entropy

MDPI

Article

The Entropy of Words—Learnability and Expressivity across More than 1000 Languages

Christian Bentz [1,2,*], Dimitrios Alikaniotis [3], Michael Cysouw [4] and Ramon Ferrer-i-Cancho [5]

1 DFG Center for Advanced Studies, University of Tübingen, Rümelinstraße 23, D-72070 Tübingen, Germany
2 Department of General Linguistics, University of Tübingen, Wilhemstraße 19-23, D-72074 Tübingen, Germany
3 Department of Theoretical and Applied Linguistics, University of Cambridge, Cambridge CB3 9DP, UK; da352@cam.ac.uk
4 Forschungszentrum Deutscher Sprachatlas, Philipps-Universität Marburg, Pilgrimstein 16, D-35032 Marburg, Germany; cysouw@uni-marburg.de
5 Complexity and Quantitative Linguistics Lab, LARCA Research Group, Departament de Ciències de la Computació, Universitat Politècnica de Catalunya, 08034 Barcelona, Catalonia, Spain; rferrericancho@cs.upc.edu
* Correspondence: chris@christianbentz.de; Tel.: +49-7071-29-77506

Academic Editor: Gunnar Pruessner
Received: 26 April 2017; Accepted: 6 June 2017; Published: 14 June 2017

Abstract: The choice associated with words is a fundamental property of natural languages. It lies at the heart of quantitative linguistics, computational linguistics and language sciences more generally. Information theory gives us tools at hand to measure precisely the average amount of choice associated with words: the word entropy. Here, we use three parallel corpora, encompassing ca. 450 million words in 1916 texts and 1259 languages, to tackle some of the major conceptual and practical problems of word entropy estimation: dependence on text size, register, style and estimation method, as well as non-independence of words in co-text. We present two main findings: Firstly, word entropies display relatively narrow, unimodal distributions. There is no language in our sample with a unigram entropy of less than six bits/word. We argue that this is in line with information-theoretic models of communication. Languages are held in a narrow range by two fundamental pressures: word learnability and word expressivity, with a potential bias towards expressivity. Secondly, there is a strong linear relationship between unigram entropies and entropy rates. The entropy difference between words with and without co-textual information is narrowly distributed around ca. three bits/word. In other words, knowing the preceding text reduces the uncertainty of words by roughly the same amount across languages of the world.

Keywords: natural language entropy; entropy rate; unigram entropy; quantitative language typology

1. Introduction

Symbols are the building blocks of information. When concatenated to strings, they give rise to surprisal and uncertainty, as a consequence of choice. This is the fundamental concept underlying information encoding. Natural languages are communicative systems harnessing this information-encoding potential. Their fundamental building blocks are words. For any natural language, the average amount of information a word can carry is a basic property, an information-theoretic fingerprint that reflects its idiosyncrasies, and sets it apart from other languages across the world. Shannon [1] defined the entropy, or average information content, as a measure for the choice associated with symbols in strings. Since Shannon's [2] original proposal, many researchers have undertaken

great efforts to estimate the entropy of written English with the highest possible precision [3–6] and to broaden the account to other natural languages [7–11].

Entropic measures in general are relevant for a wide variety of linguistic and computational subfields. For example, several recent studies engage in establishing information-theoretic and corpus-based methods for linguistic typology, i.e., classifying and comparing languages according to their information encoding potential [10,12–16], and how this potential evolves over time [17–19]. Similar methods have been applied to compare and distinguish non-linguistic sequences from written language [20,21], though it is controversial whether this helps with more fine-grained distinctions between symbolic systems and written language [22,23].

In the context of quantitative linguistics, entropic measures are used to understand laws in natural languages, such as the relationship between word frequency, predictability and the length of words [24–27], or the trade-off between word structure and sentence structure [10,13,28]. Information theory can further help to understand the complexities involved when building words from the smallest meaningful units, i.e., morphemes [29,30].

Beyond morphemes and word forms, the surprisal of words in co-text is argued to be related to syntactic expectations [31]. We use the term "co-text" here to refer to the purely linguistic, textual environment of a word type. This contrasts with the "context", which can also include any non-linguistic environment associated with the word type, such as gestures, facial expression, or any multimodal perception in general. An example of a co-text effect is the usage of complementizers such as "that", which might serve to smooth over maxima in the information density of sentences [32]. This has become known as the Uniform Information Density (UID) hypothesis. It was introduced in the 1980s by August and Gertraud Fenk [33], and developed in a series of articles (see [34] and the references therein). For a critical review of some problematic aspects of the UID hypothesis see [35,36].

In optimization models of communication, word entropy is a measure of cognitive cost [37]. These models shed light on the origins of Zipf's law for word frequencies [38,39] and a vocabulary learning bias [40]. With respect to the UID hypothesis, these models have the virtue of defining explicitly a cost function and linking the statistical regularities with the minimization of that cost function.

Finally, in natural language processing, entropy (and related measures) have been widely applied to tackle problems relating to machine translation [41,42], distributional semantics [43–46], information retrieval [47–49] and multi-word expressions [50,51].

All of these accounts crucially hinge upon estimating the probability and uncertainty associated with different levels of information encoding in natural languages. Here, we focus on the word entropy of a given text and language. There are two central questions associated with this measure: (1) What is the text size (in number of tokens) at which word entropies reach stable values? (2) How much systematic difference do we find across different texts and languages? The first question is related to the problem of data sparsity. The minimum text size at which results are still reliable is a lower bound on any word entropy analysis. The second question relates to the diversity that we find across languages of the world. From the perspective of linguistic typology, we want to understand and explain this diversity.

In this study, we use state-of-the-art methods to estimate both unigram entropies [52] (for a definition see Section 3.2.6), and the entropy rate per word according to Gao et al. [5] (Section 3.2.7). Unigram entropy is the average information content of words assuming that they are independent of the co-text. The entropy rate can be seen, under certain conditions (see Section 3.2.7), as the average information content of words assuming that they depend on a sufficiently long, preceding co-text. For both measures we first establish stabilization points for big parallel texts in 21 and 32 languages respectively. Based on these stabilization points, we select texts with sufficiently large token counts from a massively parallel corpus and estimate word entropies across more than 1000 texts and languages. Our analyses illustrate two major points:

1. Across languages of the world, unigram entropies display a unimodal distribution around a mean of ca. nine bits/word, with a standard deviation of ca. one bit/word. Entropy rates have a lower

mean of ca. six bits/word, with a standard deviation of ca. one bit/word. Hence, there seem to be strong pressures keeping the mass of languages in a relatively narrow entropy range. This is particularly salient for the difference between unigram entropy and entropy rate (Section 5.2).

2. There is a strong positive linear relationship between unigram entropies and entropy rates ($r = 0.96, p < 0.0001$). To our knowledge, this has not been reported before. We formulate a simple linear model that predicts the entropy rate of a text $\hat{h}(T)$ from the unigram entropy $\hat{H}_1(T)$ of the same text: $\hat{h}(T) = k_1 + k_2\,\hat{H}_1(T)$, where $k_1 = -1.12$ and $k_2 = 0.78$ (Section 5.3). The implication of this relationship is that uncertainty-reduction by co-textual information is approximately linear across languages of the world.

Clearly, higher or lower word entropy does not equate to "better" or "worse" communication systems in a very general sense. In any scenario where languages are used to communicate, a multitude of contextual cues (gestures, facial expression, general world knowledge, etc.) are integrated to give rise to meaning. Such contextual cues are typically not available in written language. However, in information-theoretic models of communication, the entropy of an inventory of symbols is the upper bound on mutual information between symbols and meanings (see Section 6.1). Thus, the entropy of words can be seen as the upper bound on expressivity in an information-theoretic sense.

We further argue in the discussion section that word entropies across languages of the world reflect the trade-off between two basic pressures on natural communication systems: word learnability vs. word expressivity. Hence, information-theoretic properties like entropy are not only relevant for computational linguistics and quantitative linguistics, but constitute a basic property of human languages. Understanding and modelling the differences and similarities in the information that words can carry is an undertaking at the heart of language sciences more generally.

2. Data

To estimate the entropy of words, we first need a comparable sample of texts across many languages. Ideally, the texts should have the same content, as differences in registers and style can interfere with the range of word forms used, and hence the entropy of words [53]. To control for constant content across languages, we use three sets of parallel texts: (1) the European Parliament Corpus (EPC) [54], (2) the Parallel Bible Corpus (PBC) [55] (last accessed on 2 June 2016), and the Universal Declaration of Human Rights (http://www.unicode.org/udhr/). Details about the corpora can be seen in Table 1. The general advantage of the EPC is that it is big in terms of numbers of word tokens per language (ca. 30 M) (though we only use around one M tokens for each language, since this is enough for the stabilization analyses), whereas the PBC and the UDHR are smaller (ca. 290 K and 1.3 K word tokens per language). However, the PBC and UDHR are massively parallel in terms of encompassing more than 1000 and more than 300 languages, respectively. These are numbers of texts and languages that we actually used for our analyses. The raw corpora are bigger. However, some texts and languages had to be excluded due to their small size, or due to pre-processing errors (see Section 3.1).

Table 1. Information on the parallel corpora used.

Corpus	Register	Size *	Mean Size *	Texts	Lang
EPC	Political	ca. 21 M	ca. 1 M	21	21
PBC	Religious	ca. 430 M	ca. 290 K	1525	1137
UDHR	Legal	ca. 500 K	ca. 1.3 K	370	341
	Total:	**ca. 450 M**		**1916**	**1259**

* In number of tokens.

3. Theory

3.1. Word Types and Tokens

The basic information encoding unit chosen in this study is the word. In contrast, most earlier studies on the entropy of languages [2,4,6] chose characters instead. A word type is here defined as a unique string of alphanumeric UTF-8 characters delimited by white spaces. All letters are converted to lower case and punctuation is removed. A word token is then any reoccurrence of a word type. For example, the pre-processed first verse of the Book of Genesis in English reads:

```
in the beginning god created the heavens and the earth and the earth was
waste and empty [...]
```

The set of word types (in lower case) for this sentence is:

$$\mathcal{V} = \{in, the, beginning, god, created, heavens, and, earth, was, waste, empty\}. \quad (1)$$

Hence, the number of word types in this sentence is 11, but the number of word tokens is 17, since *the, and and earth* occur several times. Note that some scripts, e.g., those of Mandarin Chinese (cmn) and Khmer (khm), delimit phrases and sentences by white spaces, rather than words. Such scripts have to be excluded for the simple word processing we propose. However, they constitute a negligible proportion of our sample (≈0.01%). In fact, ca. 90% of the texts are written in Latin-based script. For more details and caveats of text pre-processing, see Appendix A.

Though definitions of "word-hood" based on orthography are common in corpus and computational linguistic studies, they are somewhat controversial within the field of linguistic typology. Haspelmath [56] and Wray [57], for instance, point out that there is a range of orthographic, phonetic and distributional definitions for the concept "word", which can yield different results across different cases. However, a recent study on compression properties of parallel texts vindicates the usage of orthographic words as information encoding units, as it shows that these are optimal-sized for describing the regularities in languages [58].

Hence, we suggest to start with the orthographic word definition based on non-alphanumeric characters in written texts; not the least because it is a computationally feasible strategy across many hundreds of languages. Our results might then be tested against more fine-grained definitions, as long as these can be systematically applied to language production data.

3.2. Entropy Estimation

3.2.1. General Conditions

The conditions (e.g., the sample size) under which the entropy of a source can be estimated reliably are a living field of research ([59,60] and the references therein). Proper estimation of entropy requires certain conditions to be met. Stationarity and ergodicity are typically named as such. Stationarity means that the statistical properties of blocks of words do not depend on their position in the text sequence. Ergodicity means that statistical properties of a sufficiently long text sequence match the average properties of the ensemble of all possible text sequences. (e.g., [59,61]). Another condition is finiteness (i.e., a corpus has a finite set of word types). While stationarity and ergodicity tend to be presented and discussed together in research articles, finiteness is normally stated separately, typically as part of the initial setting of the problem (e.g., [59,61]). A fourth usual condition is the assumption that types are independent and identically distributed (i.i.d.) [59,61,62]. Hence, for proper entropy estimation the typical requirements are either (a) finiteness, stationarity and ergodicity, or (b) only finiteness and i.i.d., as stationarity and ergodicity follow trivially from the i.i.d. assumption.

3.2.2. Word Entropy Estimation

A crucial pre-requisite for entropy estimation is the approximation of the probabilities of word types. In a text, each word type w_i has a token frequency $f_i = freq(w_i)$. Take the first verse of the English Bible again.

```
in the beginning god created the heavens and the earth and the earth was
waste and empty [...]
```

In this example, the word type *the* occurs four times, *and* occurs three times, etc. As a simple approximation, $p(w_i)$ can be estimated via the so-called maximum likelihood method [63]:

$$\hat{p}(w_i) = \frac{f_i}{\sum_{j=1}^{V} f_j}, \tag{2}$$

where the denominator is the overall number of word tokens, i.e., the sum of type frequencies over an empirical (i.e., finite) vocabulary of size $V = |\mathcal{V}|$. We thus have probabilities of $\hat{p}(the) = \frac{4}{17}$, $\hat{p}(and) = \frac{3}{17}$, etc.

Assume a text is a random variable T created by a process of drawing (with replacement) and concatenating tokens from a set (or vocabulary) of word types $\mathcal{V} = \{w_1, w_2, ..., w_W\}$, where W is the (potentially infinite) theoretical vocabulary size, and a probability mass function $p(w) = Pr\{T = w\}$ for $w \in \mathcal{V}$. Given these definitions, the theoretical entropy of T can be calculated as [64]:

$$H(T) = -\sum_{i=1}^{W} p(w_i) \log_2 p(w_i). \tag{3}$$

In this case, $H(T)$ can be seen as the average information content of word types. A crucial step towards estimating $H(T)$ is to reliably approximate the probabilities of word types $p(w_i)$.

The simplest way to estimate $H(T)$ is the maximum likelihood estimator, also called the plug-in estimator, which is obtained replacing $p(w)$ by $\hat{p}(w)$ (Equation (2)). For instance, our example "corpus" yields:

$$\hat{H}(T) = -\left(\frac{4}{17}\log_2\left(\frac{4}{17}\right) + \frac{3}{17}\log_2\left(\frac{3}{17}\right) + \ldots + \frac{1}{17}\log_2\left(\frac{1}{17}\right)\right) \approx 3.2\ bits/word. \tag{4}$$

Notice an important detail: we have assumed that $p(w_i) = \hat{p}(w_i) = 0$ for all the types that have not appeared in our sample. Although our sample did not contain "never", the probability of "never" in an English text is not zero. We have also assumed that the relative frequency of the words that have appeared in our sample is a good estimation of their true probability. Given the small size of our sample, this is unlikely to be the case. The bottom line is that for small text sizes we will underestimate the word entropy using the maximum likelihood approach [62,65]. A range of more advanced entropy estimators have been proposed to overcome this limitation [6,52,61,63]. These are outlined in more detail in Appendix B, and tested with our parallel corpus data.

