Peer-reviewed version available at *Entropy* **2017**, *19*, , 275; doi:10.3390/e19060275

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; chris@christianbentz.de
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, United Kingdom
4 Forschungszentrum Deutscher Sprachatlas, Philipps-Universität Marburg, Pilgrimstein 16, D-35032 Marburg
5 Complexity and Quantitative Linguistics Lab, LARCA Research Group, Departament de Ciències de la Computació, Universitat Politècnica de Catalunya, Barcelona, Catalonia, Spain
* Correspondence: chris@christianbentz.de

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 three main results: 1) a text size of 50K tokens is sufficient for word entropies to stabilize throughout the text, 2) across languages of the world, word entropies display a unimodal distribution that is skewed to the right. This suggests that there is a trade-off between the learnability and expressivity of words across languages of the world. 3) There is a strong linear relationship between unigram entropies and entropy rates, suggesting that they are inherently linked. We discuss the implications of these results for studying the diversity and evolution of languages from an information-theoretic point of view.

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 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]. For instance, the usage of complementizers such as "that" 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 references therein). For a critical review of some problematic aspects of the UID hypothesis see [35] and [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 multiword expressions [50,51].

All 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 3.2.4), and the *entropy rate* per word according to Gao *et al.* [5] (Section 3.2.5). 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 3.2.5) – 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 languages. 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 three major points:

1. Both unigram entropies and entropy rates reach stable values at around 50-100K word tokens across different languages. Hence, we do not need massive corpora to meaningfully compare languages with regards to their word entropies – as long as we keep the content of texts constant (Section 5.1).
2. Across languages of the world, unigram entropies display a unimodal distribution around a mean of ca. 9 bits/word, with a standard deviation of ca. 1 bit/word. Entropy rates display a unimodal distribution around a mean of ca. 6 bits/word, with a standard deviation of ca. 1 bit/word. In both cases, the distribution is skewed to the right. Hence, there seems to be a strong pressure to keep the mass of languages in a relatively narrow entropy range, especially for the difference between unigram entropy and entropy rate (Section 5.2). This is in line with earlier findings [8,9]. There are no languages with a unigram entropy of less

Table 1. Information on the parallel corpora used.

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

*in number of tokens

than 6 bits/word (if estimated based on 50K tokens). This observation is in agreement with optimization models of communication, where entropy is regarded as a cost to minimize in conflict with mutual information maximization. Zipf's law for word frequencies emerges in a critical balance between these two forces [39]. We further argue that this trade-off reflects two basic pressures on natural communication systems: *learnability* vs. *expressivity*.

3. There is a strong positive linear relationship between unigram entropies and entropy rates ($r = 0.96, p < 0.0001$). This allows us to formulate a simple linear model that predicts one from the other (Section 5.3). Given that unigram entropy estimation is computationally much more efficient than entropy rate estimation, this finding can help to reduce computational costs.

Finally, we argue that information theoretic properties like word entropy are not only relevant for computational linguistics and quantitative linguistics, but that they 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],[1] 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. 30M),[2] whereas the PBC and the UDHR are smaller (ca. 290K and 1.3K 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).

3. Theory

3.1. Word types and tokens

The basic information encoding unit chosen in this study is the word.[3] 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:

---

1 Last accessed on 02/06/2016
2 Though we only use around 1M tokens for each language, since this is enough for the stabilization analyses.
3 Note that earlier studies on the entropy of English [2,4,6] often chose characters instead.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2017 doi:10.20944/preprints201704.0180.v1

