Computational Constancy Measures of Texts—Yule's *K* and Rényi's Entropy

Kumiko Tanaka-Ishii[*,**]
Kyushu University and JST-PRESTO

Shunsuke Aihara[†]
Gunosy Inc.

This article presents a mathematical and empirical verification of computational constancy measures for natural language text. A constancy measure characterizes a given text by having an invariant value for any size larger than a certain amount. The study of such measures has a 70-year history dating back to Yule's K, with the original intended application of author identification. We examine various measures proposed since Yule and reconsider reports made so far, thus overviewing the study of constancy measures. We then explain how K is essentially equivalent to an approximation of the second-order Rényi entropy, thus indicating its signification within language science. We then empirically examine constancy measure candidates within this new, broader context. The approximated higher-order entropy exhibits stable convergence across different languages and kinds of text. We also show, however, that it cannot identify authors, contrary to Yule's intention. Lastly, we apply K to two unknown scripts, the Voynich manuscript and Rongorongo, and show how the results support previous hypotheses about these scripts.

1. Introduction

A constancy measure for a natural language text is defined, in this article, as a computational measure that converges to a value for a certain amount of text and remains invariant for any larger size. Because such a measure exhibits the same value for any size of text larger than a certain amount, its value could be considered as a text characteristic.

The concept of such a text constancy measure was introduced by Yule (1944) in the form of his measure *K*. Since Yule, there has been a continuous quest for such measures, and various formulae have been proposed. They can be broadly categorized into three types, namely, those measuring (1) repetitiveness, (2) power law character, and (3) complexity.

[*] Kyushu University, 744 Motooka Nishiku, Fukuoka City, Fukuoka, Japan. E-mail: kumiko@ait.kyushu-u.ac.jp.

[**] JST-PRESTO, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan.

[†] Gunosy Inc., 6-10-1 Roppongi, Minato-ku, Tokyo, Japan.

Submission received: 11 July 2013; revised version received: 17 February 2015; accepted for publication: 18 March 2015.

doi:10.1162/COLI_a_00228

Yule's original intention for *K*'s utility lay in author identification, assuming that it would differ for texts written by different authors. State-of-the-art multivariate machine learning techniques are powerful, however, for solving such language engineering tasks, in which Yule's *K* is used only as one variable among many, as reported in Stamatatos, Fakotakis, and Kokkinakis (2001) and Stein, Lipka, and Prettenhofer (2010).

We believe that constancy measures today, however, have greater importance in understanding the mathematical nature of language. Although mathematical models of language have been studied in the computational linguistics milieu, via Markov models (Manning and Schuetze 1999), Zipf's law and its modifications (Mandelbrot 1953; Zipf 1965; Bell, Cleary, and Witten 1990), and Pitman-Yor models (Teh 2006) more recently, the true mathematical model of linguistic processes is ultimately unknown. Therefore, the convergence of a constancy measure must be examined through empirical verification. Because some constancy measures have a mathematical theory of convergence for a known process, discrepancies in the behavior of real linguistic data from such a theory would shed light on the nature of linguistic processes and give hints towards improving the mathematical models. Furthermore, as one application, a convergent measure would allow for comparison of different texts through a common, stable norm, provided that the measure converges for a sufficiently small amount of text. One of our goals is to discover a non-trivial measure with a certain convergence speed that distinguishes the different natures of texts.

The objective of this article is thus to provide a potential explanation of what the study of constancy measures over 70 years has been about, by answering the three following questions mathematically and empirically:

Question 1 Does a measure exhibit constancy?

Question 2 If so, how fast is the convergence speed?

Question 3 How discriminatory is the measure?

We seek answers by first showing the meaning of Yule's *K* in relation to the Rényi higher-order entropy, and by then empirically examining constancy across large-scale texts of different kinds. We finally provide an application by considering the natures of two unknown scripts, the Voynich manuscript and Rongorongo, in order to show the possible utility of a constancy measure.

The most important and closest previous work was reported in Tweedie and Baayen (1998), the first paper to have examined the empirical behavior of constancy measures on real texts. The authors used English literary texts to test constancy measure candidates proposed prior to their work. Today, the coverage and abundance of language corpora allow us to conduct a larger-scale investigation across multiple languages. Recently, Golcher (2007) tested his measure *V* (discussed later in this paper) with Indo-European languages and also programming language sources. Our papers (Kimura and Tanaka-Ishii 2011, 2014) also precede this one, presenting results preliminary to this article but with only part of our data, and neither of those provides mathematical analysis with respect to the Rényi entropy. Compared with these previous reports, our contribution here can be summarized as follows:

- Our work elucidates the mathematical relation of Yule's *K* to Rényi's higher-order entropy and explains why *K* converges.
- Our work vastly extends the corpora used for empirical examination in terms of both size and language.

