.08 (53)	0.024 (32)	0.014 (71)	0.018 (48)
Icelandic	0.637 (11)	0.704 (45)	0.09 (46)	0.035 (18)	0.022 (30)	0.029 (19)
Isoko	0.569 (70)	0.656 (81)	0.02 (116)	0.011 (99)	0.01 (111)	0.011 (102)
Italian	0.595 (40)	0.614 (106)	0.082 (52)	0.022 (41)	0.013 (87)	0.015 (65)
Japanese	0.302 (133)	0.914 (1)	0.024 (106)	0.008 (128)	0.019 (42)	0.012 (85)
Georgian	0.563 (77)	0.729 (31)	0.175 (12)	0.022 (38)	0.047 (10)	0.025 (27)
Kongo	0.534 (100)	0.619 (103)	0.022 (112)	0.009 (114)	0.009 (127)	0.01 (122)
Greenlandic	0.538 (98)	0.623 (99)	0.335 (1)	0.02 (47)	0.02 (38)	0.015 (65)
Cambodian	0.509 (117)	0.779 (12)	0.011 (129)	0.008 (129)	0.014 (69)	0.012 (96)
Kannada	0.587 (52)	0.754 (18)	0.239 (3)	0.036 (16)	0.095 (4)	0.041 (10)
Korean	0.349 (131)	0.907 (2)	0.057 (71)	0.01 (110)	0.027 (24)	0.015 (69)
Kikaonde	0.553 (83)	0.541 (127)	0.087 (48)	0.015 (68)	0.011 (99)	0.012 (96)
Kikongo	0.486 (126)	0.541 (128)	0.079 (55)	0.011 (94)	0.011 (103)	0.01 (118)
Kirghiz	0.563 (76)	0.695 (51)	0.144 (24)	0.02 (49)	0.027 (25)	0.02 (39)
Luganda	0.601 (36)	0.539 (129)	0.14 (25)	0.033 (20)	0.013 (79)	0.016 (61)
Lingala	0.526 (108)	0.633 (93)	0.04 (88)	0.01 (105)	0.011 (101)	0.01 (109)
Silosi	0.539 (97)	0.598 (112)	0.033 (97)	0.01 (103)	0.01 (119)	0.01 (116)
Lithuanian	0.637 (13)	0.706 (43)	0.167 (17)	0.067 (9)	0.033 (18)	0.041 (10)
Kiluba	0.544 (92)	0.56 (125)	0.112 (35)	0.016 (64)	0.012 (89)	0.012 (91)
Tshiluba	0.489 (123)	0.617 (105)	0.074 (60)	0.011 (96)	0.012 (94)	0.01 (107)
Luvale	0.545 (91)	0.525 (133)	0.145 (23)	0.018 (55)	0.013 (83)	0.012 (90)
Mizo	0.595 (42)	0.681 (65)	0.04 (87)	0.016 (67)	0.013 (78)	0.015 (63)
Latvian	0.582 (57)	0.745 (24)	0.123 (32)	0.022 (38)	0.036 (16)	0.027 (24)
Mauritian Creole	0.583 (55)	0.624 (98)	0.019 (117)	0.012 (90)	0.009 (127)	0.011 (103)
Plateau Malagasy	0.499 (122)	0.538 (131)	0.062 (66)	0.011 (100)	0.01 (111)	0.009 (124)