Peer-reviewed version available at Entropy 2017, 19, , 275; doi:10.3390/e19060275


```
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 ($\approx 0.01\%$). 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 taken as a given in most corpus and computational linguistic studies, they are not necessarily uncontroversial from a linguistically more informed point of view. Haspelmath [56] and Wray [57] point out that there is a whole range of orthographic, phonetic and distributional definitions for the concept "word", which can yield different results for specific 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. 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 4 times, *and* occurs 3 times, etc. As a simple approximation, $p(w_i)$ can be estimated via the so-called *maximum likelihood* method [59]:

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

where the denominator is the overall number of word tokens. 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 $\mathcal{V} = \{w_1, w_2, ..., w_V\}$, with vocabulary size $V = |\mathcal{V}|$, 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 [60]

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

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 *plug-in* estimator – which is obtained replacing $p(w)$ by $\hat{p}(w)$ (Eq. 2). For instance, our example "corpus" yields:

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

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 [61,62]. A range of more advanced entropy estimators have been proposed to overcome this limitation [6,52,59,63]. These are outlined in more detail in Appendix B, and tested with our parallel corpus data.

The conditions (e.g. the sample size) under which the entropy of a source can be estimated reliably are a living field of research ([64] and references therein). Proper estimation of entropy requires that certain conditions are met. Stationarity and ergodicity are typically named as such [4] (e.g., [63,64]). Another condition is finiteness (a finite set of 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., [63,64]). A fourth usual condition is the assumption that types are independent and identically distributed (i.i.d.) [61,63,64]. 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. In the following, we review two problems relating to the stationarity and the i.i.d. assumptions.

3.2.1. 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. Even if we captured the whole set of linguistic interactions of a speaker population in a 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 [65][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 [66–68]. Thus, we never actually sample the whole set of word types of a language. However, the concentration of tokens on a core vocabulary [67,69] potentially alleviates this problem. Still, word entropy estimation is a harder problem than entropy estimation for characters or phonemes types, since the latter have a repertoire that is finite and usually small.

3.2.2. Problem 2: Short- and long-range correlations between words

In natural languages, we find co-occurrence patterns implying 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

[4] Stationarity means that the statistical properties of blocks of words of some length do not depend on their position in the text sequence. Ergodicity means that statistical properties of a sufficiently long text sequence matches the average properties of the ensemble of all possible text sequences.

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 greater than expected by chance [70]. Collocations are examples of *short-range correlations* in text sequences. Text sequences also exhibit *long-range correlations*, i.e. correlations between types that are far away in the text [8,71].

3.2.3. Our perspective

We are fully aware of both Problem 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.

3.2.4. 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

$$H_n(T) = -\sum_{i=1}^{\mathcal{V}_n} p(w_i, w_{i+1}, ..., w_n) \times \log_2 p(w_i, w_{i+1}, ..., w_n), \tag{5}$$

where n is the block size, and $\mathcal{V}_n$ is 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 & Grassberger [6] use an English corpus of 70M words, and assert that entropy estimation beyond a block size of 5 characters (not words) is already unreliable. Our strategy is to stick with block sizes of 1, 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].

3.2.5. 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 [72, p.74], i.e.

$$h(T) = \lim_{N\to\infty} = \frac{1}{N} H_n(T) \tag{6}$$

$$= \frac{1}{N} H(t_1, t_2, ..., t_N), \tag{7}$$

where $t_1, t_2, ..., t_N$ is a block of consecutive tokens of length N. Given stationarity, this is equivalent to [72, p.75]

$$h(T) = \lim_{N\to\infty} H(t_N | t_1, t_2, ..., t_{N-1}). \tag{8}$$

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. So $h(T)$ accounts for *all statistical dependencies* between tokens [63]. Note that in the limit, i.e. as block size n approaches

infinity, the *block* entropy per token converges to the entropy *rate*. Also, for an independent and identically distributed (i.i.d) random variable, the block entropy of block size 1, i.e. $H_1(T)$, is identical to the entropy rate [63]. 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 & Lempel [73,74] 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* [72], is efficient in terms of convergence. The conditions for this estimator are stationarity and ergodicity of the process that generates a text T.