However, even with advanced entropy estimators, there are two fundamental problems: (1) the set of word types in natural languages is non-finite, (2) tokens are not i.i.d., since they exhibit short- and long-range correlations. These issues are discussed in the following.

3.2.3. Problem 1: The Infinite Productive Potential of Languages

The first problem relates to the productive potential of languages. In fact, it is downright impossible to capture the actual repertoire of word types in a language. Even if we captured the whole set of linguistic interactions of a speaker population in a massive corpus, we would still not capture the productive potential of the language beyond the finite set of linguistic interactions.

Theoretically speaking, it is always possible to expand the vocabulary of a language by compounding (recombining word types to form new word types), affixation (adding affixes to existing words), or by creating neologisms. This can be seen in parallel to Chomsky's ([66], p. 8) reappraisal of Humboldt's "make infinite use of finite means" in syntax. The practical consequence is that even for massive corpora like the British National Corpus, vocabulary growth is apparently not coming to a halt [67–69]. Thus, we never actually sample the whole set of word types of a language. However, current entropy estimators are developed only for a finite "alphabet" (i.e., repertoire of word types in our case). Note that the concentration of tokens on a core vocabulary [68,70] potentially alleviates this problem. Still, word entropy estimation is a harder problem than entropy estimation for characters or phoneme types, since the latter have a repertoire that is finite and usually small.

3.2.4. Problem 2: Short- and Long-Range Correlations between Words

In natural languages, we find co-occurrence patterns illustrating that word types in a text sequence are not independent events. This is the case for collocations, namely blocks of consecutive words that behave as a whole. Examples include place names such as "New York" or fixed expressions such as "kith and kin", but also many others that are less obvious. They can be detected with methods determining if some consecutive words co-occur with a frequency that is higher than expected by chance [71]. Collocations are examples of short-range correlations in text sequences. Text sequences also exhibit long-range correlations, i.e., correlations between types that are further apart in the text [8,72].

3.2.5. Our Perspective

We are fully aware of both Problems 1 and 2 when estimating the word type entropy of real texts, and the languages represented by them. However, our question is not so much: what is the exact entropy of a text or language? but rather: How precisely do we have to approximate it to make a meaningful cross-linguistic comparison possible? To address this practical question, we need to establish the number of word tokens at which estimated entropies reach stable values, given a threshold of our choice. Furthermore, we incorporate entropic measures which are less demanding in terms of assumptions. Namely, we include h, the entropy rate of a source, as it does not rely on the i.i.d. assumption. We elaborate on h in the next two subsections.

3.2.6. n-Gram Entropies

When estimating entropy according to Equation (3), we assume unigrams, i.e., single, independent word tokens, as "blocks" of information encoding. As noted above, this assumption is generally not met for natural languages. To incorporate dependencies between words, we could use bigrams, trigrams, or more generally n-grams of any size, and thus capture short- and long-range correlations by increasing "block" sizes to 2, 3, n. This yields what are variously called n-gram or block entropies [6], defined as:

$$H_n(T) = -\sum_{i=1}^{W^n} p({g_i}^{(n)}) \times \log_2 p({g_i}^{(n)}), \tag{5}$$

where n is the block size, $p({g_i}^{(n)})$ is the probability of an n-gram g_i of block size n, and W^n is the potentially infinite size of the "alphabet" of n-grams. However, since the number of different n-grams grows exponentially with n, very big corpora are needed to get reliable estimates. Schürmann and Grassberger [6] use an English corpus of 70 M words, and assert that entropy estimation beyond a block size of five characters (not words) is already unreliable. Our strategy is to stick with block sizes of one, i.e., unigram entropies. However, we implement a more parsimonious approach to take into account long-range correlations between words along the lines of earlier studies by Montemurro and Zanette [8,9].

3.2.7. Entropy Rate

Instead of calculating $H_n(T)$ with ever increasing block sizes n, we use an approach focusing on a particular feature of the entropy growth curve: the so-called entropy rate, or per-symbol entropy [5]. In general, it is defined as the rate at which the word entropy grows as the number of word tokens N increases ([73], p. 74), i.e.,

$$h(T) = \lim_{N\to\infty} \frac{1}{N} H_n(T) = \lim_{N\to\infty} \frac{1}{N} H(t_1, t_2, ..., t_N), \tag{6}$$

where $t_1, t_2, ..., t_N$ is a block of consecutive tokens of length N. Given stationarity, this is equivalent to ([73], p. 75)

$$h(T) = \lim_{N\to\infty} H(t_N | t_1, t_2, ..., t_{N-1}). \tag{7}$$

In other words, as the number of tokens N approaches infinity, the entropy rate $h(T)$ reflects the average information content of a token t_N conditioned on all preceding tokens. Therefore, $h(T)$ accounts for all statistical dependencies between tokens [61]. Note that in the limit, i.e., as block size n approaches infinity, the block entropy per token converges to the entropy rate. Furthermore, for an independent and identically distributed (i.i.d) random variable, the block entropy of block size one, i.e., $H_1(T)$, is identical to the entropy rate [61]. However, as pointed out above, in natural languages words are not independently distributed.

Kontoyiannis et al. [4] and Gao et al. [5] apply findings on optimal compression by Ziv and Lempel [74,75] to estimate the entropy rate. More precisely, Gao et al. [5] show that entropy rate estimation based on the so-called increasing window estimator, or LZ78 estimator [73], is efficient in terms of convergence. The conditions for this estimator are stationarity and ergodicity of the process that generates a text T. Again, its consistency has only been proven for a finite alphabet (i.e., finite set of word types in our case).

Applied to the problem of estimating word entropies, the method works as follows: for any given word token t_i find the longest match-length L_i for which the token string $s_i^{i+l-1} = (t_i, t_{i+1}, ..., t_{i+l-1})$ matches a preceding token string of the same length in $(t_1, ..., t_{i-1})$. Formally, we define L_i as:

$$L_i = 1 + \max\{0 \le l \le i : s_i^{i+l-1} = s_j^{j+l-1} \text{ for some } 0 \le j \le i-1\}. \tag{8}$$

This is an adaptation of Gao et al.'s [5] match-length definition (note that Gao et al. [5] give a more general definition that also holds for the so-called sliding window, or LZ77 estimator). To illustrate this, take the example from above again:

$\texttt{in}_1$ $\texttt{the}_2$ $\texttt{beginning}_3$ $\texttt{god}_4$ $\texttt{created}_5$ $\texttt{the}_6$ $\texttt{heavens}_7$ $\underline{\texttt{and}}_8$ $\underline{\texttt{the}}_9$ $\underline{\texttt{earth}}_{10}$ $\texttt{and}_{11}$ $\underline{\texttt{the}}_{12}$ $\underline{\texttt{earth}}_{13}$ $\texttt{was}_{14}$ $\texttt{waste}_{15}$ $\texttt{and}_{16}$ $\texttt{empty}_{17}$ [...]

For the word token *beginning*, in position $i = 3$, there is no match in the preceding token string (in the). Hence, the match-length is $L_3 = 0(+1) = 1$. In contrast, if we look at *and* in position $i = 11$, then the longest matching token string is *and the earth*. Hence, the match-length is $L_{11} = 3(+1) = 4$.

Note that the average match-lengths across all word tokens reflect the redundancy in the token string, which is the inverse of unpredictability or choice. Based on this connection, Gao et al. [5], Equation (6), show that the entropy rate of a text can be approximated as:

$$\hat{h}(T) = \frac{1}{N} \sum_{i=2}^{N} \frac{\log_2 i}{L_i}, \tag{9}$$

where N is the overall number of tokens, and i is the position in the string. Here, we approximate the entropy rate $h(T)$ based on a text T as $\hat{h}(T)$ given in Equation (9).

4. Methods

4.1. Entropy Estimation Software

We estimate unigram entropies, i.e., entropies for block sizes of one ($H_1(T)$), with nine different estimation methods, using the R package entropy [76], and the Python implementation of the Nemenman–Shafee–Bialek (NSB) [52] estimator (https://gist.github.com/shhong/1021654/). For further analyses, we especially focus on the NSB estimator, as it has been shown to have a faster convergence rate compared to other block entropy estimators [63]. Moreover, we implemented the entropy rate estimator $\hat{h}(T)$ as in Gao et al.'s [5] proposal in both Python and R, available on github (https://github.com/dimalik/Hrate/). This was inspired by an earlier implementation by Montemurro and Zanette [8] of Kontoyiannis et al.'s [4] estimator.

4.2. Stabilization Criterion

When estimating entropy with increasingly longer "prefixes" (i.e., runs of text preceding a given token) a fundamental milestone is convergence, namely, the prefix length at which the true value is reached with an error that can be neglected. However, the "true" entropy of a productive system like natural language is not known. For this reason, we replace convergence by another important milestone: the number of tokens at which the next 10 estimations of entropy have a SD (standard deviation) that is sufficiently small, e.g., below a certain threshold. If N is the number of tokens, SD is defined as the standard deviation that is calculated over entropies obtained with prefixes of lengths:

$$N, N + 1K, N + 2K, ..., N + 10K. \tag{10}$$

K represents the number 1000 here. We say that entropies have stabilized when SD $< \alpha$, where α is the threshold. Notice that this is a local stabilization criterion. Here, we choose N to run from $1K$ to $90K$. We thus get 90 SD values per language. The threshold is $\alpha = 0.1$. The same methodology is used to determine when the entropy rate has stabilized.

Two earlier studies, Montemurro and Zanette [8] and Koplenig et al. [10], use a more coarse-grained stabilization criterion. In [8] entropies are estimated for two halves of each text, and then compared to the entropy of the full text. Only texts with a maximum discrepancy of 10% are included for further analyses. Similarly, [10] compares entropies for the first 50% of the data and for the full data. Again, only texts with a discrepancy of less then 10% are included. In contrast, [11] establishes the convergence properties of different off-the-shelf compressors by estimating the encoding rate with growing text sizes. This has the advantage of giving a more fine-grained impression of convergence properties. Our assessment of entropy stabilization follows a similar rationale, though with words as information encoding units, rather than characters, and with our own implementation of Gao et al.'s entropy estimator rather than off-the-shelf compressors.

4.3. Corpus Samples

To assess the minimum text sizes at which both unigram entropies and entropy rates stabilize, we use 21 languages of the European Parliament Corpus (EPC), as well as a subset of 32 languages from different language families of the Parallel Bible Corpus (Section 5.1).

For estimating unigram entropies and entropy rates across languages of the world, we use the full PBC. Based on our stabilization analyses, we choose a conservative cut-off point: only texts with at least 50 K tokens are included. Of these, in turn, we take the first 50 K tokens for estimation. This criterion reduces the original PBC sample of 1525 texts and 1137 languages (ISO 639-3 codes) to 1499 texts and 1115 languages. This sample is used in Section 5.2 and Section 5.3.

5. Results

5.1. Entropy Stabilization throughout the Text Sequence

For both unigram entropies and entropy rates, the stabilization criterion, i.e., SD < 0.1, is met at 50 K tokens. This is the case for the 21 languages of the European Parliament Corpus (EPC) (Appendix C), as well as the 32 languages of the Parallel Bible Corpus (Appendix D). The UDHR texts are generally too small (ca. 1300 tokens on average) for stabilization.

Additionally, Appendix E illustrates that there are generally strong correlations between different entropy estimation methods, despite differences in stabilization properties. Appendix G gives Pearson correlations between unigram entropies of the PBC, EPC, and the UDHR. These are also strong (around 0.8), despite the differences in registers and styles, as well as text sizes. In sum, our analyses converge to show that entropy rankings of languages are stable at ca. 50 K tokens.

5.2. Word Entropies across More than 1000 Languages

Figure 1a shows a density plot of the estimated entropic values across all 1499 texts and 1115 languages of the PBC sample. Both unigram entropies and entropy rates show a unimodal distribution that is skewed to the right. Unigram entropies are distributed around a mean of 9.14 ($SD = 1.12$), whereas entropy rates have a lower mean of 5.97 ($SD = 0.91$). In Figure 1b, it is clearly visible that the difference between unigram entropies and entropy rates has a narrower distribution: it is distributed around a mean of ca. 3.17 bits/word with a standard deviation of 0.36.

To visually illustrate the diversity of entropies across languages of the world, Figure 2 gives a map with unigram entropies, entropy rates and the difference between them. The range of entropy values in bits per word is indicated by a colour scale from blue (low entropy) to red (high entropy). As can be seen in the upper map, there are high and low entropy areas across the world, further discussed below. A similar pattern holds for entropy rates (middle panel), though they are generally lower. The difference between unigram entropies and entropy rates, on the other hand, is even more narrowly distributed (as seen also in Figure 1b) and the colours are almost indistinguishable.

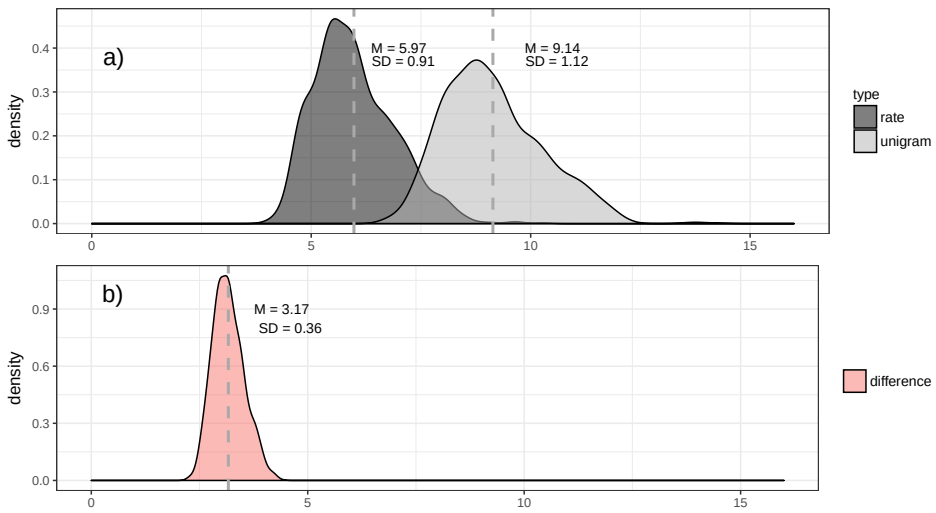

Figure 1. The distribution of entropic measures in bits. (**a**) Probability density of unigram entropies (light grey) and entropy rates (dark grey) across texts of the PBC (using 50K tokens). M and SD are, respectively, the mean and the standard deviation of the values. A vertical dashed line indicates the mean M. (**b**) The same for the difference between unigram entropies and entropy rates.