Marshallese	0.587 (51)	0.718 (38)	0.022 (113)	0.012 (86)	0.013 (76)	0.015 (67)
Macedonian	0.571 (68)	0.698 (49)	0.083 (51)	0.017 (58)	0.02 (38)	0.018 (46)
Malayalam	0.607 (32)	0.701 (47)	0.272 (2)	0.059 (11)	0.041 (14)	0.037 (15)
Moore	0.561 (79)	0.724 (34)	0.027 (104)	0.011 (96)	0.014 (66)	0.014 (75)
Marathi	0.612 (27)	0.738 (28)	0.095 (44)	0.028 (27)	0.028 (23)	0.03 (18)
Maltese	0.616 (24)	0.683 (63)	0.075 (59)	0.024 (33)	0.016 (58)	0.021 (36)
Burmese	0.514 (113)	0.75 (20)	0.016 (121)	0.009 (126)	0.014 (69)	0.012 (93)
Nepali	0.524 (109)	0.768 (15)	0.096 (43)	0.013 (76)	0.034 (17)	0.018 (44)
Niuean	0.389 (129)	0.646 (86)	0.013 (125)	0.008 (131)	0.009 (122)	0.009 (131)
Dutch	0.604 (34)	0.683 (64)	0.061 (67)	0.02 (50)	0.015 (61)	0.018 (42)
Norwegian	0.605 (33)	0.723 (35)	0.056 (72)	0.019 (53)	0.019 (42)	0.021 (34)
Sepedi	0.514 (114)	0.637 (90)	0.037 (91)	0.01 (111)	0.011 (101)	0.01 (112)
Chichewa	0.567 (72)	0.562 (124)	0.124 (31)	0.019 (52)	0.013 (81)	0.013 (80)
Eastern Oromo	0.552 (86)	0.568 (121)	0.111 (36)	0.016 (61)	0.013 (85)	0.012 (88)
Ossetian	0.575 (63)	0.688 (61)	0.077 (57)	0.017 (60)	0.017 (55)	0.017 (53)
Punjabi	0.572 (66)	0.816 (7)	0.025 (105)	0.012 (88)	0.018 (47)	0.017 (51)
Pangasinan	0.612 (28)	0.66 (78)	0.058 (70)	0.02 (46)	0.014 (73)	0.017 (49)
Papiamentu (Curaçao)	0.603 (35)	0.704 (46)	0.031 (102)	0.015 (70)	0.014 (73)	0.016 (55)
Solomon Islands Pidgin	0.642 (9)	0.64 (88)	0.013 (123)	0.015 (69)	0.009 (122)	0.014 (76)
Polish	0.617 (23)	0.745 (23)	0.152 (20)	0.047 (13)	0.047 (10)	0.045 (6)
Ponapean	0.533 (102)	0.576 (118)	0.032 (98)	0.01 (107)	0.009 (129)	0.009 (123)
Portuguese	0.595 (41)	0.697 (50)	0.075 (58)	0.02 (47)	0.019 (44)	0.02 (38)
Romanian	0.609 (31)	0.695 (52)	0.071 (63)	0.021 (43)	0.017 (51)	0.021 (36)
Russian	0.5 (121)	0.722 (37)	0.137 (26)	0.014 (74)	0.032 (20)	0.016 (57)
Kirundi	0.534 (101)	0.636 (91)	0.15 (22)	0.016 (62)	0.018 (49)	0.014 (73)