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

$$L_i = 1 + \max\{0 \leq l \leq i : s_i^{i+l-1} = s_j^{j+l-1} \text{ for some } 0 \leq j \leq i-1\}. \tag{9}$$

This is an adaptation of Gao et al.'s [5] match-length definition.[5] To illustrate this, take the example from above again:

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

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

$$\hat{h}(T) = \frac{1}{N} \sum_{i=2}^{N} \frac{\log_2 i}{L_i}, \tag{10}$$

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 10.

4. Methods

We estimate *unigram entropies* – i.e. entropies for block sizes of 1 ($H_1(T)$) – with nine different estimation methods, using the *R* package *entropy* [75], and the Python implementation of the *Nemenman-Shafee-Bialek* (NSB) [52] estimator. 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 [59].[6] 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*.[7] This was inspired by an earlier implementation by Montemurro & Zanette [8] of Kontoyiannis et al.'s [4] estimator.

As one estimates 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

---

[5] Note that Gao et al. [5] give a more general definition that also holds for the so-called *sliding window*, or *LZ77* estimator.
[6] https://gist.github.com/shhong/1021654/
[7] https://github.com/dimalik/EntropyEstimator

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 ($\alpha = 0.1$). If L is the number of tokens, SD is defined as the standard deviation that is calculated over entropies obtained with prefixes of lengths

$$L, L + 1K, L + 2K, ..., L + 10K. \quad (11)$$

K represents the number 1000 here. We say that entropies have stabilized when SD$< \alpha$, where α is the threshold. Notice that this is indeed a local stabilization criterion. Here we choose L 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.

Note that both [8] and [10] use a more coarse-grained stabilization criterion. Namely, 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.

5. Results

We first assess the minimum text sizes at which both *unigram entropies* and *entropy rates* stabilize, i.e. SD<0.1, using a subset of 21 languages (Section 5.1) of the European Parliament Corpus (EPC). The EPC was chosen for this task since it is the largest in terms of number of tokens, ranging up to several millions. We then select texts from the full Parallel Bible Corpus (PBC) which have enough tokens, estimate their entropies, and compare their spread on an entropy spectrum (Section 5.2). The *Universal Declaration of Human Rights* (UDHR) is used in Appendix D and Appendix F to illustrate that there are strong correlations between unigram entropies for different estimators and different corpora. Finally, in Section 5.3, we investigate the correlation and linear relationship between *unigram entropies* and *entropy rates*.

5.1. Entropy stabilization throughout the text sequence

5.1.1. Unigram entropies

Figure 1 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 100K 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 [59].

However, overall the stabilization behaviour of entropy estimators is similar across all languages. Namely, in the range of ca. 0 to 25K tokens, they all undergo a fast growth, and start to stabilize after 25K. This stabilization behaviour is visualized in Figure 2. 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 >50K 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 Figure 1 and 2 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 1. See Appendix D for details.

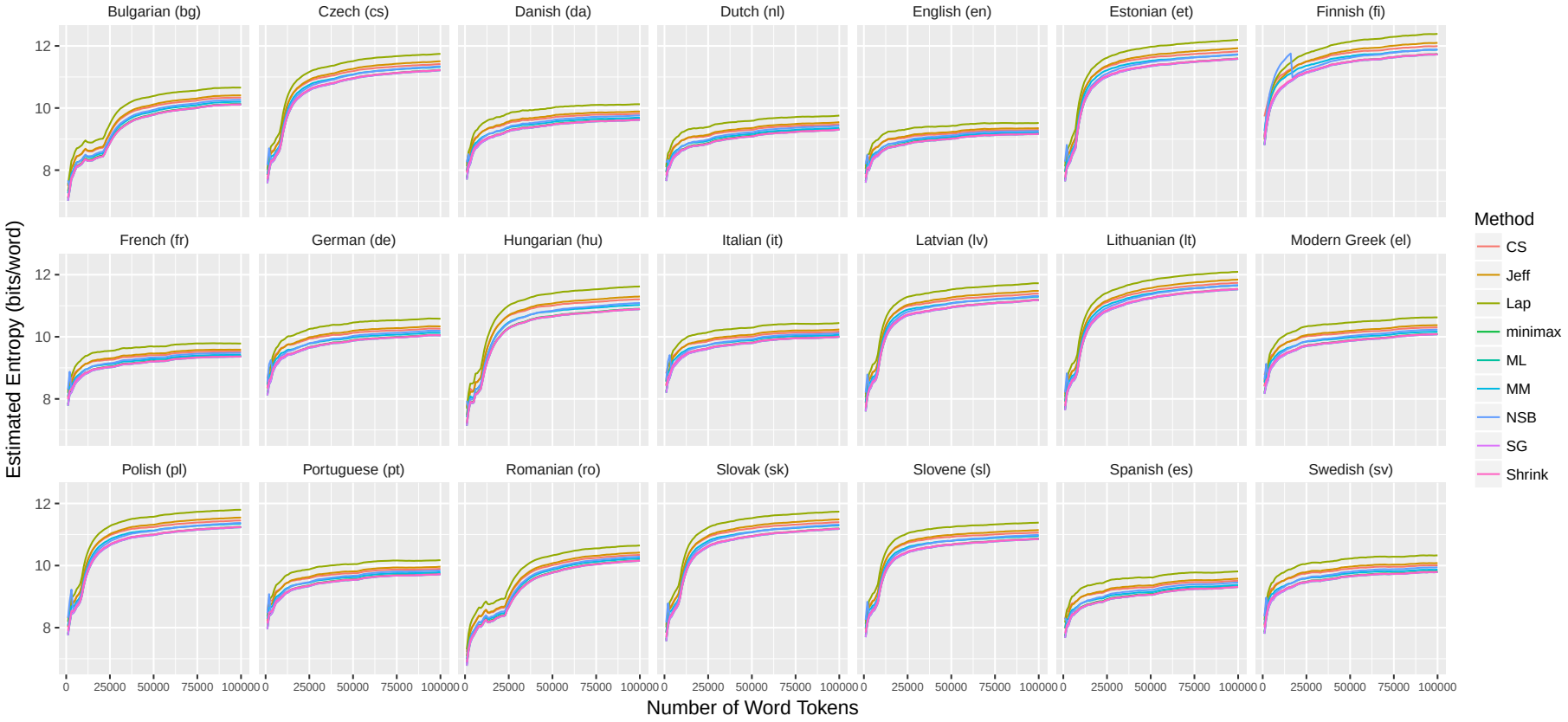

Figure 1. Unigram entropies (y-axis) as a function of text length (x-axis) across 21 languages of the EPC corpus. Unigram entropies are estimated on prefixes of the text sequence increasing by 1K tokens. Thus, the first prefix covers tokens 1 to 1K, the second prefix covers tokens 1 to 2K, etc. The number of tokens is limited to 100K, since entropy values already (largely) stabilize throughout the text sequence before that. Hence, there are 100 points along the x-axis. 9 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.

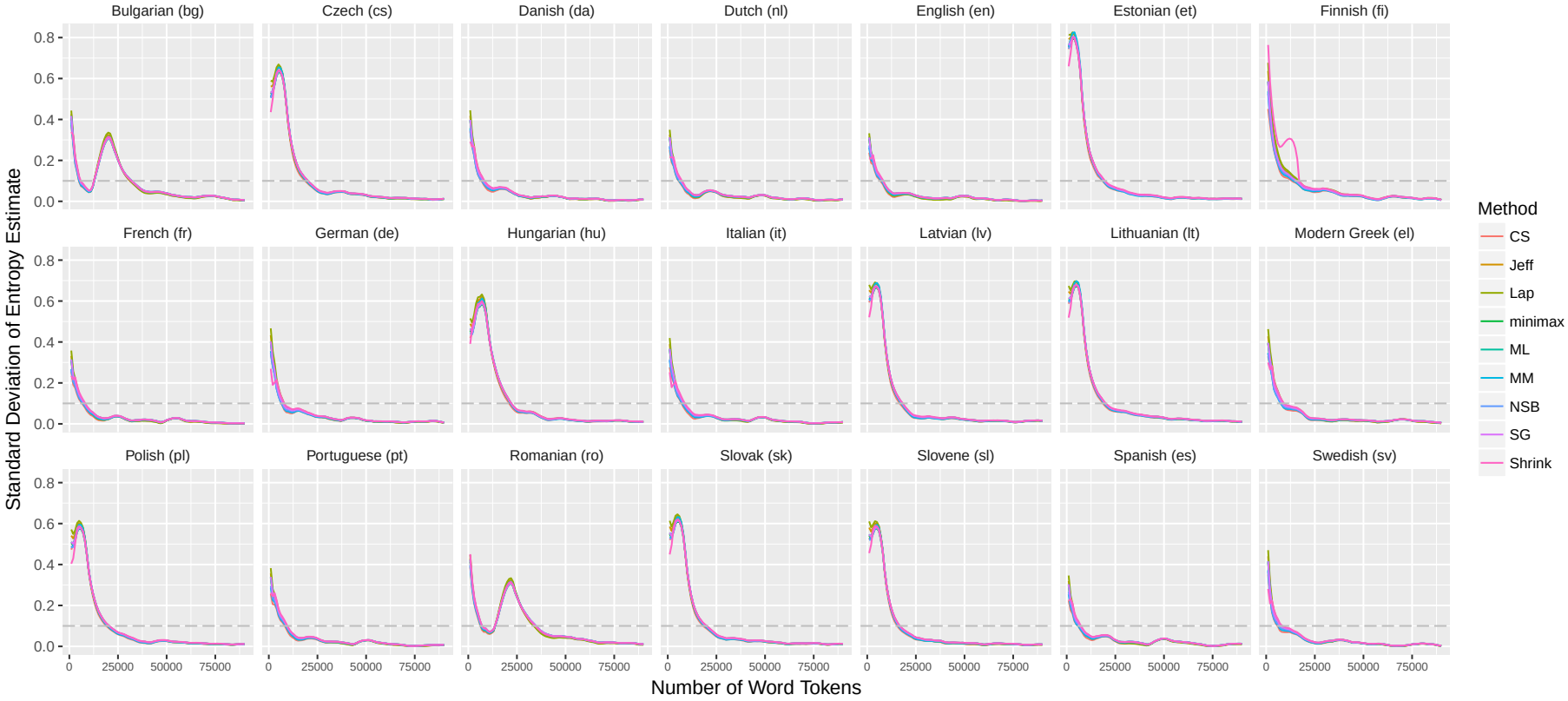

Figure 2. SDs of unigram entropies (y-axis) as a function of text length (x-axis) across 21 languages of the EPC corpus, and the 9 different estimators. Unigram entropies are estimated on prefixes of the text sequence increasing by 1K tokens as in Fig. 1. SDs are calculated over the entropies of the 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.

5.1.2. Entropy rate

The results for *entropy rates* ($\hat{h}(T)$) are visualized in Figure 3 and 4. The stabilization points and SDs are similar to the ones established for unigram entropies. Namely, below 25K 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 5 bits/word at 5K tokens, and then steeply rise to approximately 7 bits/word at 25K tokens.

In parallel to the results for unigram entropies, the 21 languages reach stable entropy values at around 50K. This is visible in Figure 4. 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 C).

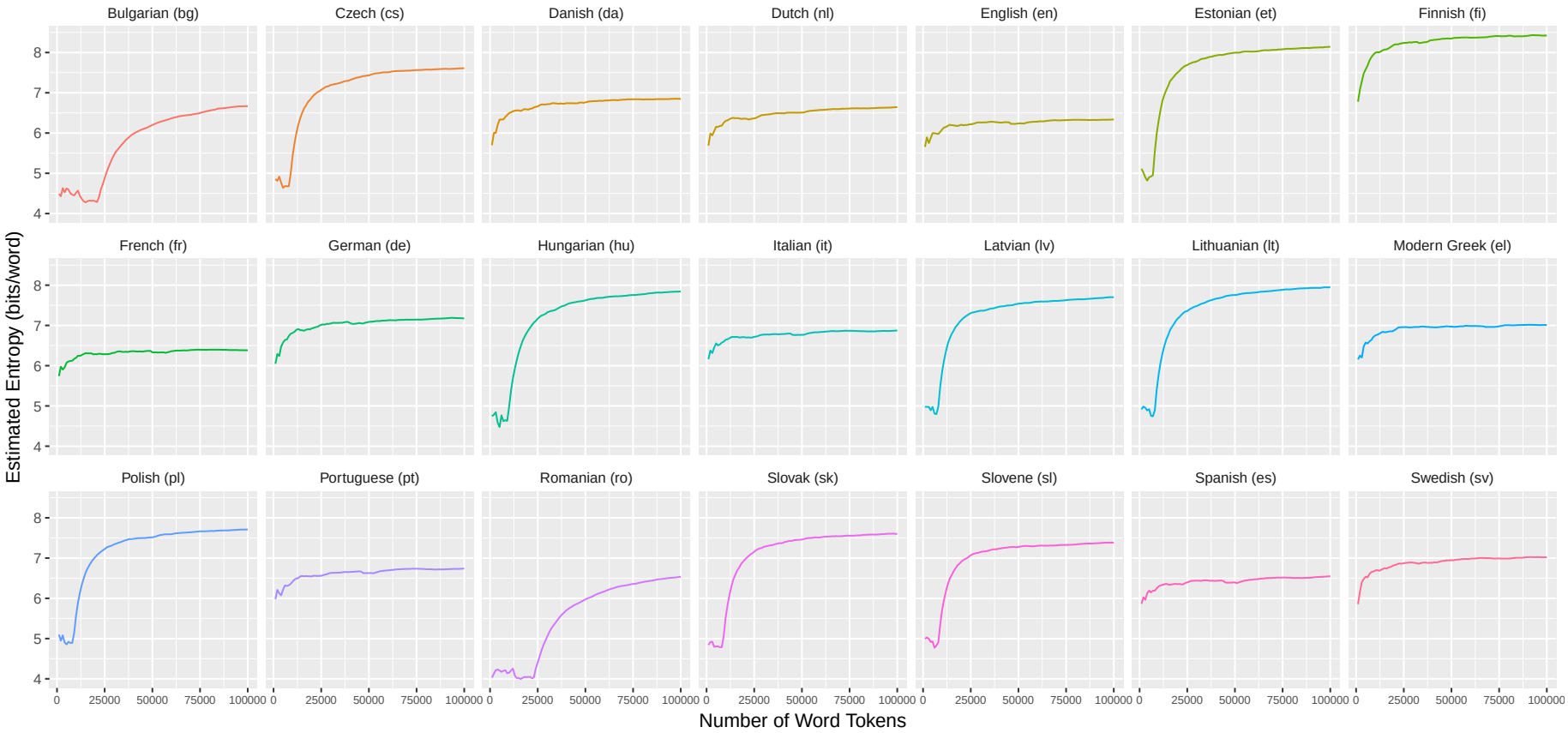

Figure 3. Entropy rates as a function of text length across 21 languages of the EPC. Entropy rates are estimated on prefixes of the text sequence increasing by 1K tokens as in Fig. 1. Hence there are 100 points along the x-axis. The language identifiers used by the EPC are given in parenthesis.