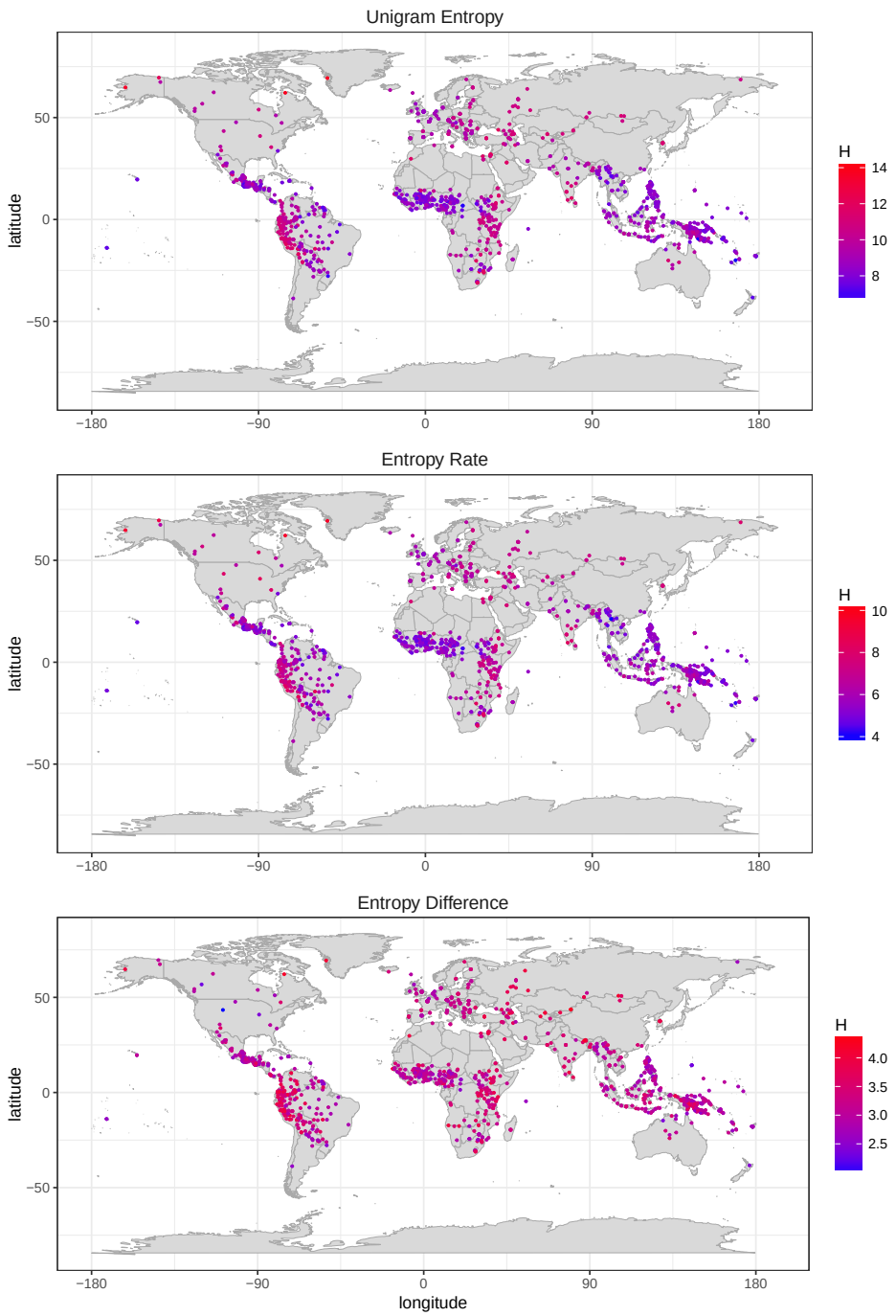

Figure 2. World maps with unigram entropies (upper panel), entropy rates (middle panel) and the difference between them (lower panel), across texts of the PBC (using 50K tokens), amounting to 1499 texts and 1115 languages.

5.3. Correlation between Unigram Entropy and Entropy Rate

The density plot for unigram entropies in Figure 1 looks like a slightly wider version of the entropy rate density plot, just shifted to the right of the scale. This suggests that unigram entropies and entropy rates might be correlated. Figure 3 confirms this by plotting unigram entropies on the x-axis versus entropy rates on the y-axis for each given text. The linearity of the relationship tends to increase as text length increases (from 30 K tokens onwards) as shown in Figure 4. This holds for all entropy estimators since they are strongly correlated (Appendix E).

In Appendix F we also give correlations between unigram entropy values for all nine estimation methods and entropy rates. The Pearson correlations are strong (between $r = 0.95$ and $r = 0.99$, with $p < 0.0001$). Hence, despite the conceptual difference between unigram entropies and entropy rates, there is a strong underlying connection between them. Namely, a linear model fitted (with the package lme4 in R [77]) through the points in Figure 3 (left panel) can be specified as:

$$\hat{h}(T) = -1.12 + 0.78\,\hat{H}_1(T). \tag{11}$$

The intercept of -1.12 is significantly smaller than zero ($p < 0.0001$), and the linear coefficient of 0.78 is significantly smaller than one ($p < 0.0001$). Via Equation (11), we can convert unigram entropies into entropy rates with a variance of 0.035. Hence, the observed difference of ca. 3.17 bits/word between unigram entropies and entropy rates (see Figure 1) is due to the interaction of two constants: First, there is a fixed difference of ca. 1.12 bit/word between the two estimators and a variable reduction of ca. 22% when converting unigram entropies into entropy rates (22% of the mean 9.14 bit/word for unigram entropies amounts to ca. 2.0 bit/word on average). More generally, for natural languages word entropy rates $\hat{h}(T)$ and unigram entropies $\hat{H}_1(T)$ are linked by a linear relationship:

$$\hat{h}(T) = k_1 + k_2\,\hat{H}_1(T), \tag{12}$$

where k_1 and k_2 are constants. The exact meaning and implications of these constants are topics for future research.

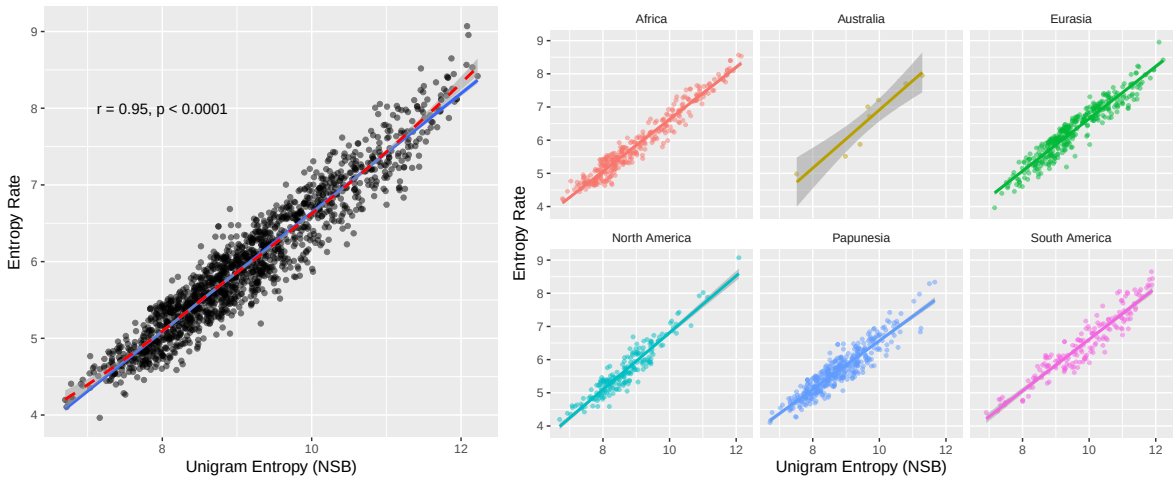

Figure 3. Linear relationship between unigram entropies approximated with the NSB estimator (x-axis) and entropy rates (y-axis) for 1495 PBC texts (50K tokens) across 1112 languages. Four texts (Ancient Hebrew (hbo), Eastern Canadian Inuktitut (ike), Kalaallisut (kal), and Northwest Alaska Inupiatun (esk)) were excluded here, since they have extremely high values of more than 13 bits/word. In the left panel, a linear regression model is given as blue line, a local regression smoother is given as red dashed line. The Pearson correlation coefficient is $r = 0.95$. In the right panels, plots are faceted by macro areas across the world. Macro areas are taken from Glottolog 2.7 [78]. Linear regression models are given as coloured lines with 95% confidence intervals.

The right panels of Figure 3 plot the same relationship between unigram entropies and entropy rates faceted by geographic macro areas taken from Glottolog 2.7 [78]. These macro areas are generally considered relevant from a linguistic typology point of view. It is apparent from this plot that the linear relationship extrapolates across different areas of the world.

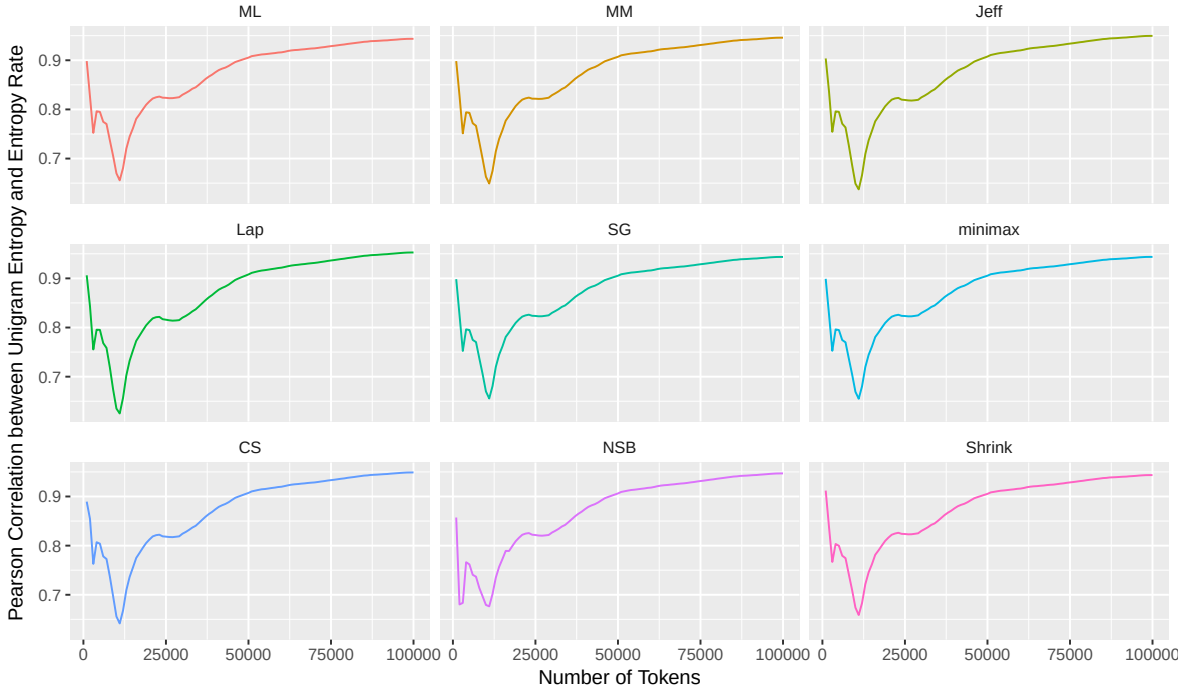

Figure 4. Pearson correlation between unigram entropy and entropy rate (y-axis) as a function of text length (x-axis). Each correlation is calculated over the 21 languages of the EPC corpus. All nine unigram entropy estimators are considered.

6. Discussion

6.1. Entropy Diversity across Languages of the World

In Section 5.2, we estimated word entropies for a sample of more than 1000 languages of the PBC. We find that unigram entropies cluster around a mean value of about nine bits/word, while entropy rates are generally lower, and fall closer to a mean of six bits/word (Figure 1). This is to be expected, since the former do not take co-textual information into account, whereas the latter do. To see this, remember that under stationarity, the entropy rate can be defined as a word entropy conditioned on a sufficiently large number of previous tokens (Equation (7)), while the unigram entropy is not conditioned on the co-text. As conditioning reduces entropy [73], it is not surprising that entropy rates tend to fall below unigram entropies.

It is more surprising that given the wide range of potential entropies, from zero to ca. 14, most natural languages fall on a relatively narrow spectrum. It is non-trivial to find an upper limit for the maximum word entropy of natural languages. In theory, it could be infinite given that the range of word types is potentially infinite. However, in practice the highest entropy languages range only up to ca. 14 bits/word. Unigram entropies mainly fall in the range between seven to 12 bits/word, and entropy rates in the range between four to nine bits/word. Thus, each only covers around 40% of the scale. The distributions are also skewed to the right, and seem to differ from the Gaussian, and therefore symmetric, distribution that is expected for the plug-in estimator under a two-fold null hypothesis: (1) that the true entropy is the same for all languages, and that (2) besides the bias in the entropy estimation, there is no additional bias constraining the distribution of entropy [62]. It needs to be clarified in further studies where this right-skewness stems from.

Overall, the distributions suggest that there are pressures at play which keep word entropies in a relatively narrow range. We argue that this observation is related to the trade-off between the learnability and expressivity of communication systems. A (hypothetical) language with maximum word entropy would have a vast (potentially infinite) number of word forms of equal probability, and

would be hard (or impossible) to learn. A language with minimum word entropy, on the other hand, would repeat the same word forms over and over again, and lack expressivity. Natural languages fall in a relatively narrow range between these extremes. This is in line with evidence from iterated learning experiments and computational simulations. For instance, Kirby et al. [79] illustrate how artificial languages collapse into states with underspecified word/meaning mappings, i.e., low word entropy states, if only pressure for learnability is given. On the other hand, when only pressure for expressivity is given, then so-called holistic strategies evolve with a separate word form for each meaning, i.e., a high word entropy state. However, if both pressures for learnability and expressivity interact then combinatoriality emerges as a coding strategy, which keeps unigram word entropies in the middle ground.

This trade-off is also obvious in optimization models of communication, which are based on two major principles: entropy minimization and maximization of mutual information between meanings and word forms [40]. Zipf's law for word frequencies, for instance, emerges in a critical balance between these two forces [39]. In these models, entropy minimization is linked with learnability: fewer word forms are easier to learn (see [80] for other cognitive costs associated with entropy). Whereas mutual information maximization is linked with expressivity via the form/meaning mappings available in a communication system. Note that a fundamental property in information theoretic models of communication is that MI, the mutual information between word forms and meanings, cannot exceed the entropy, i.e., [37]

$$MI \leq H \tag{13}$$

The lower the entropy, the lower the potential for expressivity. Hence, the entropy of words is an upper bound on the expressivity of words. This may also shed light on the right-skewness of the unigram entropy distribution in Figure 1. Displacing the distribution to the left or skewing it towards low values would compromise expressivity. In contrast, skewing it towards the right increases the potential for expressivity according to Equation (13), though this comes with a learnability cost.

Further support for pressure to increase entropy to warrant sufficient expressivity (Equation (13)) comes from the fact that there is no language with less than six bits/word unigram entropy for neither the PBC nor the EPC. There is only one language in the UDHR that has a slightly lower unigram entropy, a Zapotecan language of Meso-America (zam). It has a fairly small corpus size (1067 tokens), and the same language has more than eight bits/word in the PBC. Thus, there is no language which is more than three SDs below the mean of unigram entropies. On the other hand, there are several languages that are more than four SDs higher than the mean, i.e., around or beyond 13 bits/word.