Table A1. Cont.

Language	H ₁	H ₃	TTR	TTR+H ₁	TTR+H ₃	TTR+H ₁ +H ₃
Kinyarwanda	0.599 (38)	0.57 (120)	0.134 (28)	0.03 (24)	0.014 (73)	0.016 (59)
Sango	0.385 (130)	0.725 (33)	0.01 (130)	0.008 (133)	0.012 (91)	0.01 (111)
Sinhala	0.578 (59)	0.742 (25)	0.079 (56)	0.017 (56)	0.025 (27)	0.021 (34)
Slovak	0.614 (26)	0.767 (16)	0.124 (30)	0.036 (17)	0.043 (12)	0.042 (9)
Slovenian	0.637 (12)	0.69 (56)	0.111 (37)	0.041 (14)	0.022 (32)	0.029 (20)
Samoa	0.536 (99)	0.629 (95)	0.017 (119)	0.009 (117)	0.009 (125)	0.01 (121)
Shona	0.622 (19)	0.538 (130)	0.18 (10)	0.069 (7)	0.014 (67)	0.019 (41)
Albanian	0.648 (8)	0.723 (36)	0.073 (61)	0.029 (26)	0.021 (37)	0.029 (20)
Sranantongo	0.54 (96)	0.562 (123)	0.01 (131)	0.009 (125)	0.008 (133)	0.009 (132)
Sesotho (Lesotho)	0.465 (128)	0.58 (116)	0.033 (95)	0.009 (123)	0.009 (122)	0.009 (130)
Swedish	0.621 (20)	0.706 (44)	0.066 (64)	0.024 (34)	0.019 (44)	0.023 (31)
Swahili	0.598 (39)	0.566 (122)	0.098 (41)	0.025 (31)	0.012 (93)	0.015 (68)
Swahili (Congo)	0.562 (78)	0.586 (114)	0.098 (41)	0.017 (59)	0.013 (82)	0.013 (82)
Tamil	0.618 (21)	0.715 (40)	0.234 (6)	0.074 (5)	0.043 (12)	0.045 (7)
Telugu	0.66 (7)	0.811 (8)	0.211 (7)	0.143 (1)	0.133 (2)	0.136 (1)
Thai	0.552 (87)	0.74 (26)	0.013 (124)	0.009 (112)	0.013 (75)	0.013 (83)
Tigrinya	0.666 (5)	0.891 (3)	0.162 (18)	0.087 (4)	0.095 (4)	0.115 (2)
Tiv	0.576 (61)	0.659 (79)	0.017 (120)	0.011 (94)	0.01 (113)	0.012 (98)
Tagalog	0.514 (115)	0.676 (67)	0.054 (73)	0.011 (100)	0.014 (67)	0.012 (94)
Otetela	0.529 (105)	0.605 (110)	0.085 (49)	0.013 (78)	0.013 (88)	0.011 (100)
Setswana	0.503 (120)	0.612 (108)	0.031 (100)	0.009 (120)	0.01 (118)	0.009 (127)
Tongan	0.532 (103)	0.688 (60)	0.023 (110)	0.009 (115)	0.012 (96)	0.011 (104)
Chitonga	0.558 (81)	0.647 (85)	0.177 (11)	0.022 (41)	0.021 (35)	0.017 (50)
Tok Pisin	0.575 (62)	0.632 (94)	0.008 (133)	0.01 (104)	0.009 (130)	0.01 (109)
Turkish	0.684 (2)	0.65 (83)	0.175 (13)	0.133 (2)	0.021 (35)	0.031 (17)
Tsonga	0.572 (67)	0.571 (119)	0.036 (93)	0.012 (85)	0.009 (124)	0.011 (106)
Tatar	0.593 (45)	0.689 (58)	0.116 (34)	0.025 (30)	0.022 (31)	0.022 (32)
Chitumbuka	0.588 (49)	0.534 (132)	0.108 (40)	0.022 (38)	0.012 (97)	0.014 (78)
Twi	0.469 (127)	0.664 (72)	0.039 (89)	0.009 (116)	0.012 (90)	0.01 (107)
Tahitian	0.487 (125)	0.669 (70)	0.012 (127)	0.008 (130)	0.01 (111)	0.009 (126)
Ukrainian	0.601 (37)	0.775 (14)	0.136 (27)	0.031 (22)	0.049 (8)	0.038 (14)
Umbundu	0.531 (104)	0.56 (126)	0.048 (80)	0.011 (98)	0.01 (115)	0.01 (119)
Urdu	0.631 (15)	0.781 (11)	0.033 (96)	0.018 (54)	0.019 (42)	0.025 (28)
Venda	0.512 (116)	0.619 (101)	0.031 (103)	0.009 (118)	0.01 (114)	0.009 (125)
Vietnamese	0.344 (132)	0.692 (54)	0.014 (122)	0.008 (131)	0.011 (100)	0.01 (117)
Waray-Waray	0.586 (54)	0.665 (71)	0.042 (86)	0.014 (72)	0.013 (85)	0.014 (72)
Wallisian	0.517 (112)	0.577 (117)	0.013 (126)	0.008 (127)	0.008 (132)	0.008 (133)
Xhosa	0.615 (25)	0.647 (84)	0.237 (4)	0.069 (7)	0.023 (29)	0.027 (24)
Yapese	0.639 (10)	0.635 (92)	0.018 (118)	0.016 (65)	0.01 (120)	0.014 (76)
Yoruba	0.553 (84)	0.749 (21)	0.023 (109)	0.01 (102)	0.015 (59)	0.014 (73)
Maya	0.587 (50)	0.688 (59)	0.05 (78)	0.016 (65)	0.015 (65)	0.016 (60)
Zande	0.505 (119)	0.62 (100)	0.037 (92)	0.009 (112)	0.01 (105)	0.01 (120)
Zulu	0.57 (69)	0.609 (109)	0.235 (5)	0.027 (28)	0.018 (50)	0.016 (55)

Appendix B. Complexity Measures Bibles Corpus

Table A2. Complexity measures on the Bibles corpus (for all languages).