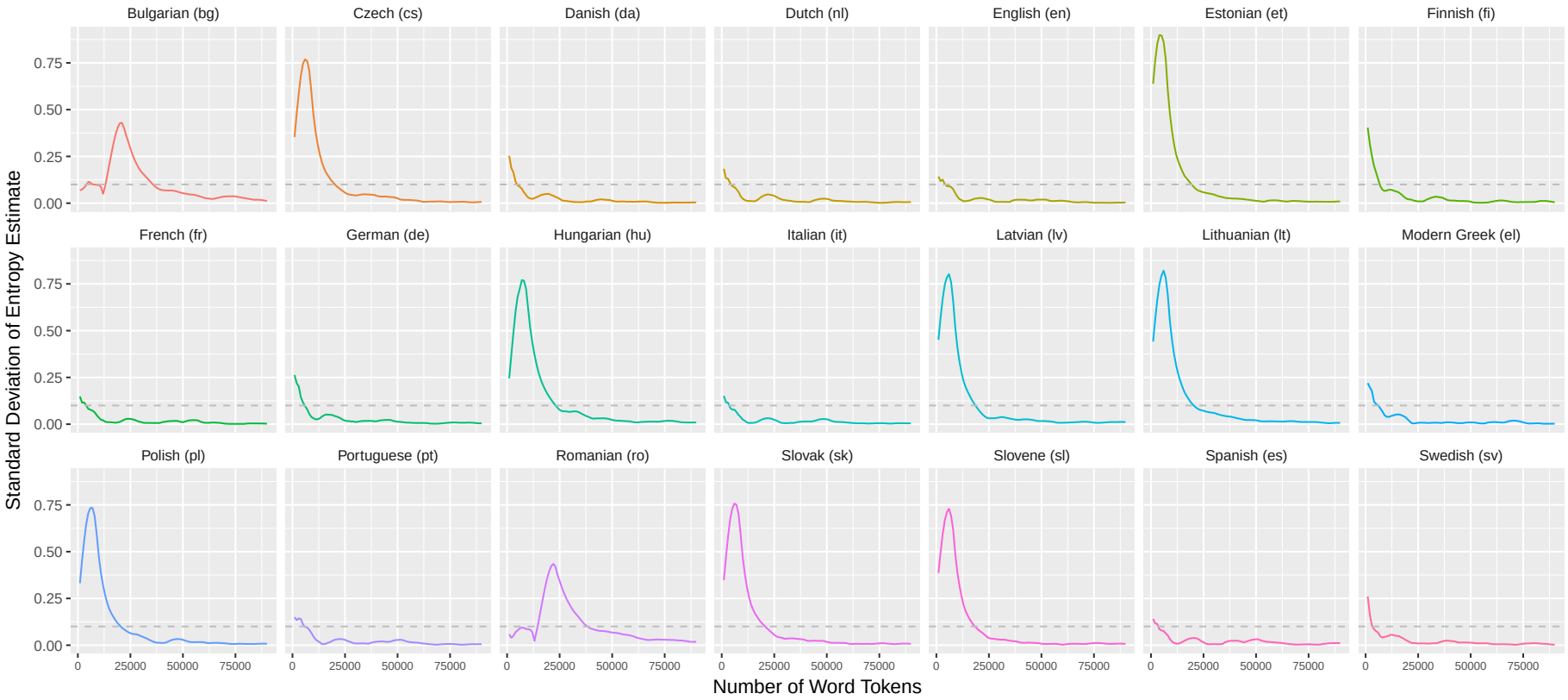

Figure 4. SDs of entropy rates as a function of text length across 21 languages of the EPC corpus. The format is the same is as in Fig. 2.

5.2. Word entropies across more than 1000 languages

To estimate word entropies across languages of the world, we need to revert to a corpus with much wider cross-linguistic coverage than the EPC. The Parallel Bible Corpus (PBC) serves this purpose. Based on our stabilization analyses, we choose a conservative cut-off point: only texts with at least 50K tokens are included. Of these, in turn, we take the first 50K 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.

Figure 5 a) shows a density plot of the estimated entropic values across all texts and languages. Both unigram entropies and entropy rates show a unimodal distribution that is skewed to the right. The distribution of entropy differs clearly 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 [61].

As we would expect, Figure 5 also shows that entropy rates are systematically lower than unigram entropies, since the former take co-text information into account, whereas the latter do not. 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 8), while the unigram entropy is not conditioned on the co-text. As conditioning reduces entropy [72], it is not surprising that entropy rates tend to fall below unigram entropies. Unigram entropies are distributed around a mean of 9.14 ($SD = 1.12$), whereas entropy rates have a lower mean of 5.97 ($SD = 0.91$). Figure 5 shows 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.

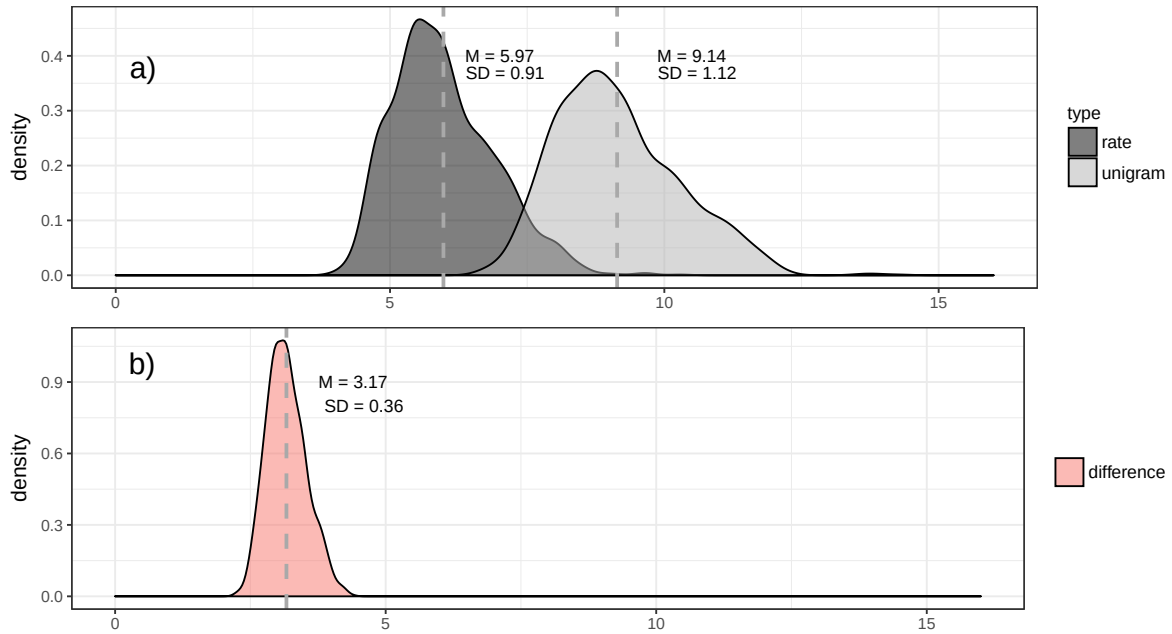

Figure 5. The distribution of entropic measures in bits. a) Probability density of unigram entropies (light grey) and entropy rates (dark grey) across texts of the PBC (using 50K tokens). M and SD are, respectively, the mean and the standard deviation of the values. A vertical dashed line indicates the mean M. b) The same for the difference between unigram entropies and entropy rates.

To further illustrate the diversity of entropies across languages of the world, Figure 6 gives a map with unigram entropies, entropy rates and the difference between them for the PBC sample. 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2017 doi:10.20944/preprints201704.0180.v1

Peer-reviewed version available at Entropy 2017, 19, , 275; doi:10.3390/e19060275

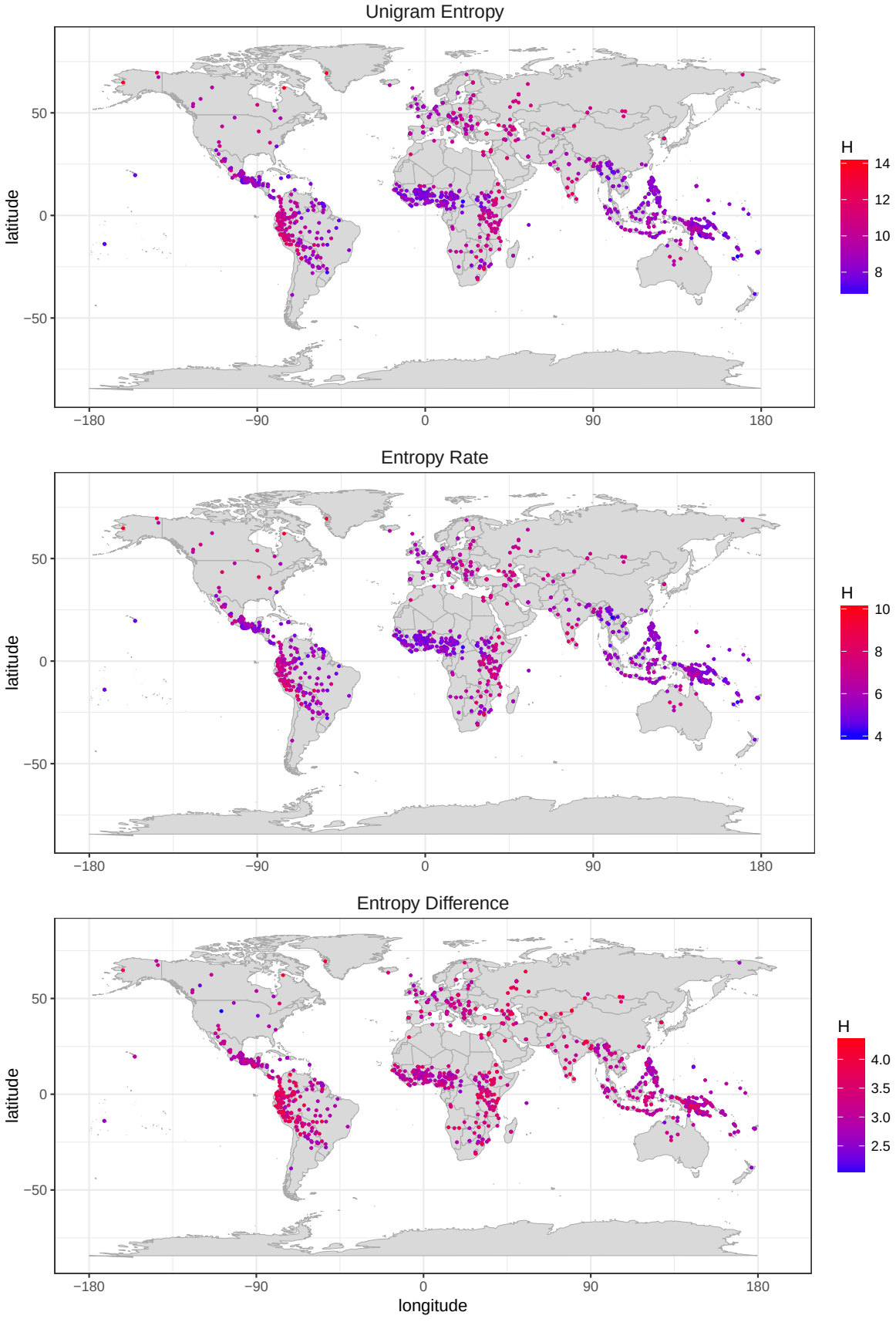

Figure 6. World maps with unigram entropies (upper panel), entropy rates (middle panel), and the difference between them (lower panel), across texts of the PBC (using 50K tokens), amounting to 1499 texts and 1115 languages.

the world. For example, languages in the Andean region of South America all have high unigram entropies (bright red). 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 [76]. 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). 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 more narrowly distributed (as seen also in Figure 5).

5.3. Correlation between unigram entropy and entropy rate

The density plot for unigram entropies in Figure 5 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 7 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 30K tokens onwards) as shown in Fig. 8. The fact that this holds for all entropy estimators is not surprising as they are strongly correlated (Appendix D).

In Appendix E we also give correlations between unigram entropy values for all 9 estimation methods and entropy rates. The Pearson correlations are strong (between $r = 0.95$ and $r = 0.99$, with $p < 0.0001$). This illustrates empirically that despite the conceptual difference between unigram entropies and entropy rates, there is a strong underlying connection between them. Namely, a linear model fitted through the points in Figure 7 (left panel) can be specified as

$$\hat{h}(T) = -1.12 + 0.78\,\hat{H}_1(T). \tag{12}$$

Via Equation 12 we can convert unigram entropies into entropy rates with a variance of 0.035. More generally, word entropy rates $\hat{h}(T)$ and unigram entropies $\hat{H}_1(T)$ are linked by a linear relationship:

$$\hat{h}(T) = k_1 + k_2\,\hat{H}_1(T), \tag{13}$$

where k_1 and k_2 are constants.

The right panels of Figure 7 plot the same relationship between unigram entropies and entropy rates facetted by geographic macro areas taken from Glottolog 2.7 [77]. 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 these different areas of the world.

6. Discussion

6.1. Entropy estimation and stabilization

In Section 5.1, we set out to establish stabilization points for both unigram entropies and entropy rates. We have illustrated that – independent of the estimation method – texts of 50K tokens are sufficient to estimate values with standard deviations below 0.1. This was shown across 21 languages of the EPC. Similar results are found for 32 languages of different language families in Appendix C. Furthermore, Appendix D illustrates that there are generally strong correlations between different entropy estimation methods – despite their differing stabilization properties. Finally, Appendix F gives Pearson correlations between unigram entropies of the PBC, EPC, and UDHR. These are also strong (around 0.8) – despite the differences in registers and styles as well as text sizes. Overall, our analyses converge to support the claim that entropy rankings of languages are stable at 50K tokens.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2017 doi:10.20944/preprints201704.0180.v1