Despite the fact that natural languages do not populate the whole range of possible word entropies, there can still be remarkable differences. Some of the languages at the low-entropy end are Tok Pisin, Bislama (Creole languages), and Sango (Atlantic-Congo language of Western Africa). These have unigram entropies of around 6.7 bits/word. Languages to the high-end include Greenlandic Inuktitut and Ancient Hebrew, with unigram entropies around 13 bits/word. Note that this is not to say that Greenlandic Inuktitut or Ancient Hebrew are "better" or "worse" communication systems than Creole languages or Sango. Such an assessment is misleading for two reasons: First, information encoding happens in different linguistic (and non-linguistic) dimensions, not just at the word level. We are only just starting to understand the interactions between these levels from an information-theoretic perspective [10]. Second, if we assume that natural languages are used for communication, then both learnability and expressivity of words are equally desirable features. Any combination of the two arises in the evolution of languages due to adaptive pressures. There is nothing "better" or "worse" about learnability or expressivity per se.

On a global scale, there seem to be high and low entropy areas. For example, languages in the Andean region of South America all have high unigram entropies (bright red in Figure 2). This is most likely due to their high morphological complexity, resulting in a wide range of word types, which were shown to correlate with word entropies [81]. Further areas of generally high entropies include Northern Eurasia, Eastern Africa, and North America. In contrast, Meso-America, Sub-Saharan

Africa und South-East Asia are areas of relatively low word entropies (purple and blue in Figure 2). Testing these global patterns for statistical significance is an immediate next step in our research. Some preliminary results for unigram entropies and their relationship with latitude can be found in [14].

We are just beginning to understand the driving forces involved when languages develop extremely high or low word entropies. For example, Bentz et al. [12] as well as Bentz and Berdicevskis [18] argue that specific learning pressures reduce word entropy over time. As a consequence, different scenarios of language learning, transmission and contact might lead to global patterns of low and high entropy areas [14].

6.2. Correlation between Unigram Entropies and Entropy Rates

In Section 5.3, we found a strong correlation between unigram entropies and entropy rates. This is surprising, as we would expect that the co-text has a variable effect on the information content of words, and that this might differ across languages too. However, what we actually find is that the co-text effect is (relatively) constant across languages. To put it differently, knowing the co-text of words decreases their uncertainty, or information content, by roughly the same amount, regardless of the language. Thus, entropy rates are systematically lower than unigram entropies, by 3.17 bits/word on average.

Notably, this result is in line with earlier findings by Montemurro and Zanette [8,9]. They have reported, for samples of eight and 75 languages respectively, that the difference between word entropy rates for texts with randomized word order and those of texts with original word order is about 3.5 bits/word. Note that the word entropy rate given randomized word order is conceptually the same as unigram entropies, since any dependencies between words are destroyed via randomization [8] (technically all the tokens of the sequence become independent and identically distributed variables ([73], p. 75). Montemurro and Zanette [8,9] also show that while the average information content of words might differ across languages, the co-text reduces the information content of words by a constant amount. They interpret this as a universal property of languages. We have shown for a considerably bigger sample of languages that the entropy difference has a smaller variance than the original unigram entropies and entropy rates, which consistent with Montemurro and Zanette's findings (Figure 1).

As a consequence, the entropy rate is a linear function of unigram entropy, and can be straightforwardly predicted from it. To our knowledge, we have provided the first empirical evidence for a linear dependency between the entropy rate and unigram entropy (Figure 3). Interestingly, we have shown that this linearity of the relationship increases as text length increases (Figure 4). A mathematical investigation of the origins of this linear relationship should be the subject of future research.

There is also a practical side to this finding: estimating entropy rates requires searching strings of length $i - 1$, where i is the index running through all tokens of a text. As i increases, the CPU time per additional word token increases linearly. In contrast, unigram entropies can be estimated based on dictionaries of word types and their token frequencies, and the processing time per additional word token is constant. Hence, Equation (11) can help to reduce processing costs considerably.

7. Conclusions

The entropy, average information content, uncertainty or choice associated with words is a core information-theoretic property of natural languages. Understanding the diversity of word entropies requires an interdisciplinary discourse between information theory, quantitative linguistics, computational linguistics, psycholinguistics, language typology, as well as historical and evolutionary linguistics.

As a first step, we have here established word entropy stabilization points for 21 languages of the European Parliament Corpus and 32 languages of the Parallel Bible Corpus. We illustrated that word entropies can be reliably estimated with text sizes of >50 K. Based on these findings, we estimated entropies across 1499 texts and 1115 languages of the Parallel Bible Corpus. These analyses shed

light on both the diversity of languages and the underlying universal pressures that shape them. While the information encoding strategies of individual languages might vary considerably across the world, they all need to adhere to fundamental principles of information transfer. In this context, word learnability and word expressivity emerge as fundamental constraints on language variation. In some scenarios of learning and usage, the pressure to be expressive might be systematically stronger than the pressure to be learnable.

Furthermore, we have shown that there is a strong linear relationship between unigram entropies and their entropy rates, which holds across different macro areas of the world's languages. The theoretical implication of this finding is that co-text effects on the predictability of words are relatively similar regardless of the language in consideration. The practical implication is that entropy rates can be approximated by using unigram entropies, thus reducing processing costs.

Acknowledgments: We are grateful to Ł. Dębowski, Richard Futrell, and Ryosuke Takahira for helpful discussions. We would also like to thank two anonymous reviewers for helpful suggestions and feedback. Christian Bentz was funded by the German Research Foundation (DFG FOR 2237: Project "Words, Bones, Genes, Tools: Tracking Linguistic, Cultural, and Biological Trajectories of the Human Past") and by the ERC Advanced Grant 324246 EVOLAEMP. Ramon Ferrer-i-Cancho was funded by the grants 2014SGR 890 (MACDA) from AGAUR (Generalitat de Catalunya) and also the APCOM project (TIN2014-57226-P) from MINECO (Ministerio de Economia y Competitividad).

Author Contributions: Christian Bentz, Dimitrios Alikaniotis and Ramon Ferrer-i-Cancho conceived of and designed the experiments. Christian Bentz and Dimitrios Alikaniotis performed the experiments. Christian Bentz and Dimitrios Alikaniotis analysed the data. Michael Cysouw contributed materials. Christian Bentz, Ramon Ferrer-i-Cancho and Michael Cysouw wrote the paper. All authors have read and approved the final manuscript.

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

Abbreviations

The following abbreviations are used in this manuscript:

EPC	European Parliament Corpus
PBC	Parallel Bible Corpus
UDHR	Universal Declaration of Human Rights
NSB	Nemenman–Shafee–Bialek

Appendix A. Text Pre-Processing

Appendix A.1. Converting Letters to Lower Case

This is done via the function tolower() in R. Cross-linguistically, it is common that upper case letters are used to indicate the beginning of a sentence, i.e., *the* and *The*. In this case, we are not dealing with different word types from a grammatical or semantic perspective, but with two alternative writings of the same word type. Therefore, it is common practice to convert all letters to lower case as a pre-processing step.

Note that, in a few cases, conversion of letters to lower case can create polysemy. Namely, when upper and lower case actually distinguish separate word types. For example, in German, verbs can be nominalized by using an upper case initial letter: fahren "drive", and das Fahren "the driving" are two different word types. This difference is neutralized when all letters are set to lower case.

Appendix A.2. Removal of Punctuation

All three corpora (EPC, PBC, UDHR) are encoded in unicode UTF-8, which generally makes them accessible to the same automated processing tools. However, there are subtle difference with regards to orthographic practices. While the EPC follows common orthographic conventions, the PBC and UDHR are specifically curated to delimit word types and punctuation by additional white spaces. Take the following example sentences:

Although, as you will have seen, the dreaded "millennium bug" failed to materialise, [...]

EPC (English, line 3)

And God said, let there be light . And there was light . [...]

PBC (English, Genesis 1:3)

Furthermore, no distinction shall be made on the basis of the political, jurisdictional or international status of the country [...]

UDHR (English, paragraph 3)

The automatically added, and manually checked, white spaces (also before punctuation) in the PBC and UDHR make cross-linguistic analyses of word types more reliable. This is especially helpful when we deal with hundreds of languages with a multitude of different writing systems and scripts. For instance, characters like apostrophes and hyphens are often ambiguous as to whether they are part of a word type, or being used as punctuation.

Apostrophes can indicate contractions as in English *she's* representing *she is*. Here, the apostrophe functions as a punctuation mark. In other cases, however, it might reflect phonetic distinctions. For instance, glottal stops or ejectives as in the Mayan language K'iche' (quc). Take the word *q'atb'altzij* meaning "everyone". Here, the apostrophes are part of the word type, and should not be removed. To disambiguate these different usages the PBC and UDHR would give *she ' s* versus *q'atb'altzij*.

Another example are raised numbers indicating tone distinctions. In the Usila Chinantec (cuc) version of the Bible, for instance, the proper name Abraham is rendered as A^3brang^{23}. These numbers indicate differences in pitch when pronouncing certain word types, and are hence part of their lexical (and grammatical) properties. If tone numbers are interpreted instead as non-alphanumeric indications of footnotes, then words might be erroneously split into separate parts, e.g. *A brang*. Again, the PBC and UDHR use white spaces to disambiguate here.

Given this difference in the usage of white spaces, we use two different strategies to remove punctuation:

1. For the EPC, we use the regular expression `\\W+` in combination with the R function strsplit() to split strings of UTF-8 characters on punctuation and white spaces.
2. For the PBC and UDHR, we define a regular expression meaning "at least one alpha-numeric character between white spaces" which would be written as: `.*[[:alpha:]].*`
   This regex can then be matched with the respective text to yield word types. This is done via the functions regexpr() and regmatches() in R.

Appendix B. Advanced Entropy Estimators

As outlined in Section 3.2, to estimate the entropy given in Equation (3) reliably, the critical part is to get a good approximation of the probability of words. Hence, $p(w_i)$ is the critical variable to be approximated with highest possible precision. Henceforth, estimated probabilities are denoted as $\hat{p}(w_i)$, and estimated entropies correspondingly $\hat{H}$.

Using frequency counts from a subsample, i.e., using:

$$\hat{p}(w_i)^{ML} = \frac{f_i}{\sum_{j=1}^{V} f_j}, \tag{A1}$$

where the denominator is the overall number of tokens, i.e., the sum of type frequencies over an empirical (i.e., finite) vocabulary size $V = |\mathcal{V}|$, is generally referred to as the maximum likelihood (ML) method. Note that the denominator here is equivalent to the total number of tokens, denoted as N in the following. Plugging Equation (A1) into Equation (3) yields the estimated entropy as [63]:

$$\hat{H}^{ML} = -\sum_{i=1}^{V} \hat{p}(w_i)^{ML} \log_2(\hat{p}(w_i)^{ML}). \tag{A2}$$

The ML method yields reliable results for situations where $N \gg V$, i.e., the number of tokens is much bigger than the empirical number of types in a given text ([63] p. 1470). In other words, it is reliable for a small ratio of word types to word tokens V/N. Since in natural language this ratio is typically big for small texts, and only decreases with N [67,82], entropy estimation tends to be unreliable for small texts. However, since Shannon's original proposal in the 1950s a range of entropy estimators have been proposed to overcome the underestimation bias ([63] see for an overview). Some of these are discussed in turn.

Appendix B.1. The Miller–Madow Estimator

For example, the Miller–Madow (MM) estimator ([63], p. 1471) tries to reduce the bias by adding a correction to the ML estimated entropy, such that:

$$\hat{H}^{MM} = \hat{H}^{ML} + \frac{M_{>0} - 1}{2N}, \tag{A3}$$

where $M_{>0}$ refers to the number of types with token frequencies > 0, i.e., V in our definition. Note that the correction $\frac{V-1}{2N}$ is relatively big for the $N < V$ scenario, and relatively small for the $N > V$ scenario. Hence, it counterbalances the underestimation bias in the $N < V$ scenario of small text sizes.

Appendix B.2. Bayesian Estimators

Another set of estimators derives from estimating $p(w_i)$ within a Bayesian framework using the Dirichlet distribution with $a_1, a_2, ..., a_V$ as priors, such that:

$$\hat{p}(w_i)^{Bayes} = \frac{f_i + a_i}{N + A}, \tag{A4}$$

where a_i values essentially "flatten out" the distribution of frequency counts to overcome the bias towards short tailed distributions (with small V). Besides N, we have $A = \sum_{i=1}^{V} a_i$ added to the denominator ([83], pp. 302–303).

Now, depending on which priors exactly we choose, we end up with different estimated entropies (see also Table 1 in (Hausser and Strimmer [63], p. 1471)). A uniform Jeffreys prior of $a_i = 1/2$ gives us $\hat{H}^{Jeff}$, a uniform Laplace prior of $a_i = 1$ gives us $\hat{H}^{Lap}$, a uniform Perks prior of $a_i = 1/V$ gives us $\hat{H}^{SG}$, after Schürmann and Grassberger [6], who proposed to use this prior. Finally, the so-called minimax prior of $a_i = \sqrt{N/V}$ yields $\hat{H}^{minimax}$.

Furthermore, the most recent, and arguably least biased, entropy estimator based on a Bayesian framework is the Nemenman–Shafee–Bialek (NSB) estimator [52]. (Nemenman et al. [52], p. 5) illustrate that the entropies estimated with the other priors proposed above will be strongly influenced by the prior distributions and only recover after a relatively big number of tokens has been sampled. Instead of directly using any specific Dirichlet prior, they form priors as weighted sums of the different Dirichlet priors, which they call infinite Dirichlet mixture priors. The resulting entropy estimates $\hat{H}^{NSB}$ turn out to be robust across the whole range of sample sizes.