Language	H ₁	H ₃	TTR	TTR+H ₁	TTR+H ₃	TTR+H ₁ +H ₃
Amele	0.568 (37)	0.59 (29)	0.134 (26)	0.031 (36)	0.036 (34)	0.032 (36)
Alamblak	0.673 (11)	0.643 (18)	0.203 (15)	0.076 (8)	0.06 (8)	0.068 (9)
Bukiyip	0.651 (16)	0.591 (28)	0.119 (32)	0.041 (25)	0.033 (37)	0.039 (28)
Apurinã	0.592 (29)	0.523 (43)	0.205 (14)	0.046 (19)	0.035 (36)	0.034 (33)
Mapudungun	0.598 (27)	0.596 (27)	0.145 (20)	0.041 (25)	0.042 (20)	0.04 (26)
Egyptian Arabic	0.725 (5)	0.748 (4)	0.31 (4)	0.222 (2)	0.25 (3)	0.23 (2)
Barasana-Eduria	0.526 (45)	0.577 (35)	0.146 (19)	0.031 (36)	0.037 (31)	0.03 (40)
Chamorro	0.678 (10)	0.663 (13)	0.13 (29)	0.051 (17)	0.046 (16)	0.056 (12)
German	0.588 (30)	0.663 (13)	0.136 (24)	0.037 (29)	0.054 (12)	0.044 (18)
Daga	0.585 (32)	0.545 (41)	0.095 (39)	0.028 (40)	0.025 (44)	0.026 (44)
Modern Greek	0.683 (9)	0.655 (16)	0.181 (17)	0.076 (8)	0.06 (8)	0.071 (7)
English	0.703 (7)	0.667 (10)	0.082 (40)	0.042 (22)	0.04 (24)	0.052 (13)
Basque	0.655 (14)	0.588 (31)	0.224 (13)	0.074 (10)	0.045 (17)	0.051 (15)

Table A2. Cont.

Language	H ₁	H ₃	TTR	TTR+H ₁	TTR+H ₃	TTR+H ₁ +H ₃
Fijian	0.568 (37)	0.519 (44)	0.048 (46)	0.024 (42)	0.022 (47)	0.023 (46)
Finnish	0.696 (8)	0.589 (30)	0.266 (6)	0.142 (5)	0.055 (10)	0.068 (9)
French	0.606 (25)	0.609 (24)	0.139 (23)	0.041 (25)	0.042 (20)	0.041 (23)
Paraguayan Guaraní	0.613 (21)	0.642 (19)	0.174 (18)	0.051 (17)	0.054 (12)	0.051 (15)
Eastern Oromo	0.652 (15)	0.573 (38)	0.196 (16)	0.064 (12)	0.037 (31)	0.043 (19)
Hausa	0.609 (24)	0.613 (23)	0.098 (38)	0.032 (32)	0.032 (39)	0.035 (30)
Hindi	0.54 (43)	0.729 (6)	0.057 (43)	0.023 (43)	0.04 (24)	0.032 (36)
Indonesian	0.661 (13)	0.598 (26)	0.115 (34)	0.042 (22)	0.033 (37)	0.041 (23)
Popti'	0.624 (20)	0.646 (17)	0.108 (37)	0.035 (30)	0.037 (31)	0.04 (26)
Kalaallisut	0.572 (35)	0.455 (47)	0.542 (2)	0.054 (15)	0.04 (24)	0.035 (30)
Georgian	0.632 (18)	0.67 (9)	0.238 (9)	0.071 (11)	0.111 (5)	0.081 (5)
West Kewa	0.573 (34)	0.583 (33)	0.113 (35)	0.028 (40)	0.029 (41)	0.029 (41)
Halh Mongolian	0.745 (3)	0.601 (25)	0.228 (11)	0.142 (5)	0.055 (10)	0.076 (6)
Korean	0.393 (47)	0.861 (1)	0.348 (3)	0.04 (27)	0.5 (2)	0.058 (11)
Lango (Uganda)	0.602 (26)	0.558 (40)	0.112 (36)	0.032 (32)	0.026 (43)	0.029 (41)
San Miguel El Grande Mixtec	0.57 (36)	0.614 (22)	0.125 (30)	0.03 (39)	0.038 (27)	0.034 (33)
Burmese	0.739 (4)	0.822 (2)	0.791 (1)	0.4 (1)	0.666 (1)	0.428 (1)
Wichí Lhamtés Güisnay	0.586 (31)	0.585 (32)	0.117 (33)	0.031 (36)	0.03 (40)	0.031 (38)
Nama (Namibia)	0.576 (33)	0.665 (11)	0.131 (28)	0.032 (32)	0.05 (14)	0.041 (23)
Western Farsi	0.67 (12)	0.705 (7)	0.135 (25)	0.054 (15)	0.062 (7)	0.068 (9)
Plateau Malagasy	0.567 (39)	0.518 (45)	0.14 (21)	0.032 (32)	0.029 (41)	0.028 (43)
Imbabura Highland Quichua	0.598 (27)	0.492 (46)	0.249 (8)	0.057 (14)	0.037 (31)	0.037 (29)
Russian	0.75 (1)	0.732 (5)	0.225 (12)	0.153 (4)	0.117 (4)	0.166 (3)
Sango	0.537 (44)	0.56 (39)	0.024 (47)	0.021 (47)	0.023 (46)	0.023 (46)
Spanish	0.647 (17)	0.656 (15)	0.133 (27)	0.045 (20)	0.047 (15)	0.05 (17)
Swahili	0.612 (22)	0.575 (36)	0.233 (10)	0.06 (13)	0.043 (18)	0.043 (19)
Tagalog	0.632 (18)	0.629 (20)	0.121 (31)	0.04 (27)	0.038 (27)	0.042 (21)
Thai	0.554 (41)	0.752 (3)	0.055 (44)	0.023 (43)	0.042 (20)	0.034 (33)
Turkish	0.705 (6)	0.629 (20)	0.297 (5)	0.181 (3)	0.08 (6)	0.096 (4)
Vietnamese	0.406 (46)	0.684 (8)	0.066 (41)	0.022 (45)	0.04 (24)	0.031 (38)
Sanumá	0.546 (42)	0.574 (37)	0.05 (45)	0.022 (45)	0.024 (45)	0.024 (45)
Yagua	0.563 (40)	0.524 (42)	0.266 (6)	0.042 (22)	0.04 (24)	0.033 (35)
Yaqui	0.748 (2)	0.579 (34)	0.14 (21)	0.086 (7)	0.036 (34)	0.052 (13)
Yoruba	0.612 (22)	0.665 (11)	0.064 (42)	0.031 (36)	0.037 (31)	0.04 (26)