Peer-reviewed version available at Entropy 2017, 19, , 275; doi:10.3390/e19060275

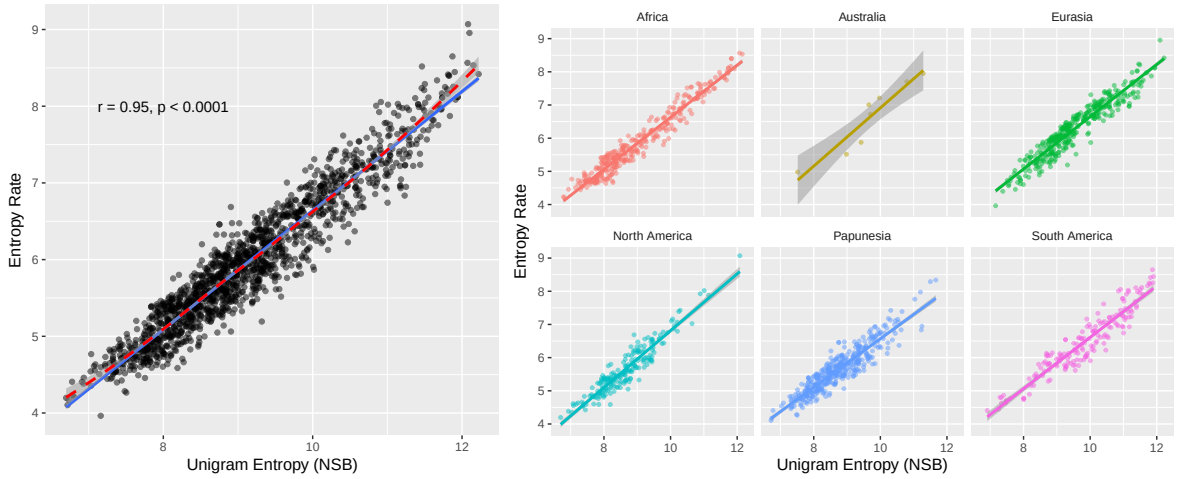

Figure 7. Linear relationship between unigram entropies approximated with the NSB estimator (x-axis) and entropy rates (y-axis) for 1495 PBC texts (50K tokens) across 1112 languages. Four texts (Ancient Hebrew (hbo), Eastern Canadian Inuktitut (ike), Kalaallisut (kal), and Northwest Alaska Inupiatun (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 facetted by macro areas across the world. Macro areas are taken from Glottolog 2.7 [77]. Linear regression models are given as coloured lines with 95% confidence intervals.

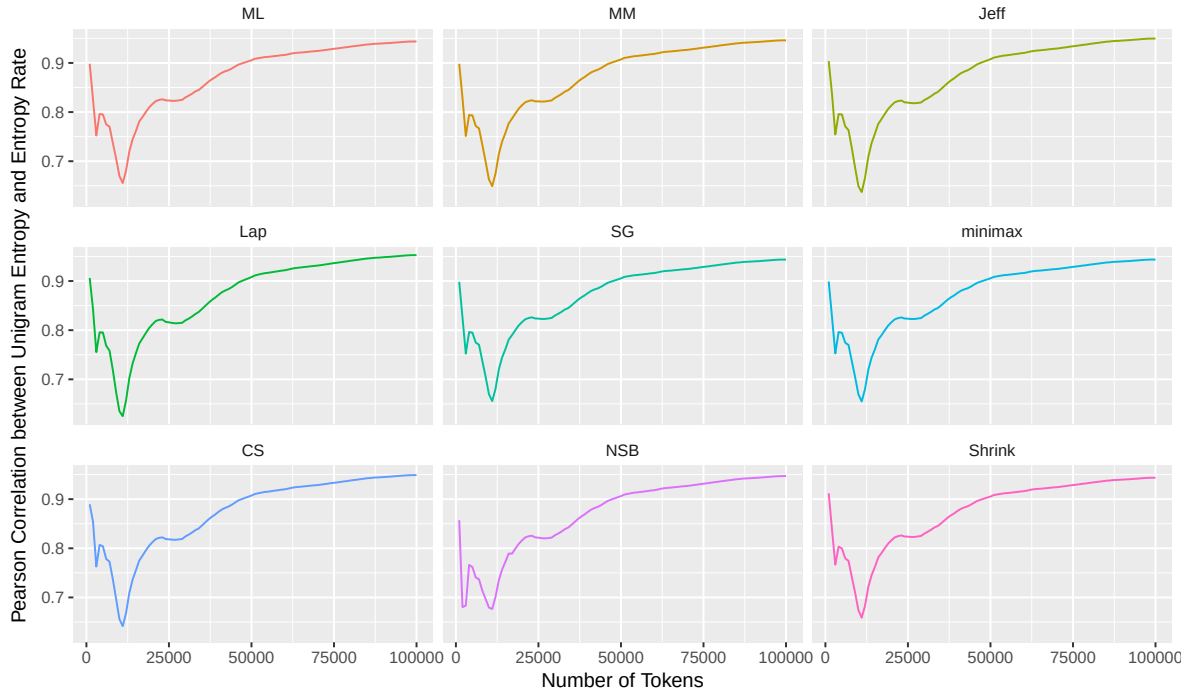

Figure 8. Pearson correlation between unigram entropy and entropy rate (y-axis) as a function of text length (x-axis). Each correlation is calculated over the 21 languages of the EPC corpus. All nine unigram entropy estimators are considered.

6.2. 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 9 bits/word, while entropy rates fall closer to a mean of 6 bits/word. It is remarkable that given the wide range of potential entropies – from 0 to ca. 14 – most natural languages fall on a relatively narrow spectrum.[8] Unigram entropies mainly fall in the range between 7 to 12 bits/word, and entropy rates in the range between 4 to 9 bits/word. Thus, each only covers around 40% of the scale.

This suggests that there are pressures at play which keep word entropies in a relatively narrow range. We argue that these pressures are 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 narrow range between these extremes. Kirby *et al.* [78] illustrate how this trade-off effects the evolution of languages by using iterated learning experiments and computational simulations.

The 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]. Entropy minimization is linked with learnability: fewer word forms are easier to learn (see [79] 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]

$$MI \leq H \tag{14}$$

The lower the entropy, the lower the potential for expressivity. This may shed light on the right-skeweness of the entropy distribution in Figure 5. 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 Eq. 14, though this comes with a learnability cost.

Further support for pressure to increase entropy to warrant sufficient expressivity (Eq. 14) comes from the fact that there is no language with less than 6 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 8 bits/word in the PBC.

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 entropies around 14 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 are equally desirable features. Any

[8] It is non-trivial to find an upper limit for the maximum word entropy of natural languages. In theory, word entropy 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*.

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*.

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]. Developing and testing entropy estimation methods is the necessary foundation to further unravel the factors and causes of word entropy evolution and change.

6.3. Correlation between unigram entropies and entropy rates

Finally, in Section 5.3, we found a strong correlation between unigram entropies and entropy rates. This is somewhat 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. But 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 8 and 75 languages respectively – that the difference between word entropy rates for texts *with randomized word order* and those of text *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 [72, 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. Also, we have shown that the entropy difference has a smaller variance than the original unigram entropies and entropy rates, again consistent with Montemurro and Zanette's findings (Fig. 5).

As a practical result, 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 (Fig. 7). Interestingly, we have shown that this linearity of the relationship increases as text length increases (Fig. 8). A mathematical investigation of the origins of this linear relationship should be the subject of future research.

Note that 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 12 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 using the European Parliament Corpus. We illustrated that word entropies can be reliably estimated with text sizes of $>$ 50K. 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, *learnability* and *expressivity* seem to emerge as fundamental constraints on language variation.

Furthermore, we have shown that there is a strong linear relationship between block entropies and 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 information content of words are relatively similar regardless of the language we are looking at. 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 for helpful discussions. CB 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. RFC 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: CB, DA and RFC conceived and designed the experiments; CB and DA performed the experiments; CB and DA analyzed the data; MC contributed materials; CB and RFC wrote the paper.