Appendix B.3. The Chao–Shen Estimator

Chao and Shen [84] (p. 432) propose to overcome the problem of overestimating the probability of each type (in their case species instead of word types) by first estimating the so-called sample coverage as:

$$\hat{C} = 1 - \frac{m_1}{N}, \tag{A5}$$

where m_1 is the number of types with frequency one in the sample (i.e., hapax legomena). The idea is that the number of types not represented by tokens is roughly the same as the number of types with frequency one. In this case, the sample coverage reflects the conditional probability of getting a new type if a token is added to the sample N. This probability is then multiplied with the simple ML estimate $\hat{p}(w_i)^{ML}$ to get the so-called Good-Turing estimated probability of a type:

$$\hat{p}(w_i)^{GT} = \left(1 - \frac{m_1}{N}\right) \hat{p}(w_i)^{ML}. \tag{A6}$$

Furthermore, (Chao and Shen [84], p. 431) suggest to use the Horvitz–Thompson estimator to modify the estimated entropy $\hat{H}^{ML}$. This estimator is based on the rationale that if N tokens have been sampled with replacement, then the probability of the i-th type not being represented by a specific token is $1 - \hat{p}(w_i)^{GT}$. Thus, the probability of the i-th type not being represented by any token is $(1 - \hat{p}(w_i)^{GT})^N$, and, inversely, the probability of appearing at least once in a sample of N tokens is $1 - (1 - p(w_i)^{GT})^N$. The full specification of the Horvitz–Thompson estimator, with Good-Turing probability estimates, is then:

$$\hat{H}^{CS} = -\sum_{i=1}^{V} \frac{\hat{p}(w_i)^{GT} \log_2(\hat{p}(w_i)^{GT})}{1 - (1 - \hat{p}(w_i)^{GT})^N}. \tag{A7}$$

Appendix B.4. The James–Stein Shrinkage Estimator

Finally, (Hausser and Strimmer [63], p. 1472) put forward an entropy estimator based on the so-called James–Stein shrinkage. According to this approach the estimated probability per type is:

$$\hat{p}(w_i)^{shrink} = \lambda \hat{p}(w_i)^{target} + (1 - \lambda)\hat{p}(w_i)^{ML}, \tag{A8}$$

where $\lambda \in [0, 1]$ is the shrinkage intensity and $\hat{p}(w_i)^{target}$ is the so-called "shrinkage target". (Hausser and Strimmer [63], p. 1473) suggest to use the maximum entropy distribution as a target, i.e., $\hat{p}(w_i)^{target} = \frac{1}{V}$. This yields:

$$\hat{p}(w_i)^{shrink} = \frac{\lambda}{V} + (1 - \lambda)\hat{p}(w_i)^{ML}. \tag{A9}$$

The idea here is that the estimated probability $\hat{p}(w_i)^{shrink}$ consists of two additive components, $\frac{\lambda}{V}$ and $(1 - \lambda)\hat{p}(w_i)^{ML}$ respectively. In the full shrinkage case ($\lambda = 1$) Equation (A9) yields:

$$\hat{p}(w_i)^{shrink} = \frac{1}{V}, \tag{A10}$$

i.e., uniform probabilities that will yield maximum entropy. In the lowest shrinkage case ($\lambda = 0$) Equation (A9) yields:

$$\hat{p}(w_i)^{shrink} = \hat{p}(w_i)^{ML}, \tag{A11}$$

i.e., the ML estimation that is biased towards low entropy. Given empirical data, the true probability is very likely to lie somewhere in between these two cases and hence $0 < \lambda < 1$. In fact, (Hausser and Strimmer [63], p. 1481) show that the optimal shrinkage λ^* can be calculated analytically and without knowing the true probabilities $p(w_i)$. Given the optimal shrinkage, the probability $\hat{p}(w_i)^{shrink}$ can then be plugged into the original entropy equation to yield:

$$\hat{H}^{shrink} = -\sum_{i=1}^{V} \hat{p}(w_i)^{shrink} \log_2 \hat{p}(w_i)^{shrink}. \tag{A12}$$

Overall, nine different entropy estimators were outlined here, from the most simple and "naive" maximum likelihood estimator to very advanced and arguably much less biased estimators such as the NSB estimator. All of these are available for estimation via the R package entropy [76].

Appendix C. Stabilization of the Entropy Rates for 21 Languages of the European Parliament Corpus

Appendix C.1. Unigram Entropies

Figure A1 illustrates the stabilization of unigram entropies ($H_1(T)$) across 21 languages of the EPC. Nine different entropy estimators are used here, including the simple maximum likelihood estimator (ML). The estimation methods are further detailed in Appendix B. Some slight differences between estimators are visible. For instance, Bayesian estimation with a Laplace prior (Lap) yields systematically higher results compared to the other estimators. However, this difference is in the decimal range. If we take the first panel with Bulgarian (bg) as an example: at 100 K tokens, the lowest estimated entropy value is found for maximum likelihood estimation ($\hat{H}^{ML} = 10.11$), whereas the highest values are obtained for Bayesian estimation with Laplace and Jeffrey's priors respectively ($\hat{H}^{Lap} = 10.66$ and $\hat{H}^{Lap} = 10.41$). This reflects the expected underestimation-bias of the ML estimator, and the overestimation-biases of the Laplace and Jeffrey's priors, which have been reported in an earlier study [63].

However, overall the stabilization behaviour of entropy estimators is similar across all languages. Namely, in the range of ca. zero to 25 K tokens, they all undergo a fast growth, and start to stabilize after 25 K. This stabilization behaviour is visualized in Figure A2. Here, standard deviations of estimated entropy values (y-axis) are plotted against text sizes in number of tokens (x-axis). A standard deviation of 0.1 is given as a reference line (dashed grey). Across the 21 languages, SDs fall below this line at 50K tokens at the latest.

Remember that, relating to Problem 1, we were asking for the text size at which a meaningful cross-linguistic comparison of word entropies is possible. We can give a pragmatic answer to this question: If we want to estimate the unigram entropy of a given text and language with a local precision of one decimal place after the comma, we have to supply >50 K tokens. We use the term "local" to emphasize that this is not precision with respect to the true entropic measure, but with respect to prefixes of a length within a certain window (recall Section 4).

Furthermore, note that the curves of different estimation methods in both Figures A1 and A2 have similar shapes, and, in some cases, are largely parallel. This suggests that despite the methodological differences, the values are strongly correlated. This is indeed the case. In fact, even the results of the simple ML method are strongly correlated with the results of all other methods, with Pearson's r ranging from 0.94 to one. See Appendix E for details.

Appendix C.2. Entropy Rate

The results for entropy rates ($\hat{h}(T)$) are visualized in Figures A3 and A4. The stabilization points and SDs are similar to the ones established for unigram entropies. Namely, below 25 K tokens the entropy rate is generally lower than for longer prefixes, in some cases extremely, as for Polish (pl) and Czech (cs), where the values still range around five bits/word at five K tokens, and then steeply rise to approximately seven bits/word at 25 K tokens.

In parallel to the results for unigram entropies, the 21 languages reach stable entropy values at around 50 K. This is visible in Figure A4. Again, the grey dashed line indicates a standard deviation of 0.1, i.e., values that are precise to the first decimal place. Moreover, to double-check whether entropy rate stabilization is robust across a wider range of languages, we also tested the estimator on a sample of 32 languages from the major language families represented in the PBC corpus (see Appendix D).

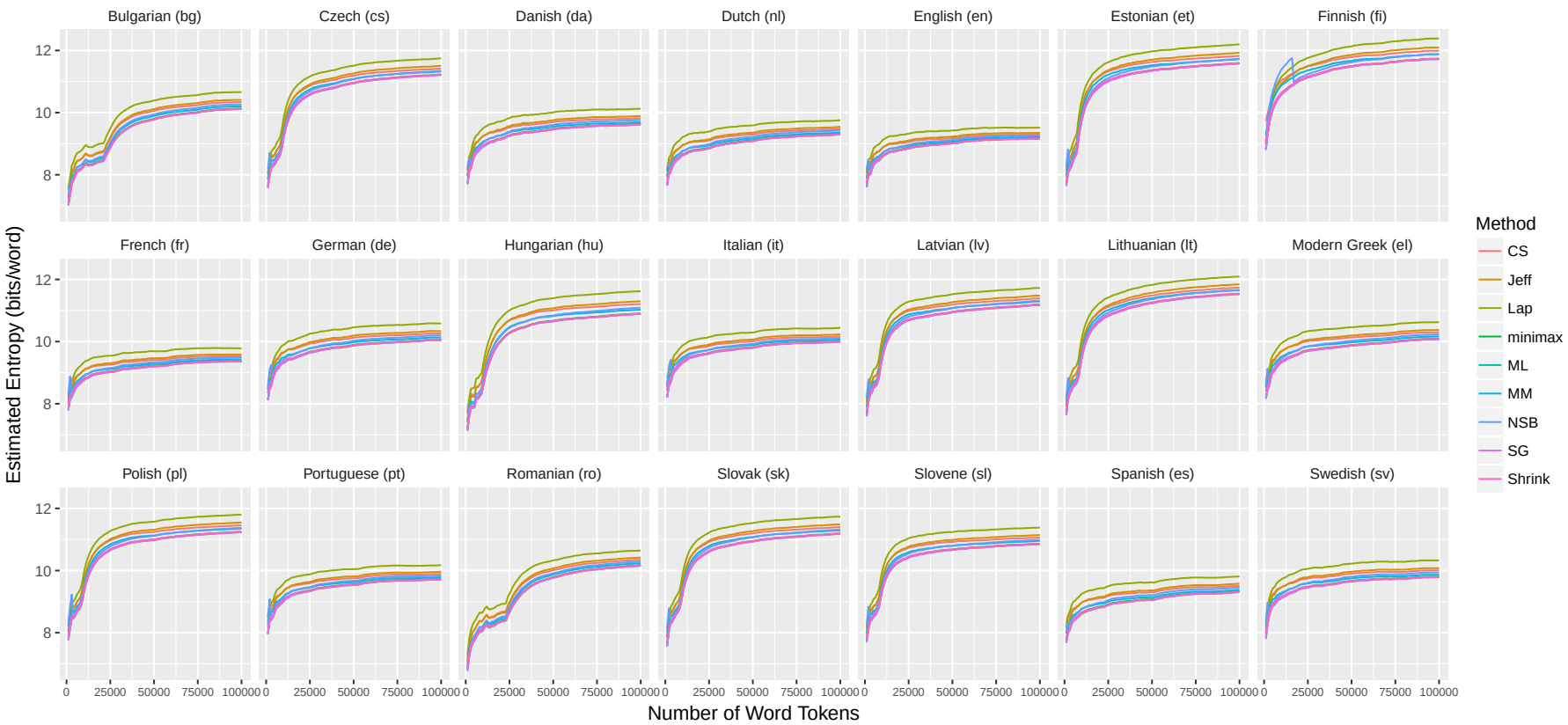

Figure A1. Unigram entropies (y-axis) as a function of text length (x-axis) across 21 languages of the EPC corpus. Unigram entropies are estimated on prefixes of the text sequence increasing by one K tokens. Thus, the first prefix covers tokens one to one K, the second prefix covers tokens one to two K, etc. The number of tokens is limited to 100 K, since entropy values already (largely) stabilize throughout the text sequence before that. Hence, there are 100 points along the x-axis. Nine different methods of entropy estimation are indicated with colours. CS: Chao–Shen estimator, Jeff: Bayesian estimation with Jeffrey's prior, Lap: Bayesian estimation with Laplace prior, minimax: Bayesian estimation with minimax prior, ML: maximum likelihood, MM: Miller–Madow estimator, NSB: Nemenman–Shafee–Bialek estimator, SG: Schürmann–Grassberger estimator, Shrink: James–Stein shrinkage estimator. Detailed explanations for these estimators are given in Appendix B. Language identifiers used by the EPC are given in parenthesis.

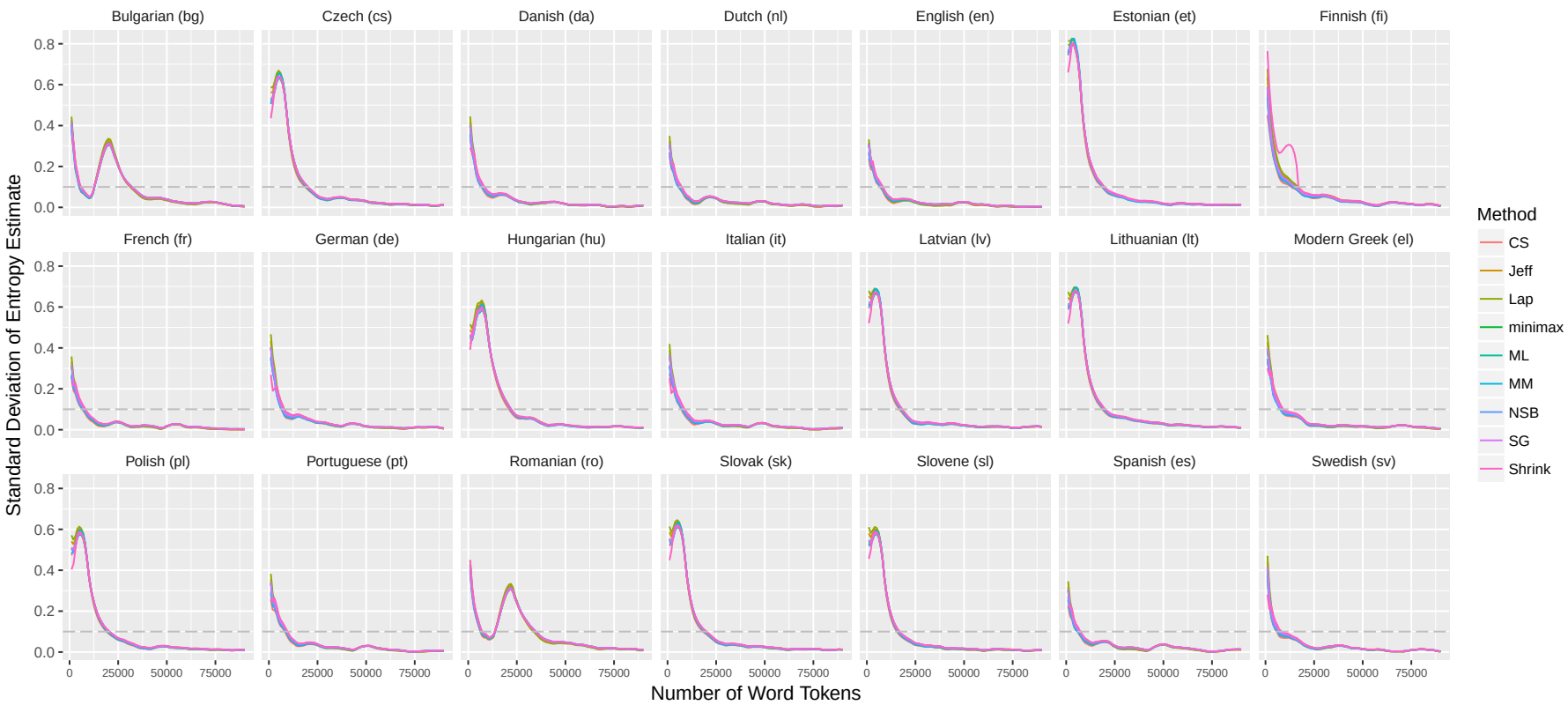

Figure A2. SDs of unigram entropies (y-axis) as a function of text length (x-axis) across 21 languages of the EPC corpus, and the nine different estimators. Unigram entropies are estimated on prefixes of the text sequence increasing by one K tokens as in Figure A1. SDs are calculated over the entropies of the next 10 prefixes as explained in Section 4. Hence, there are 90 points along the x-axis. The horizontal dashed line indicates $SD = 0.1$ as a threshold.