Appendix C. Complexity Measures Using C_{WALS} Table A3. C_{WALS} complexity for the subset of languages shared with the Bibles corpus.

Language	C_{WALS}	H ₁	H ₃	TTR	TTR+H ₁	TTR+H ₃	TTR+H ₁ +H ₃
Amele	0.456 (9)	0.568 (17)	0.59 (13)	0.134 (13)	0.066 (17)	0.076 (16)	0.069 (18)
Apurinã	0.573 (5)	0.592 (15)	0.523 (17)	0.205 (8)	0.087 (12)	0.08 (14)	0.075 (16)
Basque	0.647 (4)	0.655 (8)	0.588 (14)	0.224 (7)	0.133 (5)	0.095 (9)	0.103 (7)
Eastern Oromo	0.487 (8)	0.652 (9)	0.573 (16)	0.196 (9)	0.111 (9)	0.08 (14)	0.088 (11)
Egyptian Arabic	0.563 (6)	0.725 (3)	0.748 (1)	0.31 (1)	0.5 (1)	1.0 (1)	0.6 (1)
English	0.329 (15)	0.703 (5)	0.667 (4)	0.082 (17)	0.09 (10)	0.095 (9)	0.115 (6)
German	0.397 (13)	0.588 (16)	0.663 (6)	0.136 (12)	0.071 (14)	0.111 (5)	0.088 (11)
Halh Mongolian	0.516 (7)	0.745 (2)	0.601 (11)	0.228 (5)	0.285 (3)	0.125 (4)	0.166 (4)
Hausa	0.322 (16)	0.609 (13)	0.613 (10)	0.098 (16)	0.069 (15)	0.076 (16)	0.076 (15)
Imbabura Quichua	0.662 (3)	0.599 (14)	0.492 (19)	0.25 (3)	0.117 (8)	0.09 (12)	0.083 (14)
Indonesian	0.336 (14)	0.661 (7)	0.598 (12)	0.115 (15)	0.09 (10)	0.074 (18)	0.088 (11)
Modern Greek	0.452 (11)	0.683 (6)	0.655 (8)	0.181 (10)	0.125 (6)	0.111 (5)	0.125 (5)
Plateau Malagasy	0.309 (17)	0.567 (18)	0.518 (18)	0.14 (11)	0.069 (15)	0.069 (19)	0.063 (19)
Russian	0.453 (10)	0.751 (1)	0.732 (2)	0.225 (6)	0.285 (3)	0.25 (2)	0.333 (2)
Spanish	0.44 (12)	0.647 (10)	0.656 (7)	0.133 (14)	0.083 (13)	0.095 (9)	0.096 (8)
Swahili	0.675 (2)	0.612 (11)	0.575 (15)	0.233 (4)	0.125 (6)	0.105 (7)	0.096 (8)
Turkish	0.775 (1)	0.705 (4)	0.629 (9)	0.297 (2)	0.333 (2)	0.181 (3)	0.2 (3)
Vietnamese	0.141 (19)	0.406 (19)	0.684 (3)	0.066 (18)	0.054 (19)	0.095 (9)	0.075 (16)
Yoruba	0.178 (18)	0.612 (11)	0.665 (5)	0.064 (19)	0.066 (17)	0.083 (13)	0.085 (13)

Appendix D. Complexity Measures Using MCC

Table A4. MCC complexity for the subset of languages shared with the JW300 corpus.