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 in the PBC and UDHR make cross-linguistic analyses of word types more reliable. This is especially helpful since we deal with hundreds of languages, representing 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 UTF-8 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

$$\hat{p}(w_i)^{ML} = \frac{f_i}{\sum_{j=1}^{V} f_j}, \tag{15}$$

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 15 into 3 yields the estimated entropy as [59]

$$\hat{H}^{ML} = -\sum_{i=1}^{V} \hat{p}(w_i)^{ML} \log_2(\hat{p}(w_i)^{ML}). \tag{16}$$

The ML method yields reliable results for situations where $N \gg V$, i.e. the number of tokens is much bigger than the number of types [59, 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 [66,80], 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 [see 59, for an overview]. Some of these are discussed in turn.

Appendix B.1 The Miller-Madow estimator

For example, the *Miller-Madow* (MM) estimator [59, p. 1471] tries to reduce the bias by adding a correction to the ML estimated entropy such that

$$\hat{H}^{MM} = \hat{H}^{ML} + \frac{M_{>0} - 1}{2N}, \tag{17}$$

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

$$\hat{p}(w_i)^{Bayes} = \frac{f_i + a_i}{N + A}, \tag{18}$$

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 [81, p. 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 [59, 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 & 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 [82, 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

$$\hat{C} = 1 - \frac{m_1}{N}, \tag{19}$$

where m_1 is the number of types with frequency 1 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 1. 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

$$\hat{p}(w_i)^{GT} = \left(1 - \frac{m_1}{N}\right) \hat{p}(w_i)^{ML}. \tag{20}$$

Furthermore, Chao and Shen [82, 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

$$\hat{H}^{CS} = -\sum_{i=1}^{V} \frac{\hat{p}(w_i)^{GT} \log_2(\hat{p}(w_i)^{GT})}{1 - (1 - \hat{p}(w_i)^{GT})^N}. \tag{21}$$

Appendix B.4 The James-Stein shrinkage estimator

Finally, Hausser and Strimmer [59, 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

$$\hat{p}(w_i)^{shrink} = \lambda \hat{p}(w_i)^{target} + (1 - \lambda)\hat{p}(w_i)^{ML}, \tag{22}$$

where $\lambda \in [0, 1]$ is the shrinkage intensity and $\hat{p}(w_i)^{target}$ is the so-called "shrinkage target". Hausser and Strimmer [59, p. 1473] suggest to use the maximum entropy distribution as a target, i.e. $\hat{p}(w_i)^{target} = \frac{1}{V}$. This yields

$$\hat{p}(w_i)^{shrink} = \frac{\lambda}{V} + (1 - \lambda)\hat{p}(w_i)^{ML}. \tag{23}$$

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 23 yields

$$\hat{p}(w_i)^{shrink} = \frac{1}{V}, \tag{24}$$

i.e. uniform probabilities that will yield maximum entropy. In the lowest shrinkage case ($\lambda = 0$) Equation 23 yields

$$\hat{p}(w_i)^{shrink} = \hat{p}(w_i)^{ML}, \tag{25}$$

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 [59, 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

$$\hat{H}^{shrink} = -\sum_{i=1}^{V} \hat{p}(w_i)^{shrink} \log_2 \hat{p}(w_i)^{shrink}. \quad (26)$$

Overall, 9 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* [75].

Appendix C 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 (Figure 9 and 10). 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.

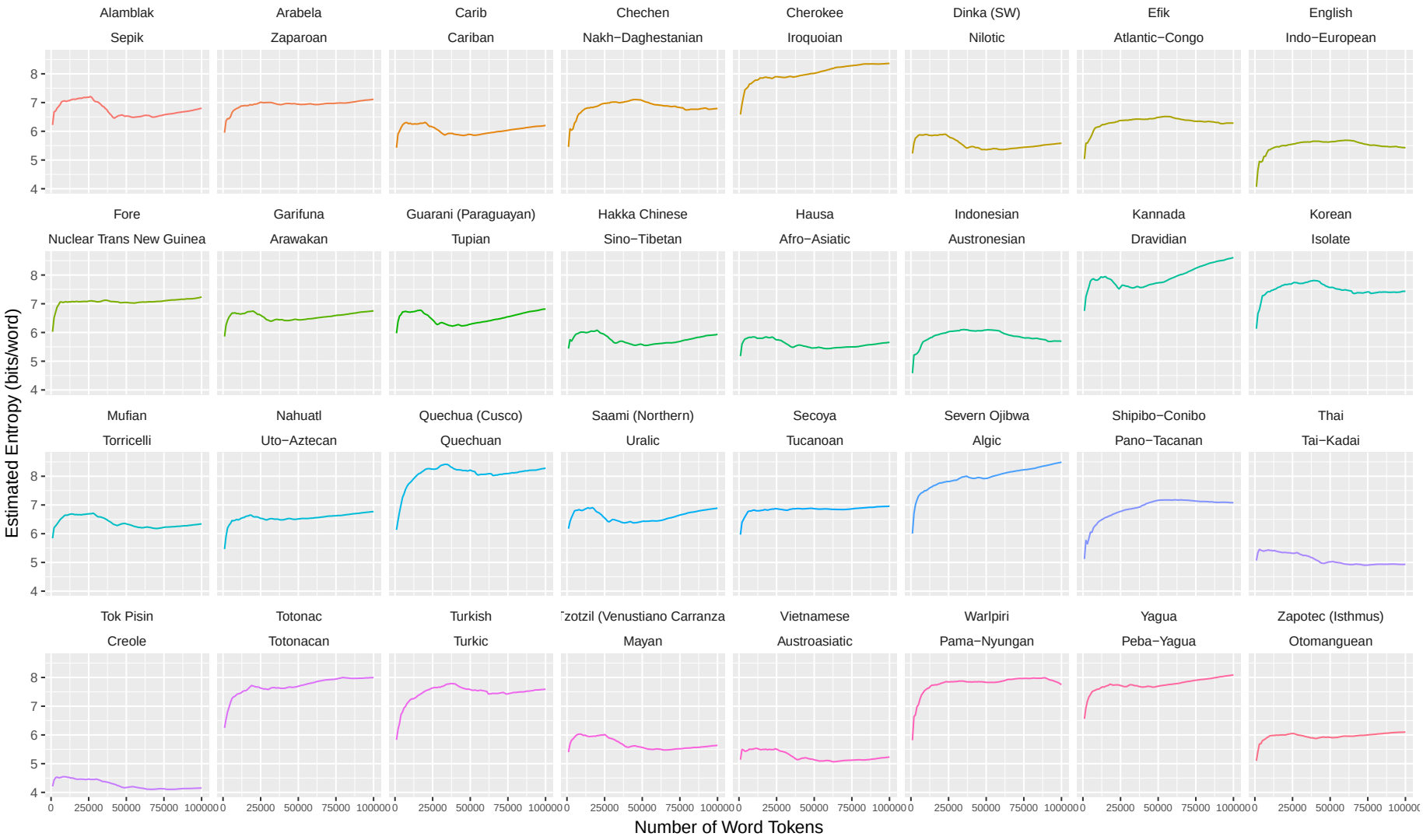

Figure 9. Entropy rates as a function of text length across 32 languages of the PBC. Languages were chosen to represent some of the major language families across the world. Entropy rates are estimated on prefixes of the text sequence increasing by 1K tokens as in Fig. 1. Hence there are 100 points along the x-axis. The language names and families are taken from Glottolog 2.7 [77] and given above the plots.

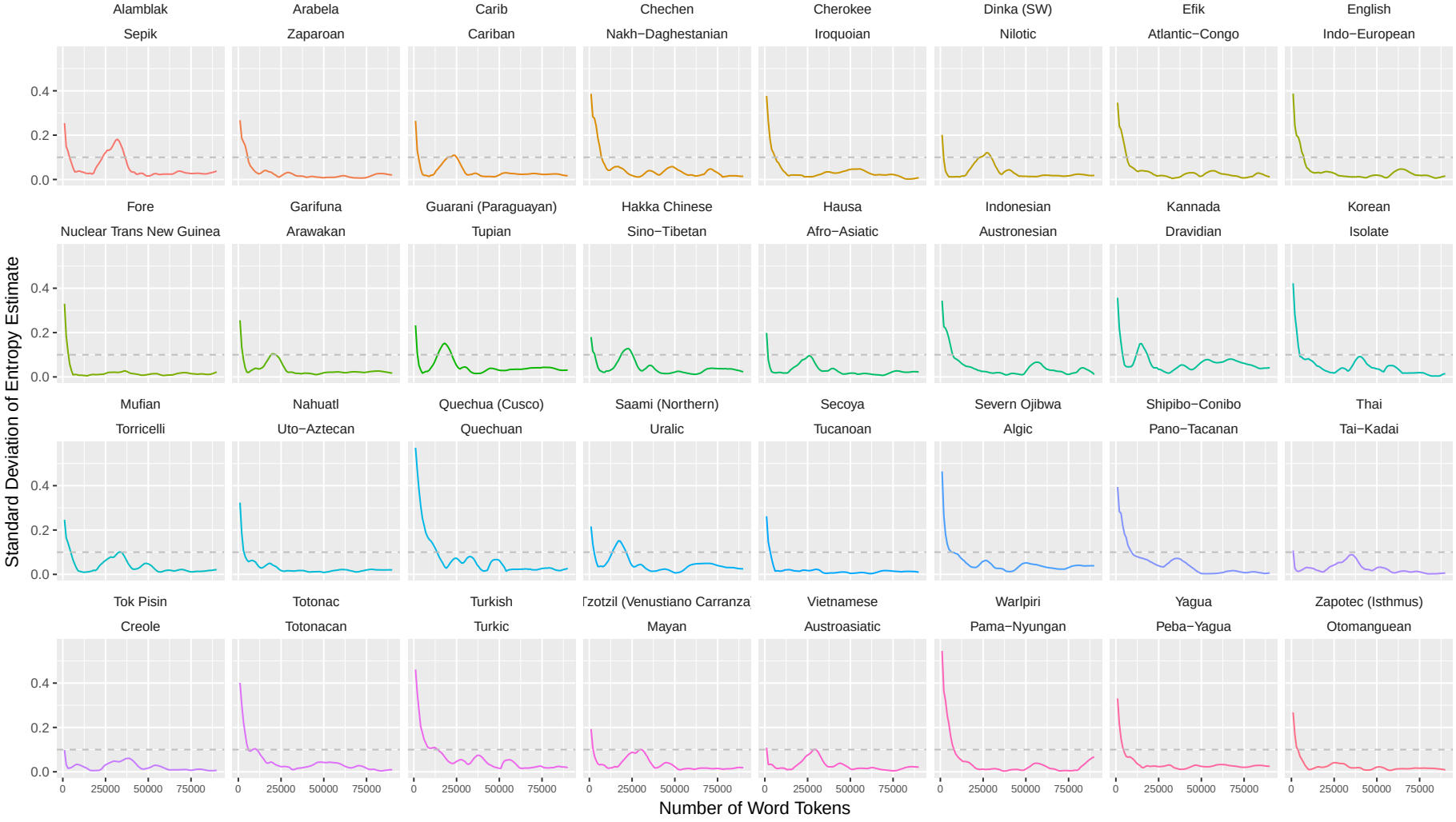

Figure 10. SDs of entropy rates as a function of text length across 32 languages of the PBC. Languages were chosen to represent some of the major language families across the world. The format is as in Fig. 2. The language names and families are given above the plots.

Appendix D 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 2, 3, and 4. For visual illustration a plot of pairwise correlations of the ML estimator with all other estimators is given in Figure 11.

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 [59]. Also, 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.

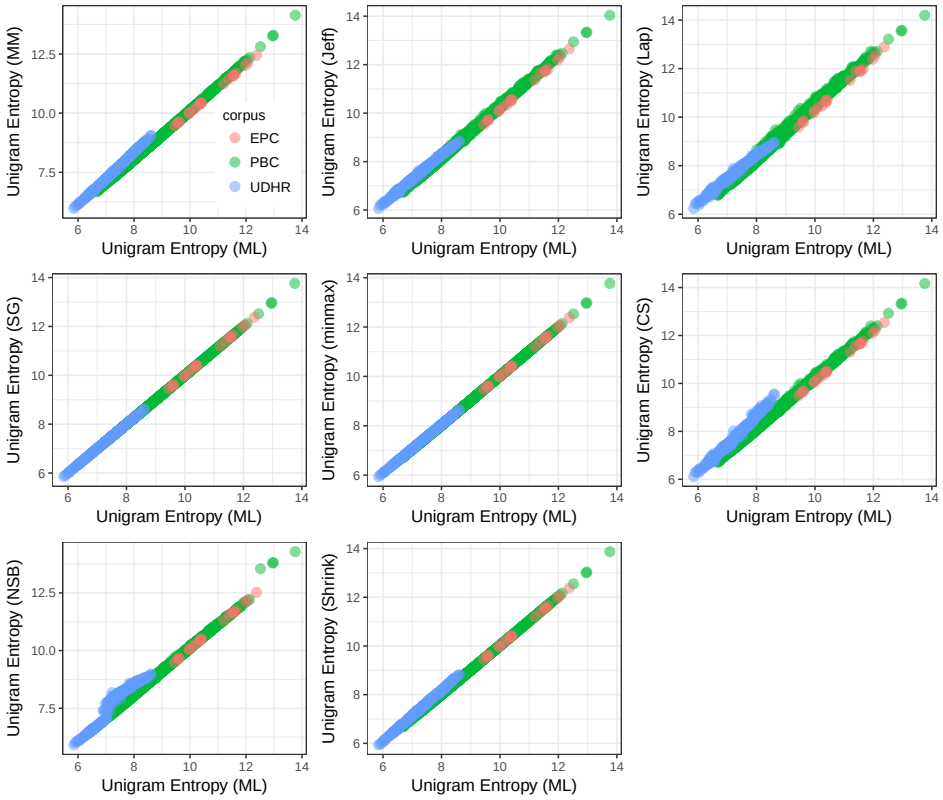

Figure 11. Pairwise correlations of estimated unigram entropy values for three different corpora: *Europarl Corpus* (EPC), *Parallel Bible Corpus* (PBC), and *Universal Declaration of Human Rights* (UDHR). Results of the maximum likelihood (ML) method are here taken as a baseline and correlated with 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 2. 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 3. 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 4. 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	-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2017 doi:10.20944/preprints201704.0180.v1

Peer-reviewed version available at Entropy 2017, 19, , 275; doi:10.3390/e19060275

Appendix E 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 9 methods – and entropy rates in Figure 12. The left panel plot in Figure 7 corresponds to the third plot from the right in the last row of Figure 12. 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.