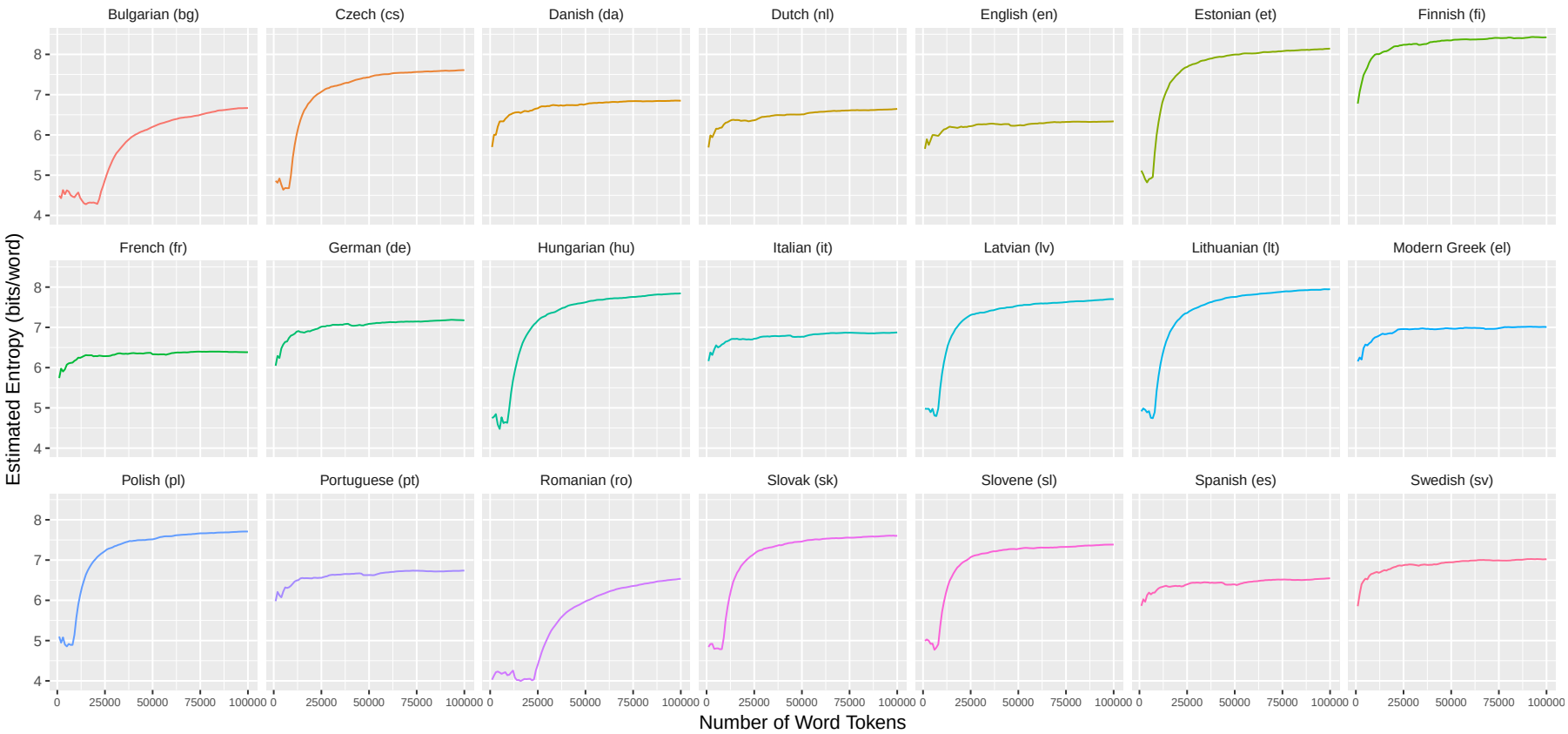

Figure A3. Entropy rates as a function of text length across 21 languages of the EPC. Entropy rates are estimated on prefixes of the text sequence increasing by one K tokens as in Figure A1. Hence there are 100 points along the *x*-axis. The language identifiers used by the EPC are given in parenthesis.

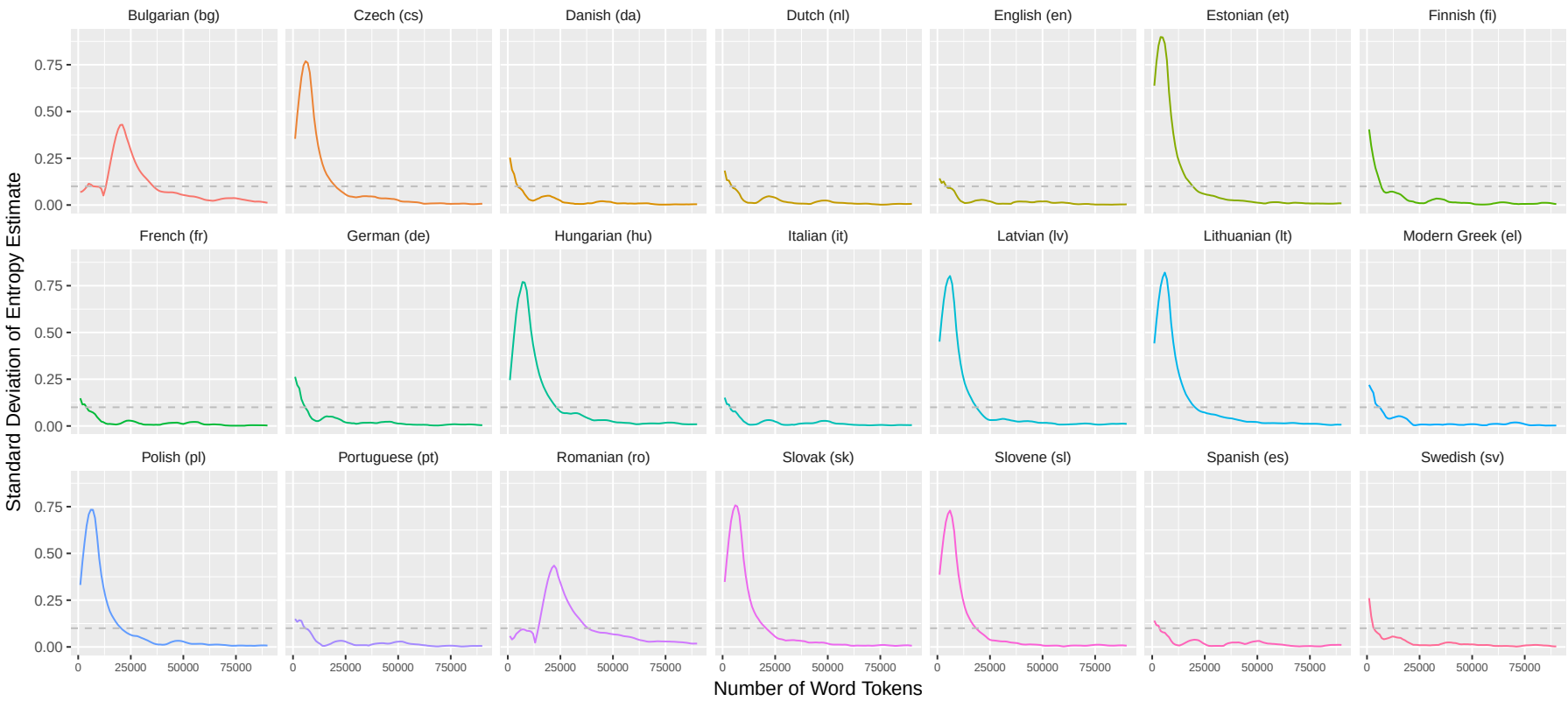

Figure A4. SDs of entropy rates as a function of text length across 21 languages of the EPC corpus. The format is the same is as in Figure A2.

Appendix D. Stabilization of Entropy Rates for 32 Languages of the Parallel Bible Corpus

Here we report entropy rate stabilization results for 32 languages of the PBC (Figures A5 and A6). These 32 languages were chosen to represent the major language families across the world. They give a better overview of the linguistic diversity than the 21 mainly Indo-European languages represented in the EPC. The methodology is the same as for the original analyses with the 21 EPC languages described in the main paper.

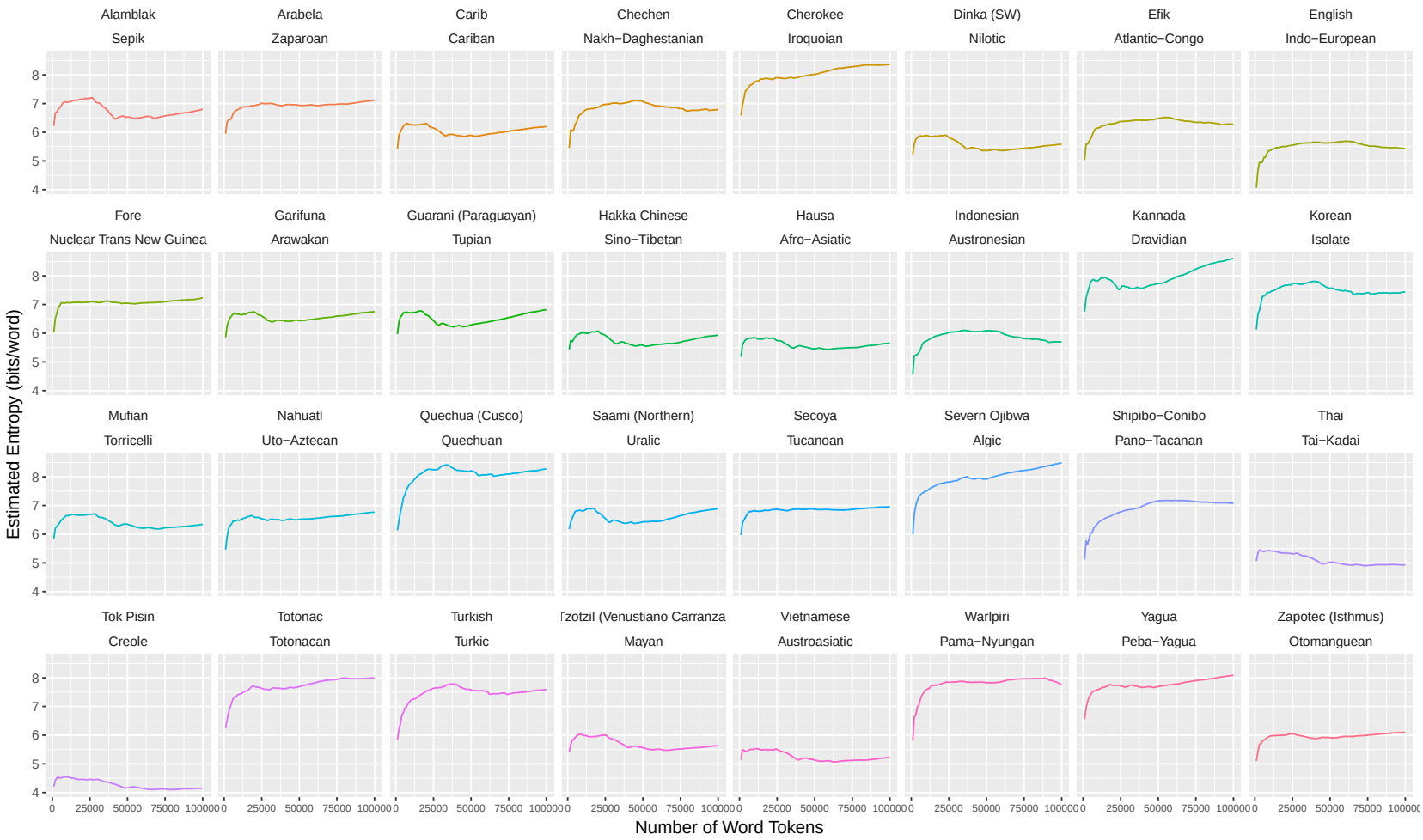

Figure A5. Entropy rates as a function of text length across 32 languages of the PBC. Languages were chosen to represent some of the major language families across the world. Entropy rates are estimated on prefixes of the text sequence increasing by one K tokens as in Figure A1. Hence there are 100 points along the x-axis. The language names and families are taken from Glottolog 2.7 [78] and given above the plots.

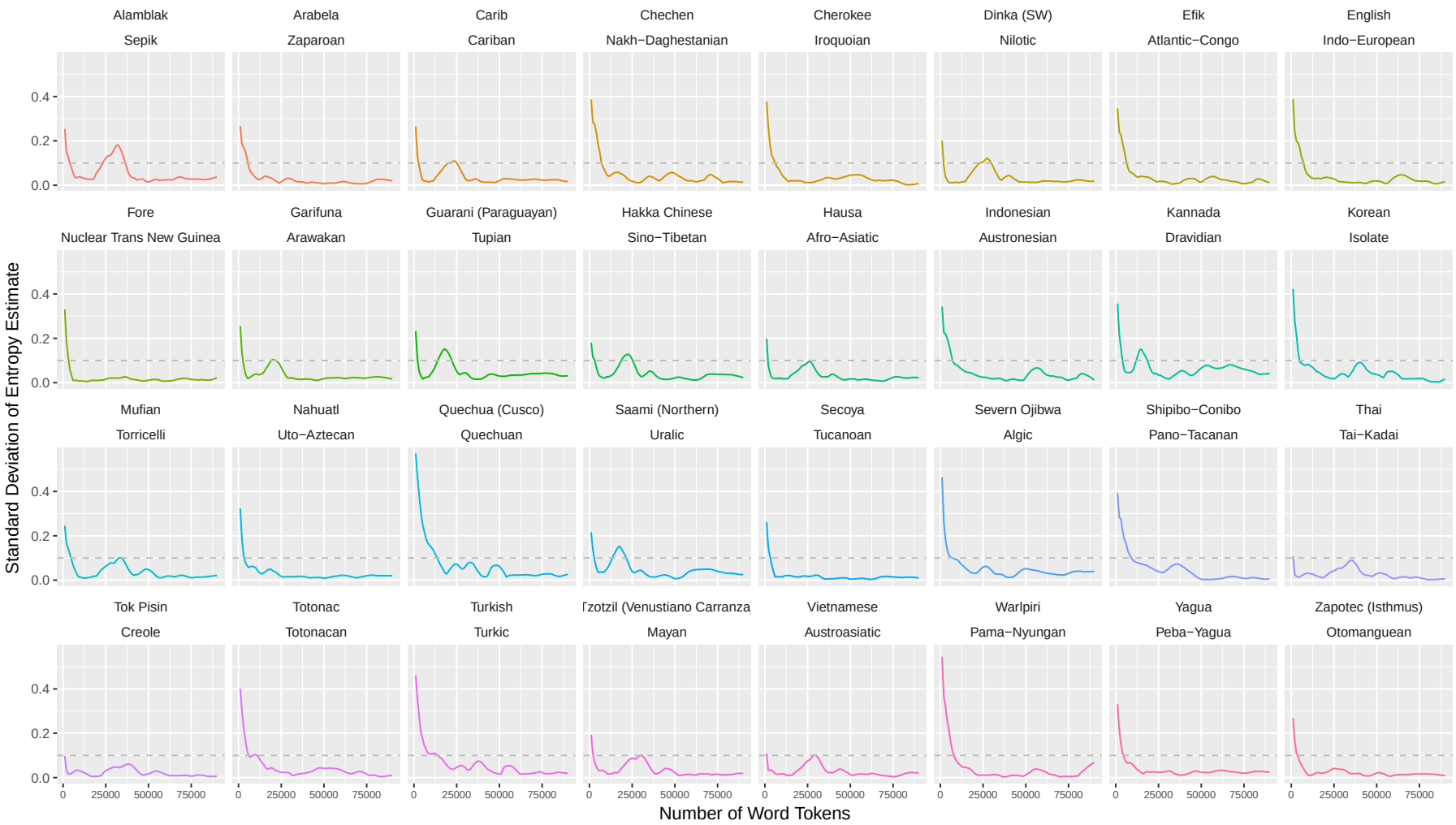

Figure A6. SDs of entropy rates as a function of text length across 32 languages of the PBC. Languages were chosen to represent some of the major language families across the world. The format is as in Figure A2. The language names and families are given above the plots.