Language	MCC	H ₁	H ₃	TTR	TTR+H ₁	TTR+H ₃	TTR+H ₁ +H ₃
Bulgarian	96.0 (7)	0.56 (20)	0.68 (16)	0.091 (10)	0.016 (20)	0.018 (13)	0.016 (18)
Czech	195.0 (2)	0.668 (3)	0.777 (1)	0.125 (6)	0.061 (3)	0.048 (2)	0.065 (2)
Danish	15.0 (20)	0.617 (9)	0.695 (11)	0.063 (19)	0.023 (12)	0.017 (16)	0.021 (13)
Dutch	26.0 (19)	0.604 (13)	0.683 (15)	0.061 (20)	0.02 (18)	0.015 (19)	0.018 (16)
English	6.0 (21)	0.682 (2)	0.713 (7)	0.053 (21)	0.026 (9)	0.017 (16)	0.025 (10)
Estonian	110.0 (5)	0.623 (7)	0.663 (18)	0.155 (4)	0.056 (4)	0.022 (8)	0.027 (8)
Finnish	198.0 (1)	0.563 (19)	0.628 (20)	0.184 (1)	0.024 (10)	0.019 (11)	0.017 (17)
French	30.0 (18)	0.522 (21)	0.674 (17)	0.072 (16)	0.012 (21)	0.015 (19)	0.012 (21)
German	38.0 (16)	0.636 (6)	0.686 (14)	0.084 (12)	0.031 (8)	0.018 (13)	0.024 (11)
Hungarian	94.0 (8)	0.694 (1)	0.747 (4)	0.172 (2)	0.133 (1)	0.056 (1)	0.081 (1)
Italian	52.0 (13)	0.595 (14)	0.614 (21)	0.082 (13)	0.022 (14)	0.013 (21)	0.015 (20)
Latvian	81.0 (9)	0.582 (18)	0.745 (5)	0.123 (8)	0.022 (14)	0.036 (5)	0.027 (8)
Lithuanian	152.0 (3)	0.637 (4)	0.706 (8)	0.167 (3)	0.067 (2)	0.033 (6)	0.041 (5)
Modern Greek	50.0 (14)	0.594 (16)	0.753 (3)	0.09 (11)	0.022 (14)	0.03 (7)	0.027 (8)
Polish	112.0 (4)	0.617 (9)	0.745 (5)	0.152 (5)	0.047 (5)	0.047 (3)	0.045 (3)
Portuguese	77.0 (10)	0.595 (14)	0.697 (10)	0.075 (15)	0.02 (18)	0.019 (11)	0.02 (15)
Romanian	60.0 (12)	0.609 (12)	0.695 (11)	0.071 (17)	0.021 (16)	0.017 (16)	0.021 (13)
Slovak	40.0 (15)	0.614 (11)	0.767 (2)	0.124 (7)	0.036 (7)	0.043 (4)	0.042 (4)
Slovenian	100.0 (6)	0.637 (4)	0.69 (13)	0.111 (9)	0.041 (6)	0.022 (8)	0.029 (6)
Spanish	71.0 (11)	0.59 (17)	0.65 (19)	0.079 (14)	0.02 (18)	0.015 (19)	0.016 (18)
Swedish	35.0 (17)	0.621 (8)	0.706 (8)	0.066 (18)	0.024 (10)	0.019 (11)	0.023 (12)

Appendix E. Correlation Using Typological Classifications

For each language in the intersection set between the Bibles and JW300 corpora, we extracted its information about the feature 20A: “Fusion of Selected Inflectional Formatives” (WALS database). We focused on the languages classified as “concatenative” or “isolating”. For each corpus, we calculated the correlations within complexity measures for concatenative languages and the correlations within the isolating ones (Tables A5 and A6).

Table A5. Spearman’s correlation between complexity measures in concatenative and isolating languages (Bibles corpus).