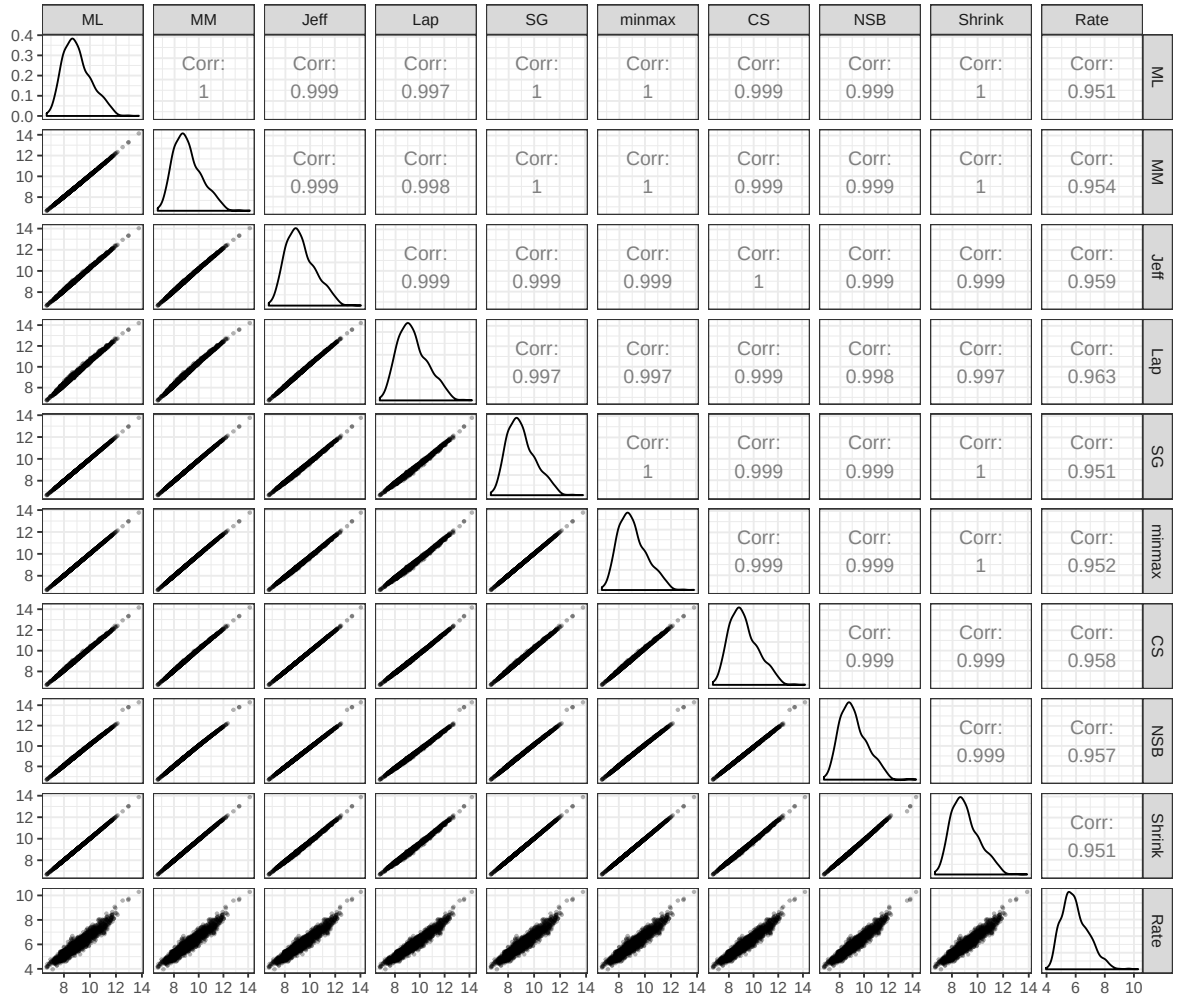

Figure 12. Correlations between the 9 unigram entropy estimators and entropy rates for the PBC. The panels in the lower half of the plot give scatterplots, the panels in the upper half give corresponding Pearson correlations. The diagonal panels give density plots.

Appendix F Correlations between PBC, EPC, and UDHR unigram entropies

Figure 13 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.

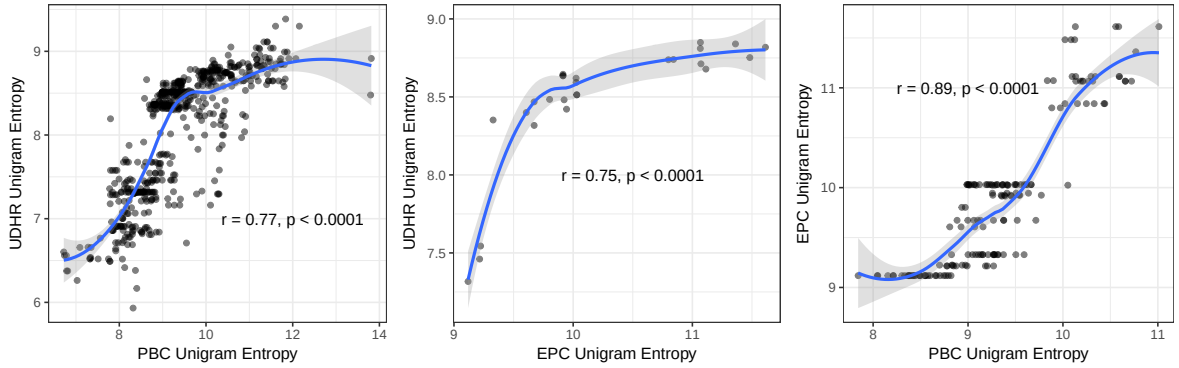

Figure 13. Correlations between unigram entropies (NSB estimated) for texts of the PBC and UDHR (left panel), texts of the EPC and UDHR (middle panel), and texts of the PBC and EPC (right panel). Local regression smoothers are given (blue lines) with 95% confidence intervals.

References

1. Shannon, C.E. A mathematical theory of communication. *The Bell Systems Technical Journal* **1948**, *27*, 379–423.
2. Shannon, C.E. Prediction and entropy of printed English. *The Bell System Technical Journal* **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. *Computational Linguistics* **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. *Information Theory, IEEE Transactions on* **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: An Interdisciplinary Journal of Nonlinear Science* **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. Data Compression Conference, 2003. Proceedings. DCC 2003. IEEE, 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, 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, 2016.
14. Bentz, C. The Low-Complexity-Belt: evidence for large-scale language contact in human prehistory? The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11); Roberts, S.G.; Cuskley, C.; McCrohon, L.; Barceló-Coblijn, L.; Feher, O.; Verhoef, T., Eds., 2016.

15. Juola, P. Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics* **1998**, *5*, 206–213.
16. Juola, P. Assessing linguistic complexity. In *Language complexity: typology, contact, change*; Miestamo, M.; Sinnemäki, K.; Karlsson, F., Eds.; Amsterdam: John Benjamins, 2008; pp. 89–108.
17. Gerlach, M.; Font-Clos, F.; Altmann, E.G. Similarity of symbol frequency distributions with heavy tails. *Physical Review X* **2016**, *6*, 021009.
18. Bentz, C.; Berdicevskis, A. Learning pressures reduce morphological complexity: linking corpus, computational and experimental evidence. Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), 26th International Conference on Computational Linguistics, 2016.
19. Bochkarev, V.; Solovyev, V.; Wichmann, S. Universals versus historical contingencies in lexical evolution. *Journal of The Royal Society 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**, pp. 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. *Computational Linguistics* **2010**, *36*, 795–805.
24. Piantadosi, S.T.; Tily, H.; Gibson, E. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences* **2010**, *108*.
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 preprint* **2015**, p. 1504.04884.
27. Bentz, C.; Ferrer-i-Cancho, R. Zipf's law of abbreviation as a language universal. Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics; Bentz, C.; Jäger, G.; Yanovich, I., Eds. University of Tübingen, 2016.
28. Futrell, R.; Mahowald, K.; Gibson, E. Quantifying word order freedom in dependency corpora. Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015), 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, 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. *Cognitive psychology* **2010**, *61*, 23–62.
33. Fenk, A.; Fenk, G. Konstanz im Kurzzeitgedächtnis - Konstanz im sprachlichen Informationsfluß. *Zietschrift für experimentelle and angewandte Pshychologie*, *XXVII*, 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, 2001; pp. 431–448.
35. Ferrer-i Cancho, R.; del Prado Martín, F.M. Information content versus word length in random typing. *Journal of Statistical Mechanics: Theory and Experiment* **2011**, *L12002*.
36. Ferrer-i Cancho, R.; Dębowski, Ł.; del Prado Martín, F.M. Constant conditional entropy and related hypotheses. *Journal of Statistical Mechanics: Theory and Experiment* **2013**, *L07001*.
37. Ferrer i Cancho, R.; Díaz-Guilera, A. The global minima of the communicative energy of natural communication systems. *Journal of Statistical Mechanics: Theory and Experiment* **2007**, *2007*, P06009.
38. Ferrer-i-Cancho, R.; Solé, R.V. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences USA* **2003**, *100*, 788–791.
39. Ferrer i Cancho, R. Zipf's law from a communicative phase transition. *European Physical Journal 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. *Computational linguistics* **1996**, *22*, 39–71.
42. Och, F.J.; Ney, H. Discriminative training and maximum entropy models for statistical machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002, pp. 295–302.
43. Herbelot, A.; Ganesalingam, M. Measuring semantic content in distributional vectors. ACL (2), 2013, pp. 440–445.
44. Padó, S.; Palmer, A.; Kisselew, M.; Šnajder, J. Measuring Semantic Content To Assess Asymmetry in Derivation. Workshop on Advances in Distributional Semantics, 2015.
45. Santus, E.; Lenci, A.; Lu, Q.; Im Walde, S.S. Chasing Hypernyms in Vector Spaces with Entropy. EACL, 2014, pp. 38–42.