- Our work compares the convergent values for these corpora.
- Our work also presents results for unknown language data, specifically from the Voynich manuscript and Rongorongo.

We start by summarizing the potential constancy measures proposed so far.

2. Constancy Measures

The measures proposed so far can broadly be categorized into three types, calculating the repetitiveness, power-law distribution, or complexity of text. This section mathematically analyzes these measures and summarizes them.

2.1 Measures Based on Repetitiveness

The study of text constancy started with proposals for simple text measures of vocabulary repetitiveness. The representative example is Yule's K (Yule 1944), while Golcher recently proposed V as another candidate (Golcher 2007).

2.1.1 Yule's K. To the best of our knowledge, the oldest mention of constancy values was made by Yule with his notion of K (Yule 1944). Let N be the total number of words in a text, $V(N)$ be the number of distinct words, $V(m,N)$ be the number of words appearing m times in the text, and m_{max} be the largest frequency of a word. Yule's K is then defined as follows, through the first and second moments of the vocabulary population distribution of $V(m,N)$, where $S_1 = N = \sum_m mV(m,N)$, and $S_2 = \sum_m m^2 V(m,N)$ (Yule 1944; Herdan 1964):

$$\begin{aligned} K &= C\frac{S_2 - S_1}{S_1^2} \\ &= C\Big[-\frac{1}{N} + \sum_{m=1}^{m_{max}} V(m,N)(\frac{m}{N})^2 \Big] \end{aligned} \tag{1}$$

where C is a constant enlarging of the value of K, defined by Yule as $C = 10^4$. K is designed to measure the vocabulary richness of a text: The larger Yule's K, the less rich the vocabulary is. The formula can be intuitively understood from the main term of the sum in the formula. Because the square of $(\frac{m}{N})^2$ indicates the degree of recurrence of a word, the sum of such degrees for all words is small if the vocabulary is rich, or large in the opposite case. Another simple example can be given in terms of S_2 in this formula. Suppose a text is 10 words long: if each of the 10 tokens is distinct (high diversity), then $S_2 = 1 \times 1 \times 10 = 10$; whereas, if each of the 10 tokens is identical (low diversity), then $S_2 = 10 \times 10 \times 1 = 100$.

Measures that are slightly different but essentially equivalent to Yule's K have appeared here and there. For example, Herdan defined V_m as follows (Herdan 1964, pp. 67, 79):

$$V_m = \sqrt{\sum_{m=1}^{m_{max}} V(m,N)(\frac{m}{N})^2 - \frac{1}{V(N)}}$$

Likewise, Simpson (1949) derived the following formula as a measure to capture the diversity of a population:

$$D = \sum_{m=1}^{m_{max}} V(m,N) \frac{m}{N} \frac{m-1}{N-1}$$

which is equivalent to Yule's K, as Simpson noted.

2.1.2 Other Measures Based on Simple Text Statistics. Apart from Yule's K, various measures have been proposed from simple statistical observation of text, as detailed in Tweedie and Baayen (1998). One genre is based on the so-called **token-type relation** (i.e., the ratio of the vocabulary size $V(N)$ and the text size N, in log) as formulated by Guiraud (1954) and Herdan (1964) as a law. Because this simple ratio is not stable, the measure was modified numerous times to formulate Herdan's C (Herdan 1964), Dugast's k and U (Dugast 1979), Maas' a^2 (Maas 1972), Tuldava's LN (Tuldava 1977), and Brunet's W (Brunet 1978).

Another genre of measures concerns the proportion of hapax legomena, that is $V(1,N)$. Honoré noted that $V(1,N)$ increases linearly with respect to the log of a text's vocabulary size $V(N)$ (Honoré 1979). Another ratio, of $V(2,N)$ to $V(N)$, was proposed as a text characteristic by Sichel (1975) and Maas (1972).