		H ₁	H ₃	TTR
Concatenative	H ₁	1.0	0.233	0.618
	H ₃	-	1.0	−0.121
	TTR	-	-	1.0
Isolating	H ₁	1.0	−0.355	0.513
	H ₃	-	1.0	−0.178
	TTR	-	-	1.0

Table A6. Spearman’s correlation between complexity measures in concatenative and isolating languages (JW300 corpus).

		H ₁	H ₃	TTR
Concatenative	H ₁	1.0	−0.12	0.296
	H ₃	−12	1.0	−0.369
	TTR	-	-	1.0
Isolating	H ₁	1.0	−0.011	0.438
	H ₃	-	1.0	−0.741
	TTR	-	-	1.0

Appendix F. Correlation Using Average Word Length

We calculate the average word length per language in both corpora. This is formulated as the average of the number of characters per word. Tables A7 and A8 show the correlations of the average word length with the other measures for the Bibles and JW300 corpora, respectively.

Table A7. Spearman’s correlation between complexity measures and the average length per word in the Bibles corpus.

	H ₁	H ₃	TTR	TTR+H ₁	TTR+H ₃	TTR+H ₁ +H ₃
Average Word Length	0.354	−0.421	0.697	0.628	0.141	0.278

Table A8. Spearman’s correlation between complexity measures and the average length per word in the JW300 corpus.

	H ₁	H ₃	TTR	TTR+H ₁	TTR+H ₃	TTR+H ₁ +H ₃
Average Word Length	0.296	−0.359	0.735	0.606	0.265	0.315

References

1. Sampson, G.; Gil, D.; Trudgill, P. *Language Complexity as an Evolving Variable*; Oxford University Press: Oxford, UK, 2009; Volume 13.
2. Meinhardt, E.; Malouf, R.; Ackerman, F. *Morphology Gets More and More Enumeratively Complex, Unless It Doesn’t*; LSA Summer Institute: Lexington, KY, USA, 2017.
3. Miestamo, M.; Sinnemäki, K.; Karlsson, F. Grammatical complexity in a cross-linguistic perspective. *Lang. Complexity Typol. Contact Chang.* **2008**, 23–41. [\[CrossRef\]](#)
4. Simon, H.A. *The Architecture of Complexity*; MIT Press: Cambridge, MA, USA, 1996.
5. Sinnemäki, K. *Language Universals and Linguistic Complexity: Three Case Studies in Core Argument Marking*; University of Helsinki: Helsinki, Finland, 2011.
6. Baerman, M.; Brown, D.; Corbett, G.G. *Understanding and Measuring Morphological Complexity*; Oxford University Press: New York, NY, USA, 2015.
7. Haspelmath, M.; Sims, A.D. *Understanding Morphology*; Hodder Education: London, UK, 2010.
8. Montermini, F.; Bonami, O. Stem spaces and predictability in verbal inflection. *Lingue e Linguaggio* **2013**, 12, 171–190.
9. Ackerman, F.; Malouf, R. Morphological organization: The low conditional entropy conjecture. *Language* **2013**, 89, 429–464. [\[CrossRef\]](#)
10. Cotterell, R.; Kirov, C.; Hulden, M.; Eisner, J. On the complexity and typology of inflectional morphological systems. *Trans. Assoc. Comput. Linguist.* **2019**, 7, 327–342. [\[CrossRef\]](#)
11. Bentz, C.; Ruzsics, T.; Koplenig, A.; Samardzic, T. A comparison between morphological complexity measures: Typological data vs. language corpora. In Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (cl4lc), Osaka, Japan, 11 December 2016; pp. 142–153.
12. Blevins, J.P. The information-theoretic turn. *Psihologija* **2013**, 46, 355–375. [\[CrossRef\]](#)
13. Bonami, O.; Beniamine, S. Joint predictiveness in inflectional paradigms. *Word Struct.* **2016**, 9, 156–182. [\[CrossRef\]](#)
14. Kettunen, K. Can type-token ratio be used to show morphological complexity of languages? *J. Quant. Linguist.* **2014**, 21, 223–245. [\[CrossRef\]](#)
15. Mayer, T.; Cysouw, M. Creating a massively parallel bible corpus. *Oceania* **2014**, 135, 40.
16. Dryer, M.S.; Haspelmath, M. The World Atlas of Language Structures Online. 2013. Available online: <https://wals.info/> (accessed on 8 December 2019).
17. Agić, Ž.; Vulić, I. JW300: A wide-coverage parallel corpus for low-resource languages. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019.
18. Bybee, J. *Language, Usage and Cognition*; Cambridge University Press: Cambridge, UK, 2010.