Appendix E. Correlations between Estimated Unigram Entropies

Given all the different entropy estimators proposed in the literature, we want to assess how strongly values given by different estimators correlate. Here we give the pairwise Pearson correlations of estimated unigram entropies for the nine proposed estimators and three parallel corpora. Hence, for each corpus there are 36 pairwise correlations. These can be seen in Tables A1–A3. For visual illustration a plot of pairwise correlations of the ML estimator with all other estimators is given in Figure A7.

Some of the lowest correlations are found for unigram entropy values as estimated by the NSB estimator and all the other estimators for the UDHR corpus. This makes sense considering that the NSB estimator is designed to converge early [52], while other estimators, e.g., the Laplace and Jeffrey's prior in a Bayesian framework, have been shown to overestimate the entropy for small sample sizes [63]. Hence, it is to be expected that the biggest divergence is found in the smallest corpus, as entropy estimation is harder with smaller sample sizes. This is the case for the UDHR. However, note that even given the small number of tokens in the UDHR, and the differences in entropy estimation, the lowest correlation is remarkably strong ($r = 0.946$ for NSB and CS). In fact, for the EPC and PBC two correlations are perfect $r = 1$, even between estimators that differ conceptually, e.g., ML and SG.

Overall, this illustrates that in practice the choice of unigram estimators is a very minor issue for cross-linguistic comparison, as long as we deal with text sizes at which estimations have reached stable values.

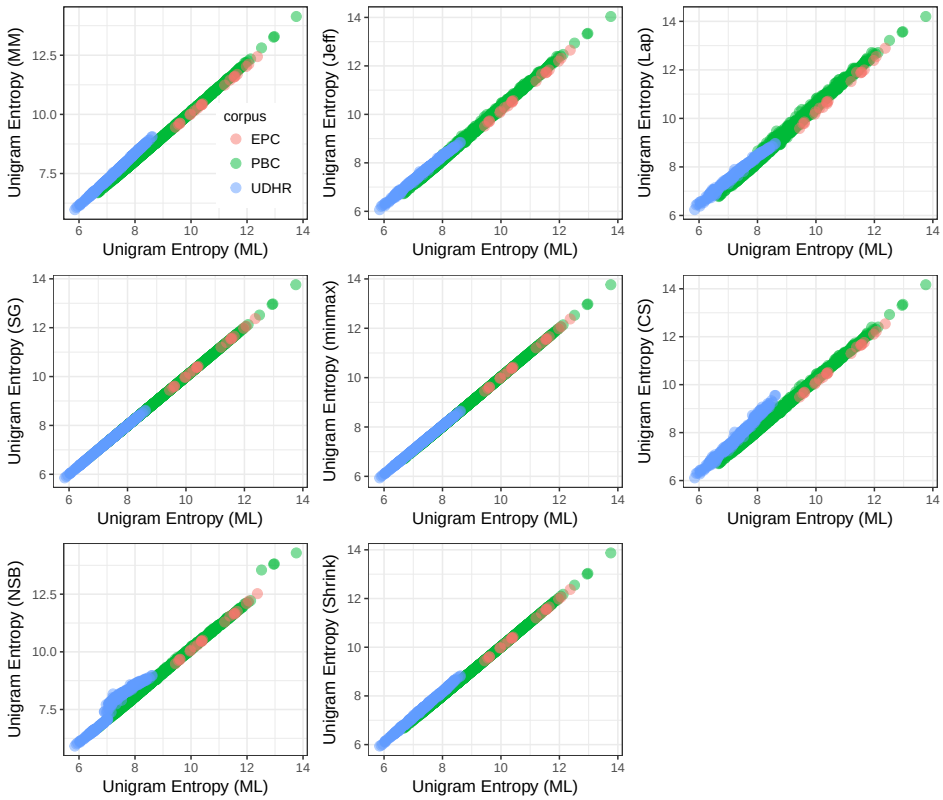

Figure A7. Pairwise correlations of estimated unigram entropy values for three different corpora: Europarl Corpus (EPC), Parallel Bible Corpus (PBC), and Universal Declaration of Human Rights (UDHR). Results of the maximum likelihood (ML) method are here taken as a baseline and correlated with all other methods. CS: Chao–Shen estimator, Jeff: Bayesian estimation with Jeffrey's prior, Lap: Bayesian estimation with Laplace prior, minimax: Bayesian estimation with minimax prior, MM: Miller–Madow estimator, NSB: Nemenman–Shafee–Bialek estimator, SG: Schürmann–Grassberger estimator, Shrink: James–Stein shrinkage estimator.

Table A1. Pairwise correlations of unigram entropies for the EPC corpus.

-	ML	MM	Jeff	Lap	SG	Minmax	CS	NSB	Shrink
ML	-								
MM	0.9999405	-							
Jeff	0.9994266	0.9996888	-						
Lap	0.9983479	0.998819	0.9997185	-					
SG	1	0.9999405	0.9994267	0.998348	-				
minmax	0.9999999	0.9999445	0.9994415	0.9983733	0.9999999	-			
CS	0.9993888	0.9996607	0.9999867	0.9997199	0.9993889	0.9994037	-		
NSB	0.9997953	0.9999065	0.9998969	0.9992965	0.9997954	0.9998041	0.9998719	-	
Shrink	0.9999945	0.9999059	0.9993348	0.9981998	0.9999945	0.9999935	0.9992906	0.9997419	-

Table A2. Pairwise correlations of unigram entropies for the PBC corpus.

-	ML	MM	Jeff	Lap	SG	Minmax	CS	NSB	Shrink
ML	-								
MM	0.9998609	-							
Jeff	0.9989818	0.9993252	-						
Lap	0.9969406	0.9975655	0.999441	-					
SG	1	0.9998611	0.9989821	0.9969412	-				
minmax	0.9999988	0.9998743	0.9990352	0.9970343	0.9999989	-			
CS	0.998965	0.999388	0.9999208	0.9992828	0.9989654	0.9990176	-		
NSB	0.999173	0.9994438	0.9992162	0.9979161	0.9991732	0.9992024	0.9993134	-	
Shrink	0.9999805	0.9998643	0.9988525	0.9967172	0.9999806	0.9999785	0.9988745	0.9991464	-

Table A3. Pairwise correlations of unigram entropies for the UDHR corpus.

-	ML	MM	Jeff	Lap	SG	Minmax	CS	NSB	Shrink
ML	-								
MM	0.9979922	-							
Jeff	0.9975981	0.9952309	-						
Lap	0.9928889	0.9895672	0.9986983	-					
SG	0.9999999	0.9980081	0.9976072	0.9929003	-				
minmax	0.999976	0.9979101	0.9980311	0.9936525	0.9999768	-			
CS	0.9854943	0.9932518	0.9826871	0.9763849	0.985522	0.9853826	-		
NSB	0.9623212	0.9621106	0.9663655	0.964961	0.9623601	0.962801	0.9459217	-	
Shrink	0.9986898	0.9984146	0.9942842	0.9877932	0.9986974	0.9984619	0.9866643	0.9607329	-

Appendix F. Correlations between Unigram Entropies and Entropy Rates for the PBC

In Section 5.3, we report the correlation for unigram entropies as estimated with the NSB method, and entropy rates. For completeness, here we give Pearson correlations between unigram entropies, as estimated with all nine methods, and entropy rates in Figure A8. The left panel plot in Figure 3 corresponds to the third plot from the right in the last row of Figure A8. However, the four texts with extreme values (beyond 13 bits/word) are not excluded here, they can be seen in the upper right corners of the scatterplots.

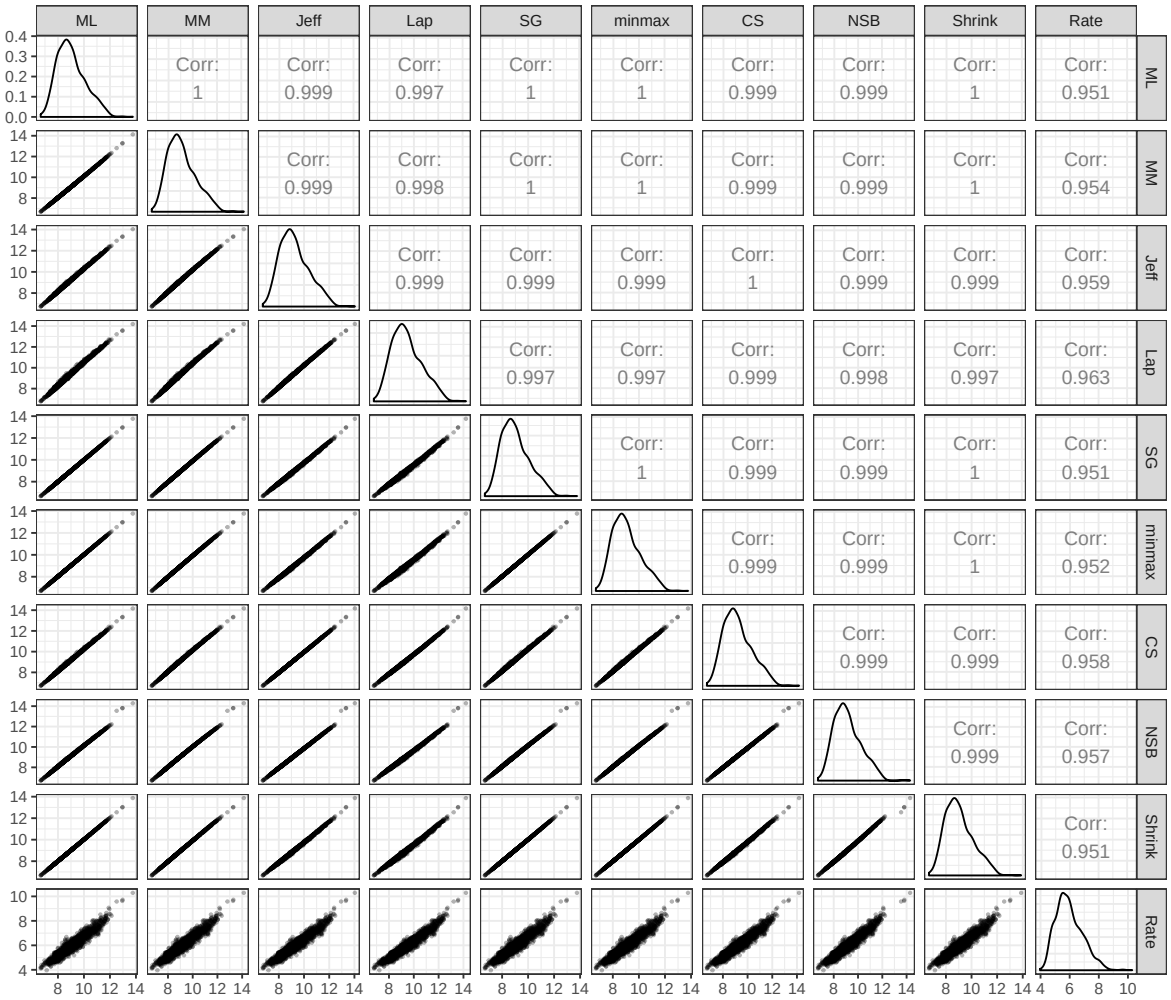

Figure A8. Correlations between the nine unigram entropy estimators and entropy rates for the PBC. The panels in the lower half of the plot give scatterplots, the panels in the upper half give corresponding Pearson correlations. The diagonal panels give density plots.

Appendix G. Correlations between PBC, EPC and UDHR Unigram Entropies

Figure A9 illustrates the Pearson correlations between unigram entropies (as estimated with the NSB method) for texts of the PBC, EPC, and the UDHR. The datasets are merged by ISO 639-3 codes, i.e., by languages represented in a corpus. This yields a sample of 599 texts for the PBC/UDHR (left panel) which share the same languages, 26 texts for the EPC/UDHR (middle panel), and 160 texts for the PBC/EPC. The Pearson correlations between the estimated unigram entropies of these texts in the three corpora are strong ($r = 0.77, p < 0.0001$; $r = 0.75, p < 0.0001$; $r = 0.89, p < 0.0001$). It is visible that the relationship, especially for the PBC/UDHR and EPC/UDHR comparison, is non-linear. This is most likely due to the fact that the smaller UDHR corpus (ca. 2000 tokens per text) results in underestimated entropies especially for high entropy languages. However, overall the correlations are strong, suggesting that rankings of languages according to unigram entropies of texts are stable even across different corpora.

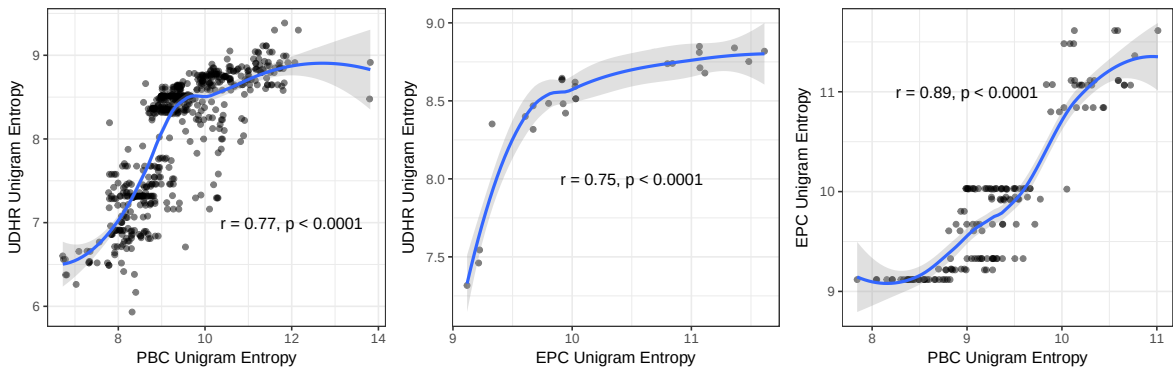

Figure A9. Correlations between unigram entropies (NSB estimated) for texts of the PBC and UDHR (left panel), texts of the EPC and UDHR (middle panel), and texts of the PBC and EPC (right panel). Local regression smoothers are given (blue lines) with 95% confidence intervals.