Each of these values, however, was found *not* to be convergent according to the extensive study conducted by Tweedie and Baayen (1998). In common with Yule's intention to apply such measures for author identification, they examined all of the measures discussed here, in addition to two measures explained later: Orlov's Z, and the Shannon entropy upper bound obtained from the relative frequencies of unigrams. They examined these measures with English novels (such as *Alice's Adventures in Wonderland*) and empirically found that only Yule's K and Orlov's Z were convergent. Given their report, we consider K the only true candidate among the constancy measures examined so far.

2.1.3 Golcher's V. Golcher's V is a string-based measure calculated on the suffix tree of a text (Golcher 2007). Letting the length of the string be N and the number of inner nodes of the (Patricia) suffix tree (Gusfield 1997) be k, V is defined as:

$$V = \frac{k}{N} \tag{2}$$

Golcher empirically showed how this measure converges to almost the same value across Indo-European languages for about 30 megabytes of data. He also showed how the convergent values differ from those calculated for programming language texts.

Golcher explains in his paper that the possibility of constancy of V does *not* yet have mathematical grounding and has only been shown empirically. He does not report values for texts larger than about 30 megabytes nor for those of *non*-Indo-European languages. A simple conjecture on this measure is that because a suffix tree for a string of length N has at most $N-1$ inner nodes, V must end up at some value $0 \leq V < 1$, for any given text.

Our group tested V with larger-scale data and concluded that V could be a constancy measure, although we admitted to observing a gradual increase (Kimura and Tanaka-Ishii 2014). Because V requires further verification on larger-scale data before ruling it out, we include it as a constancy measure candidate.

2.2 Measures Based on Power Law Distributions

Since Zipf (1965), power laws have been reported as an underlying statistical characteristic of text. The famous Zipf's law is defined as:

$$f(n) \propto n^{-\gamma} \tag{3}$$

where $\gamma \approx 1$, and $f(n)$ is the frequency of the nth most frequent word in a text. Various studies have sought to explain mathematically how the exponent could differ depending on the kind of text. To the best of our knowledge, however, there has been a limited number of reports related to text constancy.

An exception is the study on Orlov's Z (Orlov and Chitashvili 1983). Orlov and Chitashvili attempted to obtain explicit mathematical forms for $V(N)$ and $V(m, N)$ by more finely considering the long tails of vocabulary distributions for which Zipf's law does not hold. They obtained these forms through a parameter Z, defined as the potential text length minimizing the square error of the estimated $V(m, N)$, with its actual value as follows:

$$Z = arg \min_{N} \frac{1}{m_{max}} \sum_{m=1}^{m_{max}} \{ \frac{E[V(m,N)] - V(m,N)}{V(N)} \}^2 \tag{4}$$

Thus defining Z, they mathematically deduced for $V(N)$ the following formula:

$$V(N) = \frac{Z}{\log(m_{max}Z)} \frac{N}{N-Z} \log\left(\frac{N}{Z}\right) \tag{5}$$

Two ways to obtain Z can be formulated through approximation: one through Good-Turing smoothing (Good 1953), which assumes Zipf's law to hold, and the other using Newton's method. Tweedie and Baayen showed how the value of Z is stable at the size of an English novel by a single author and thus suggested that it could form a text characteristic. The empirical results, however, were not significantly convergent with respect to text size, and, moreover, Tweedie and Baayen provided their results without giving an estimation method (Tweedie and Baayen 1998). Calculation using Good-Turing smoothing, which is derived directly from Zipf's law, would cause Z to converge, but this does not take Orlov's original intention into consideration. Alternatively, our group (Kimura and Tanaka-Ishii 2014) verified Z through Newton's method by setting $g(Z) = 0$, where $g(Z)$ is the following function:

$$g(Z) = \frac{Z}{\log(m_{max}Z)} \frac{N}{N-Z} \log\left(\frac{N}{Z}\right) - V(N) \tag{6}$$

We also showed how the value of Z increases rapidly when the text size is larger than 10 megabytes.

The major problem with measures based on power laws lies in the skewed head and tail of the vocabulary population distribution. Because these exceptions constitute important parts of the population, parameter estimation by fitting to Equation (3) is sensitive to the estimation method. For example, the estimated value of the exponent for Zipf's law depends on the method used for dealing with these exceptions. We tested several simple methods of estimating the Zipf law's exponent γ with different ways of handling the head and tail of a distribution. There were settings that led to convergence, but the convergence depended on the settings. Such difficulty could be one reason why there has been no direct proposal for γ as a text constancy measure. Hence, due care must be taken in relating text constancy to a power law. We chose another path by considering text constancy through a random Zipf distribution, as described later in the experimental section.