19. Covington, M.A.; McFall, J.D. Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *J. Quant. Linguist.* **2010**, *17*, 94–100. [\[CrossRef\]](#)
20. Tweedie, F.J.; Baayen, R.H. How variable may a constant be? Measures of lexical richness in perspective. *Comput. Humanit.* **1998**, *32*, 323–352. [\[CrossRef\]](#)
21. Kelih, E. The type-token relationship in Slavic parallel texts. *Glottometrics* **2010**, *20*, 1–11.
22. Mayer, T.; Wälchli, B.; Rohrdantz, C.; Hund, M. From the extraction of continuous features in parallel texts to visual analytics of heterogeneous areal-typological datasets. *Lang. Process. Grammars. Role Funct. Oriented Comput. Model.* **2014**, 13–38. [\[CrossRef\]](#)
23. Gerz, D.; Vulić, I.; Ponti, E.M.; Reichart, R.; Korhonen, A. On the relation between linguistic typology and (limitations of) multilingual language modeling. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 316–327.
24. Baerman, M. Paradigmatic chaos in Nuer. *Language* **2012**, *88*, 467–494. [\[CrossRef\]](#)
25. Smit, P.; Virpioja, S.; Grönroos, S.A.; Kurimo, M. Morfessor 2.0: Toolkit for statistical morphological segmentation. In Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, 26–30 April 2014; pp. 21–24.
26. Sennrich, R.; Haddow, B.; Birch, A. Neural machine translation of rare words with subword units. *arXiv* **2015**, arXiv:1508.07909.
27. Baayen, R.H.; Chuang, Y.Y.; Blevins, J.P. Inflectional morphology with linear mappings. *Ment. Lex.* **2018**, *13*, 230–268. [\[CrossRef\]](#)
28. Vania, C.; Lopez, A. From characters to words to in between: Do we capture morphology? *arXiv* **2017**, arXiv:1704.08352.
29. Bengio, Y.; Ducharme, R.; Vincent, P.; Jauvin, C. A neural probabilistic language model. *J. Mach. Learn. Res.* **2003**, *3*, 1137–1155.
30. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
31. Freedman, A. Convergence Theorem for Finite Markov Chains. *Proc. REU* **2017**. Available online: <http://math.uchicago.edu/~may/REU2017/REUPapers/Freedman.pdf> (accessed on 24 December 2019).
32. Ding, C.; Aye, H.T.Z.; Pa, W.P.; Nwet, K.T.; Soe, K.M.; Utiyama, M.; Sumita, E. Towards Burmese (Myanmar) Morphological Analysis: Syllable-based Tokenization and Part-of-speech Tagging. *ACM Trans. Asian Low-Resour. Lang. Inf. Process. (TALLIP)* **2019**, *19*, 5. [\[CrossRef\]](#)
33. Spearman, C. The proof and measurement of association between two things. *Am. J. Psychol.* **1987**, *100*, 441–471. [\[CrossRef\]](#)
34. Cotterell, R.; Mielke, S.J.; Eisner, J.; Roark, B. Are all languages equally hard to language-model? *arXiv* **2018**, arXiv:1806.03743.
35. Kirov, C.; Sylak-Glassman, J.; Knowles, R.; Cotterell, R.; Post, M. A Rich Morphological Tagger for English: Exploring the Cross-Linguistic Tradeoff Between Morphology and Syntax. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, 3–7 April 2017; Short Papers; Association for Computational Linguistics: Valencia, Spain, 2017; Volume 2, pp. 112–117.
36. Blevins, J.P.; Milin, P.; Ramscar, M. The Zipfian paradigm cell filling problem. In *Perspectives on Morphological Organization*; Brill: Leiden, The Netherlands, 2017; pp. 139–158.
37. Mel'čuk, I.A. On suppletion. *Linguistics* **1976**, *14*, 45–90. [\[CrossRef\]](#)
38. Peters, A.M.; Menn, L. False Starts and Filler Syllables: Ways to Learn Grammatical Morphemes. *Language* **1993**, *69*, 742–777. [\[CrossRef\]](#)
39. Coupé, C.; Oh, Y.M.; Dediu, D.; Pellegrino, F. Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Sci. Adv.* **2019**, *5*, eaaw2594. [\[CrossRef\]](#)