46. Resnik, P. Using information content to evaluate semantic similarity in a taxonomy. IJCAI'95 Proceedings of the 14th international joint conference on Artificial intelligence, 1995, Vol. 1, pp. 448–453.
47. Boon, M. Information density, Heaps' Law, and perception of factiness in news. *ACL 2014* **2014**, p. 33.
48. Stetson, P.D.; Johnson, S.B.; Scotch, M.; Hripcsak, G. The sublanguage of cross-coverage. Proceedings of the AMIA Symposium. American Medical Informatics Association, 2002, p. 742.
49. McFarlane, D.J.; Elhadad, N.; Kukafka, R. Perplexity analysis of obesity news coverage. AMIA Annual Symposium Proceedings. American Medical Informatics Association, 2009, Vol. 2009, p. 426.
50. Zhang, Y.; Kordoni, V.; Villavicencio, A.; Idiart, M. Automated multiword expression prediction for grammar engineering. Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties. Association for Computational Linguistics, 2006, pp. 36–44.
51. Ramisch, C.; Schreiner, P.; Idiart, M.; Villavicencio, A. An evaluation of methods for the extraction of multiword expressions. Proceedings of the LREC Workshop-Towards a Shared Task for Multiword Expressions (MWE 2008), 2008, pp. 50–53.
52. Nemenman, I.; Shafee, F.; Bialek, W. Entropy and inference, revisited. *Advances in neural information processing systems* **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. *International Journal of Bifurcation and Chaos* **2012**, *22*, 1250223.
54. Koehn, P. Europarl: A parallel corpus for statistical machine translation. MT summit, 2005, Vol. 5, pp. 79–86.
55. Mayer, T.; Cysouw, M. Creating a massively parallel Bible corpus. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 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), 2014, pp. 3158–3163.
56. Haspelmath, M. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* **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, 2014; chapter 42.
58. Geertzen, J.; Blevins, J.P.; Milin, P. Informativeness of linguistic unit boundaries. *Italian journal of linguistics* **2016**, *28*, 25–47.
59. Hausser, J.; Strimmer, K. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *The Journal of Machine Learning Research* **2009**, *10*, 1469–1484.
60. Shannon, C.E.; Weaver, W. *The mathematical theory of communication*; The University of Illinois Press: Urbana, 1949.
61. Basharin, G.P. On a statistical estimate for the entropy of a sequence of independent random variables. *Theory of Probability & Its Applications* **1959**, *4*, 333–336.
62. Harris, B. The statistical estimation of entropy in the non-parametric case. In *Topics in Information Theory*; Csiszar, I., Ed.; North-Holland: Amsterdam, 1975; p. 323–355.
63. Lesne, A.; Blanc, J.L.; Pezard, L. Entropy estimation of very short symbolic sequences. *Physical Review E* **2009**, *79*, 046208.

64. Dębowski, Ł. Consistency of the plug-in estimator of the entropy rate for ergodic processes. Proceedings of the 2016 IEEE International Symposium on Information Theory. (ISIT), 2016, p. 1651–1655.
65. Chomsky, N. *Aspects of the theory of syntax*; MIT Press, 1965.
66. Baroni, M. The zipfR package for lexical statistics : A tutorial introduction.
67. Ferrer, R.; Solé, R.V. Two Regimes in the Frequency of Words and the Origins of Complex Lexicons : Zipf ' s Law Revisited. *Journal of Quantitative Linguistics* **2001**, *8*, 165–173.
68. Gerlach, M.; Altmann, E.G. Stochastic model for the vocabulary growth in natural languages. *Physical Review X* **2013**, *3*, 021006.
69. 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. *Scientific Reports* **2012**, *2*.
70. Manning, C.D.; Schütze, H., Collocations. In *Foundations of statistical natural language processing*; MIT Press: Cambridge, MA, 1999; chapter 5.
71. Montemurro, M.; Pury, P.A. Long-range fractal correlations in literary corpora. *Fractals* **2002**, *10*, 451–461.
72. Cover, T.M.; Thomas, J.A. *Elements of information theory*; John Wiley & Sons, 2006.
73. Ziv, J.; Lempel, A. A universal algorithm for sequential data compression. *IEEE Transactions on information theory* **1977**, *23*, 337–343.
74. Ziv, J.; Lempel, A. Compression of individual sequences via variable-rate coding. *Information Theory, IEEE Transactions on* **1978**, *24*, 530–536.
75. Hausser, J.; Strimmer, K. *entropy: Estimation of Entropy, Mutual Information and Related Quantities*, 2014. R package version 1.2.1.
76. Bentz, C.; Ruzsics, T.; Koplenig, A.; Samaržić, T. A comparison between morphological complexity measures: typological data vs. language corpora. Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), 26th International Conference on Computational Linguistics, 2016.
77. Hammarström, H.; Forkel, R.; Haspelmath, M.; Bank, S., Eds. *Glottolog 2.7*; Jena: Max Planck Institute for the Science of Human History, 2016.
78. Kirby, S.; Tamariz, M.; Cornish, H.; Smith, K. Compression and communication in the cultural evolution of linguistic structure. *Cognition* **2015**, *141*, 87–102.
79. Ferrer-i-Cancho, R. Optimization models of natural communication. *Journal of Quantitative Linguistics* **2017**.
80. Baayen, H.R. *Word frequency distributions*; Kluwer: Dordrecht, Boston & London, 2001.
81. Agresti, A.; Hitchcock, D.B. Bayesian inference for categorical data analysis. *Statistical Methods and Applications* **2005**, *14*, 297–330.
82. Chao, A.; Shen, T.J. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and ecological statistics* **2003**, *10*, 429–443.