2.3 Measures Based on Complexity

With respect to measures based on complexity, multiple reports have already examined the Shannon entropy (Shannon 1948; Cover and Thomas 2006). In addition, we introduce the Rényi higher-order entropy (Rényi 1960) as another possible measure.

2.3.1 Shannon Entropy Upper Bound. Let X be the random variable of a sequence $X = X_1, X_2, \ldots, X_i, \ldots$, where X_i represents the ith element of X: $X_i = x \in \mathbb{X}$, and where $\mathbb{X}$ represents a given set (e.g., a set of words or characters) whose members constitute the sequence. Let X_i^j $(i < j)$ denote the random variable indicating its subsequence $X_i, X_{i+1}, X_{i+2}, \ldots, X_j$. Let $P(X)$ indicate the probability function of a sequence X. The Shannon entropy is then defined as:

$$H(X) = -\sum_{X} P(X) \log P(X) \tag{7}$$

Tweedie and Baayen directly calculated an approximation of this formula in terms of the relative frequencies (for P) of unigrams (for X), and they concluded that the measure would continue increasing with respect to text size and would not converge for short, literary texts (Tweedie and Baayen 1998). Because we are interested in the measure's behavior on a larger scale, we replicated their experiment, as discussed later in the section on empirical constancy. We denote this measure as H_1 in this article.

Apart from that report, many have studied the entropy rate, defined as:

$$h^* = \lim_{n \to \infty} \frac{H(X_1^n)}{n} \tag{8}$$

Theoretically, the behavior of the entropy rate with respect to text size has been controversial. On the one hand, there have been indications of **entropy rate constancy** (Genzel and Charniak 2002; Levy and Jaeger 2007). These reports argue that the entropy rate of natural language could be constant. Due to the inherent difficulty in obtaining the true value of h^* from a text, however, these arguments are based only on indirect clues with respect to convergence. On the other hand, Hilberg conjectured a decrease in the human conditional entropy, as follows (Hilberg 1990):

$$H(X_n | X_1^{n-1}) \propto n^{-1+\beta}$$

He obtained this through an examination of Shannon's original experimental data and suggested that $\beta \approx 0.5$. From this formula, Dębowski induces that $H(X_1^n) \propto n^{\beta}$ and that the entropy rate can be formulated generally as follows (Dębowski 2014):

$$\frac{H(X_1^n)}{n} \approx An^{-1+\beta} + h^* \tag{9}$$

Note that at the limit of $n \to \infty$, this rate goes to h^*, a constant, provided that $\beta < 1.0$. Hilberg's conjecture is deemed compatible with entropy rate constancy at its asymptotic limit, provided that $h^* > 0$ holds.[1] We are therefore interested in whether this h^* forms a text characteristic, and if so, whether $h^* > 0$.

Empirically, many have attempted to calculate the upper bound of the entropy rate. Brown's report (Brown et al. 1992) is representative in showing a good estimation of the entropy rate for English from texts, as compared with values obtained from humans (Cover and King 1978). Subsequently, there have been important studies on calculating the entropy rate, as reported thoroughly in Schümann and Grassberger (1996). The questions related to h^*, however, remain unsolved. Recently, Dębowski used a Lempel-Ziv compressor and examined Hilberg's conjecture for texts by single authors (Dębowski 2013). He showed an exponential decrease in the entropy rate with respect to text size, supporting the validity of Equation (9). Following these previous works, we examine the entropy rate by using an algorithm proposed by Grassberger (1989) and later on by Farach et al. (1995). This method is based on universal coding. The algorithm has a theoretical background of convergence to the true h^*, provided the sequence is stationary, but has been proved by Shields (1992) to be inconsistent—that is, it does not converge to the entropy rate for certain non-Markovian processes. We still chose to apply this method, because it requires no arbitrary parameters for calculation and is applicable to large-scale data within a reasonable time.

The Grassberger algorithm (Grassberger 1989; Farach et al. 1995) can be summarized as follows. Consider a sequence X of length N. The maximum matching length L_i is defined as:

$$L_i = \max\{k : X_j^{j+k} = X_i^{i+k}\}$$

for $j \in \{1, \ldots, i-1\}, 1 \le j \le j+k \le i-1$. In other words, L_i is the maximum common subsequence before and after i. If $\bar{L}$ is the average length of L_i, given by

$$\bar{L} = \frac{1}{N} \sum_{i=1}^{i=N} L_i$$

then the method obtains the entropy rate h_1 as

$$h_1 = \frac{\log_2 N}{\bar{L}} \tag{10}$$

Given the true entropy rate h^*, convergence has been mathematically proven for a stationary process, such that $|h^* - h_1| = O(1)$ when $N \to \infty$. In this article, we consider this entropy rate h_1 as a constancy measure candidate.