References

1. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
2. Shannon, C.E. Prediction and entropy of printed English. *Bell Syst. Tech. J.* **1951**, *30*, 50–65.
3. Brown, P.F.; Pietra, V.J.D.; Mercer, R.L.; Pietra, S.A.D.; Lai, J.C. An estimate of an upper bound for the entropy of English. *Comput. Linguist.* **1992**, *18*, 31–40.
4. Kontoyiannis, I.; Algoet, P.H.; Suhov, Y.M.; Wyner, A.J. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. *IEEE Trans. Inf. Theory* **1998**, *44*, 1319–1327.
5. Gao, Y.; Kontoyiannis, I.; Bienenstock, E. Estimating the entropy of binary time series: Methodology, some theory and a simulation study. *Entropy* **2008**, *10*, 71–99.
6. Schürmann, T.; Grassberger, P. Entropy estimation of symbol sequences. *Chaos Interdiscip. J. Nonlinear Sci.* **1996**, *6*, 414–427.
7. Behr, F.H.; Fossum, V.; Mitzenmacher, M.; Xiao, D. Estimating and comparing entropies across written natural languages using PPM compression. In Proceedings of the DCC 2003 Data Compression Conference, Snowbird, UT, USA, 25–27 March 2003; p. 416.
8. Montemurro, M.A.; Zanette, D.H. Universal entropy of word ordering across linguistic families. *PLoS ONE* **2011**, *6*, e19875.
9. Montemurro, M.A.; Zanette, D.H. Complexity and universality in the long-range order of words. In *Creativity and Universality in Language*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 27–41.
10. Koplenig, A.; Meyer, P.; Wolfer, S.; Müller-Spitzer, C. The statistical trade-off between word order and word structure–Large-scale evidence for the principle of least effort. *PLoS ONE* **2017**, *12*, e0173614.
11. Takahira, R.; Tanaka-Ishii, K.; Dębowski, Ł. Entropy Rate Estimates for Natural Language—A New Extrapolation of Compressed Large-Scale Corpora. *Entropy* **2016**, *18*, 364.
12. Bentz, C.; Verkerk, A.; Kiela, D.; Hill, F.; Buttery, P. Adaptive communication: Languages with more non-native speakers tend to have fewer word forms. *PLoS ONE* **2015**, *10*, e0128254.
13. Ehret, K.; Szmrecsanyi, B. An information-theoretic approach to assess linguistic complexity. In *Complexity and Isolation*; Baechler, R., Seiler, G., Eds.; de Gruyter: Berlin, Germany, 2016.
14. Bentz, C. The Low-Complexity-Belt: Evidence for large-scale language contact in human prehistory? In Proceedings of the 11th International Conference (EVOLANG11) on The Evolution of Language, New Orleans, LA, USA, 21 March 2016.
15. Juola, P. Measuring linguistic complexity: The morphological tier. *J. Quant. Linguist.* **1998**, *5*, 206–213.
16. Juola, P. Assessing linguistic complexity. In *Language Complexity: Typology, Contact, Change*; Miestamo, M., Sinnemäki, K., Karlsson, F., Eds.; John Benjamins: Amsterdam, The Netherlands, 2008; pp. 89–108.
17. Gerlach, M.; Font-Clos, F.; Altmann, E.G. Similarity of symbol frequency distributions with heavy tails. *Phys. Rev. X* **2016**, *6*, 021009.

18. Bentz, C.; Berdicevskis, A. Learning pressures reduce morphological complexity: Linking corpus, computational and experimental evidence. In Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), 26th International Conference on Computational Linguistics, Osaka, Japan, 11 December 2016.
19. Bochkarev, V.; Solovyev, V.; Wichmann, S. Universals versus historical contingencies in lexical evolution. *J. R. Soc. Interface* **2014**, *11*, 20140841.
20. Rao, R.P.N.; Yadav, N.; Vahia, M.N.; Joglekar, H.; Adhikari, R.; Mahadevan, I. Entropic evidence for linguistic structure in the Indus script. *Science* **2009**, *324*, 1165.
21. Rao, R.P.N. Probabilistic analysis of an acient undeciphered script. *Computer* **2010**, *43*, 76–80.
22. Sproat, R. A statistical comparison of written language and nonlinguistic symbol systems. *Language* **2014**, *90*, 457–481.
23. Rao, R.P.; Yadav, N.; Vahia, M.N.; Joglekar, H.; Adhikari, R.; Mahadevan, I. Entropy, the Indus script, and language: A reply to R. Sproat. *Comput. Linguist.* **2010**, *36*, 795–805.
24. Piantadosi, S.T.; Tily, H.; Gibson, E. Word lengths are optimized for efficient communication. *Proc. Natl. Acad. Sci. USA* **2010**, *108*, 3526–3529.
25. Mahowald, K.; Fedorenko, E.; Piantadosi, S.T.; Gibson, E. Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition* **2013**, *126*, 313–318.
26. Ferrer-i Cancho, R.; Bentz, C.; Seguin, C. Compression and the origins of Zipf's law of abbreviation. *arXiv* **2015**, arXiv:1504.04884.
27. Bentz, C.; Ferrer-i-Cancho, R. Zipf's law of abbreviation as a language universal. In Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics, Leiden, The Netherlands, 26–30 October 2015; Bentz, C., Jäger, G., Yanovich, I., Eds.; University of Tübingen: Tübingen, Germany, 2016.
28. Futrell, R.; Mahowald, K.; Gibson, E. Quantifying word order freedom in dependency corpora. In Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015), Uppsala, Sweden, 24–26 August 2015; pp. 91–100.
29. Ackerman, F.; Malouf, R. Morphological organization: The low conditional entropy conjecture. *Language* **2013**, *89*, 429–464.
30. Milin, P.; Kuperman, V.; Kostic, A.; Baayen, R.H. Paradigms bit by bit: An information theoretic approach to the processing of paradigmatic structure in inflection and derivation. In *Analogy in Grammar: Form and Acquisition*; Blevins, J.P., Blevins, J., Eds.; Oxford University Press: Oxford, UK, 2009; pp. 214–252.
31. Levy, R. Expectation-based syntactic comprehension. *Cognition* **2008**, *106*, 1126–1177.
32. Jaeger, T.F. Redundancy and reduction: Speakers manage syntactic information density. *Cognit. Psychol.* **2010**, *61*, 23–62.
33. Fenk, A.; Fenk, G. Konstanz im Kurzzeitgedächtnis-Konstanz im sprachlichen Informationsfluß. *Z. Exp. Angew. Pshychol.* **1980**, *27*, 400–414.
34. Fenk-Oczlon, G. Familiarity, information flow, and linguistic form. In *Frequency and the Emergence of Linguistic Structure*; Bybee, J.L., Hopper, P.J., Eds.; John Benjamins: Amsterdam, The Netherlands, 2001; pp. 431–448.
35. Ferrer-i Cancho, R.; del Prado Martín, F.M. Information content versus word length in random typing. *J. Stat. Mech. Theory Exp.* **2011**, *L12002*, doi:10.1088/1742-5468/2011/12.
36. Ferrer-i Cancho, R.; Dębowski, Ł.; del Prado Martín, F.M. Constant conditional entropy and related hypotheses. *J. Stat. Mech. Theory Exp.* **2013**, *L07001*, doi:10.1088/1742-5468/2013/07/L07001.
37. Ferrer-i-Cancho, R.; Díaz-Guilera, A. The global minima of the communicative energy of natural communication systems. *J. Stat. Mech. Theory Exp.* **2007**, *2007*, P06009.
38. Ferrer-i-Cancho, R.; Solé, R.V. Least effort and the origins of scaling in human language. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 788–791.
39. Ferrer-i-Cancho, R. Zipf's law from a communicative phase transition. *Eur. Phys. J. B* **2005**, *47*, 449–457.
40. Ferrer-i-Cancho, R. The optimality of attaching unlinked labels to unlinked meanings. *Glottometrics* **2016**, *36*, 1–16.
41. Berger, A.L.; Pietra, V.J.D.; Pietra, S.A.D. A maximum entropy approach to natural language processing. *Comput. Linguist.* **1996**, *22*, 39–71.
42. Och, F.J.; Ney, H. Discriminative training and maximum entropy models for statistical machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 295–302.

43. Herbelot, A.; Ganesalingam, M. Measuring semantic content in distributional vectors. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Short Papers (ACL Short Papers 2013), Sofia, Bulgaria, 4–9 August 2013; pp. 440–445.
44. Padó, S.; Palmer, A.; Kisselew, M.; Šnajder, J. Measuring Semantic Content To Assess Asymmetry in Derivation. In Proceedings of the Workshop on Advances in Distributional Semantics, London, UK, 14 April 2015.
45. Santus, E.; Lenci, A.; Lu, Q.; Im Walde, S.S. Chasing Hypernyms in Vector Spaces with Entropy. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, 26–30 April 2014; pp. 38–42.
46. Resnik, P. Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of the IJCAI'95 Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 20–25 August 1995; Volume 1, pp. 448–453.
47. Boon, M. Information density, Heaps' Law, and perception of factiness in news. In Proceedings of the ACL Workshop on Language Technology and Computational Social Science (ACL LACSS 2014), Baltimore, MD, USA, 26 June 2014; p. 33.
48. Stetson, P.D.; Johnson, S.B.; Scotch, M.; Hripcsak, G. The sublanguage of cross-coverage. *Proc. AMIA Symp.* **2002**, *2002*, 742–746.
49. McFarlane, D.J.; Elhadad, N.; Kukafka, R. Perplexity analysis of obesity news coverage. *AMIA Annu. Symp. Proc.* **2009**, *2009*, 426–430.
50. Zhang, Y.; Kordoni, V.; Villavicencio, A.; Idiart, M. Automated multi-word expression prediction for grammar engineering. In Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, Association for Computational Linguistics, Sydney, Australia, 23 July 2006; pp. 36–44.
51. Ramisch, C.; Schreiner, P.; Idiart, M.; Villavicencio, A. An evaluation of methods for the extraction of multi-word expressions. In Proceedings of the LREC Workshop-Towards a Shared Task for Multiword Expressions (MWE 2008), Marrakech, Morocco, 1 June 2008; pp. 50–53.
52. Nemenman, I.; Shafee, F.; Bialek, W. Entropy and inference, revisited. *Adv. Neural Inf. Process. Syst.* **2002**, *1*, 471–478.
53. Kalimeri, M.; Constantoudis, V.; Papadimitriou, C.; Karamanos, K.; Diakonos, F.K.; Papageorgiou, H. Entropy analysis of word-length series of natural language texts: Effects of text language and genre. *Int. J. Bifurc. Chaos* **2012**, *22*, 1250223.
54. Koehn, P. Europarl: A parallel corpus for statistical machine translation. In Proceedings of the tenth Machine Translation Summit, Phuket, Thailand, 12–16 September 2005; Volume 5, pp. 79–86.
55. Mayer, T.; Cysouw, M. Creating a massively parallel Bible corpus. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, 26–31 May 2014; Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., Eds.; European Language Resources Association (ELRA): Paris, France, 2014; pp. 3158–3163.
56. Haspelmath, M. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguist.* **2011**, *45*, 31–80.
57. Wray, A. Why are we so sure we know what a word is? In *The Oxford Handbook of the Word*; Taylor, J., Ed.; Oxford University Press: Oxford, UK, 2014; Chapter 42.
58. Geertzen, J.; Blevins, J.P.; Milin, P. Informativeness of linguistic unit boundaries. *Ital. J. Linguist.* **2016**, *28*, 25–47.
59. Dębowski, Ł. Consistency of the plug-in estimator of the entropy rate for ergodic processes. In Proceedings of the 2016 IEEE International Symposium on Information Theory (ISIT), Barcelona, Spain, 10–15 July 2016; pp. 1651–1655.
60. Jiao, J.; Venkat, K.; Han, Y.; Weissman, T. Minimax estimation of functionals of discrete distributions. *IEEE Trans. Inf. Theory* **2015**, *61*, 2835–2885.
61. Lesne, A.; Blanc, J.L.; Pezard, L. Entropy estimation of very short symbolic sequences. *Phys. Rev. E* **2009**, *79*, 046208.
62. Basharin, G.P. On a statistical estimate for the entropy of a sequence of independent random variables. *Theory Probab. Appl.* **1959**, *4*, 333–336.
63. Hausser, J.; Strimmer, K. Entropy inference and the James–Stein estimator, with application to nonlinear gene association networks. *J. Mach. Learn. Res.* **2009**, *10*, 1469–1484.

64. Shannon, C.E.; Weaver, W. *The Mathematical Theory of Communication*; The University of Illinois Press: Urbana, IL, USA, 1949.
65. Harris, B. The statistical estimation of entropy in the non-parametric case. In *Topics in Information Theory*; Csiszar, I., Ed.; North-Holland: Amsterdam, The Netherlands, 1975; pp. 323–355.
66. Chomsky, N. *Aspects of the Theory of Syntax*; MIT Press: Cambridge, MA, USA, 1965.
67. Baroni, M. Distributions in text. In *Corpus Linguistics. An international handbook*; Lüdeling, A., Kytö, M., Eds.; De Gruyter: Berlin, Germany, 2009; pp. 803–821.
68. Ferrer, R.; Solé, R.V. Two Regimes in the Frequency of Words and the Origins of Complex Lexicons: Zipf 's Law Revisited. *J. Quant. Linguist.* **2001**, *8*, 165–173.
69. Gerlach, M.; Altmann, E.G. Stochastic model for the vocabulary growth in natural languages. *Phys. Rev. X* **2013**, *3*, 021006.
70. Petersen, A.M.; Tenenbaum, J.; Havlin, S.; Stanley, H.E.; Perc, M. Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Sci. Rep.* **2012**, *2*, 943.
71. Manning, C.D.; Schütze, H. In *Foundations of Statistical Natural Language Processing*; MIT Press: Cambridge, MA, USA, 1999; Chapter 5.
72. Montemurro, M.; Pury, P.A. Long-range fractal correlations in literary corpora. *Fractals* **2002**, *10*, 451–461.
73. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley and Sons: Hoboken, NJ, USA, 2006.
74. Ziv, J.; Lempel, A. A universal algorithm for sequential data compression. *IEEE Trans. Inf. Theory* **1977**, *23*, 337–343.
75. Ziv, J.; Lempel, A. Compression of individual sequences via variable-rate coding. *IEEE Trans. Inf. Theory* **1978**, *24*, 530–536.
76. Hausser, J.; Strimmer, K. *Entropy: Estimation of Entropy, Mutual Information and Related Quantities*; R Package Version 1.2.1, 2014. Available online: https://CRAN.R-project.org/package=entropy (accessed on 13 June 2017).
77. Bates, D.; Mächler, M.; Bolker, B.; Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **2015**, *67*, 1–48.
78. Hammarström, H.; Forkel, R.; Haspelmath, M.; Bank, S. *Glottolog 2.7*; Max Planck Institute for the Science of Human History: Jena, Germany, 2016.
79. Kirby, S.; Tamariz, M.; Cornish, H.; Smith, K. Compression and communication in the cultural evolution of linguistic structure. *Cognition* **2015**, *141*, 87–102.
80. Ferrer-i-Cancho, R. Optimization models of natural communication. *arXiv* **2014**, arXiv:1412.2486.
81. Bentz, C.; Ruzsics, T.; Koplenig, A.; Samaržić, T. A comparison between morphological complexity measures: typological data vs. language corpora. In Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), 26th International Conference on Computational Linguistics, Osaka, Japan, 11 December 2016.
82. Baayen, H.R. *Word Frequency Distributions*; Kluwer: Dordrecht, The Netherlands; Boston, MA, USA; London, UK, 2001.
83. Agresti, A.; Hitchcock, D.B. Bayesian inference for categorical data analysis. *Stat. Methods Appl.* **2005**, *14*, 297–330.
84. Chao, A.; Shen, T.J. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environ. Ecol. Stat.* **2003**, *10*, 429–443.

© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).