[1] According to Dębowski (2009), $h^* = 0$ suggests that the next element of a linguistic process is deterministic, that is, a function of the corpus observed before, under the two conditions that (1) the number of possible choices for the element is finite, and (2) the corpus observed before is infinite. In reality, the finiteness of linguistic sequences has the opposite tendency (i.e., the size of the observed corpus is finite, and the possible vocabulary size is infinite).

2.3.2 Approximation of Rényi Entropy H_α. The Rényi entropy is a generalization of the Shannon entropy, defined as follows (Rényi 1960; Rényi 1970; Cover and Thomas 2006; Bromiley, Thacker, and Bouhova-Thacker 2010):

$$H_\alpha(X) = \frac{1}{1-\alpha}\log(\sum_X P^\alpha(X)) \tag{11}$$

where $\alpha \geq 0, \alpha \neq 1$. $H_\alpha(X)$ represents different ideas of sequence complexity for different α. For example:

- When $\alpha = 0$, $H_0(X)$ indicates the number of distinct occurrences of X.
- When the limit $\alpha \rightarrow 1$ is taken, Equation (11) reduces to the Shannon entropy.

The formula for $\alpha = 0$ becomes equivalent to the so-called **topological entropy** (hence, it is another notion of **entropy**) for certain probability functions (Kitchens 1998) (Cover and Thomas 2006). Note that the number of distinct tokens (i.e., the cardinality of a set) has been used widely as a rough approximation of complexity in computational linguistics. Indeed, in Section 2.1.2, we saw how some candidate constancy measures are based on a token-type relation, such that the number of types is related to the complexity of a text. For texts, note also that the value grows with respect to the text size, unless X is considered, for example, in terms of unigrams of a phonographic alphabet.

For $\alpha \rightarrow 1$, there is controversy regarding convergence, as noted in the previous section. Such difficulty in convergence for these α values lies in the nature of linguistic processes, in which the vocabulary set evolves.

This view motivates us to consider $\alpha > 1$ for $H_\alpha(X)$, since the formula captures complexity by considering linguistic hapax legomena to a lesser degree, thus giving the possibility of convergence. In fact, an approximation of the probability by the relative frequencies of unigrams at $\alpha = 2$ immediately shows the essential equivalence to Yule's K, since K from Equation (1) can be rewritten as follows:

$$\sum_{m=1}^{m_{max}} V(m,N)(\frac{m}{N})^2 = \sum_{x \in \mathbb{X}}(\frac{freq(x)}{N})^2$$

where $freq(x)$ is the frequency of $x \in \mathbb{X}$. Therefore, Yule's K has significance within the context of complexity.

This relation of Yule's K to the Rényi entropy H_2 is reported for the first time here, to the best of our knowledge. This mathematical relation clarifies both why Yule's K should converge and what the convergent value means; specifically, the value represents the gross complexity underlying the language system. As noted earlier, the higher-order entropy considers hapax legomena to a lesser degree and calculates the gross entropy only from the representative vocabulary population. This simple argument shows that Yule's K captures not only the simple repetitiveness of vocabulary but also the more profound signification of its equivalence with the approximated second-order entropy. Because K has been previously reported as a stable text constancy measure, we consider it here once again, but this time within the broader context of H_α.

2.4 Summary of Constancy Measure Candidates

Based on the previous reports (Tweedie and Baayen 1998; Kimura and Tanaka-Ishii 2014) and the discussion so far, we consider the following four measures as candidates for text constancy measures.

- Repetitiveness-based measures:
  Yule's K (Equation (1)); and Golcher's V (Equation (2)).
- Complexity-based measures:
  The Shannon entropy upper bound (h_1 as the entropy rate (Equations (10) and (8)) and H_1 (Equation (7), with X in terms of unigrams and the probability function in terms of relative frequencies); and the approximated Rényi entropy, denoted as H_α ($\alpha > 1$) (Equation (11), again with X and the probability function in terms of unigrams and relative frequencies, respectively).

In addition, we empirically consider how these measures can be understood in the context of the power-law feature of language. As noted in the Introduction, for the convergent measures the speed of attaining convergence with respect to text size is examined as well. Among the candidates, K and H_1 have been previously applied in a word-based manner, whereas V is string based. The Shannon entropy rate h_1 has been considered in both ways. Because we should be able to consider a text in terms of both words and characters, we examine the constancy of each measure in both ways.

Furthermore, because we have seen the mathematical equivalence of Yule's K and H_2, in the following we only consider H_2. As for H_α, we consider $\alpha = 3, 4$ only in comparison with H_2. Because H_1 is based on relative frequencies and can be considered together with H_2, we first focus on the convergence of the three measures V, h_1, and H_2, and then we consider H_1 in comparison with H_2, H_3, and H_4.

3. Data

3.1 Real Texts

Table 1 lists the data used in our experimental examination. The table indicates the data identifier (by which we refer to the data in the rest of the article), language, source, number of distinct tokens, data length by total number of tokens, and size in bytes. The first block contains relatively large-scale natural language corpora consisting of texts written by multiple authors, and the second block contains smaller corpora consisting of texts by single authors. The third block contains programming language corpora, and the fourth block contains corpora of unknown scripts, which we examine at the end of this article in Section 4.3.

For the large-scale natural language data, we considered five languages: English, Japanese, Chinese, Arabic, and Thai. These languages were chosen to represent different language families and writing systems. The large-scale corpora in English, Japanese, and Chinese consist of newspapers in chronological order, and the Thai and Arabic corpora include other kinds of texts. The markers 'w', 'c', and 'cr' appearing at the end of every identifier in Table 1 (e.g., Enews-c, Enews-w, and Jnews-cr) indicate text processed through words, characters, and transliterated Roman characters, respectively.

Table 1
Our data.

Identifier	Language kind	Source	Number of distinct tokens	Data length by tokens	Data size in bytes
		Large scale corpora			
Enews-c	English	WSJ Corpus(1987)	87	112,868,099	108MB
Enews-w			137,466	22,679,512	
Jnews-c	Japanese	2000–2009 Mainichi Newspaper	5,758	475,101,506	1.3GB
Jnews-cr			94	1,087,919,430	
Jnews-w			468,818	289,032.862	
Cnews-c	Chinese	1995 People's Daily Newspaper	5,963	24,696,511	67MB
Cnews-cr			88	68,325,519	
Cnews-w			144,336	14,965,501	
Atext-c	Arabic	Watan-2004 corpus	59	42,174,262	73MB
Atext-w			298,370	7,450,442	
Ttext-c	Thai	NECTEC corpus	159	1,444,536	3.9MB
Ttext-w			16,291	280,602	
		Small scale corpora			
Ebook1-w	English	Ulysses	34,359	325,692	1.5MB
Ebook2-w	English	Les Miserables	25,994	677,163	3MB
Fbook-w	French	Les Miserables	31,956	691,407	3MB
Gbook-w	German	Kritik der reinen Vernunft	10,604	215,299	1.3MB
Jbook-w	Japanese	Dohyo	19,179	502,137	2MB
Cbook-w	Chinese	Hong Lou Meng	18,450	701,255	2.5MB
Abook-w	Arabic	Quaran	16,121	75,185	728KB
Sbook-w	Sanskrit	Ramayana	62,318	213,736	1.9MB
		Corpora of programming languages			
Python-w	Python	python library sources	1,517,424	48,704,374	214MB
Cplus-w	C++	C++ library sources	127,332	15,617,801	64MB
Lisp-w	Common Lisp	sbcl and Clozure CL	164,248	2,326,270	32MB
		Corpora of Unknown scripts			
VoynichA-c	Unknown	Voynich Manuscript	22	44,360	44KB
VoynichB-c	Unknown	Voynich Manuscript	25	117,105	115KB
VoynichA-w	Unknown	Voynich Manuscript	2,628	7,460	44KB
VoynichB-w	Unknown	Voynich Manuscript	4,609	18,495	115KB
RongoA-c	Unknown	Rongorongo script	3,546	10,376	60KB
RongoB-c	Unknown	Rongorongo script	656	14,003	60KB

As for the small-scale corpora in the second block, the texts were only considered in terms of words, since verification via characters produced findings consistent with those obtained with the large-scale corpora. The texts were chosen because each was written by a single author but is relatively large.

Here, we summarize our preprocessing procedures. For the annotated Thai NECTEC corpus, texts were tokenized according to the annotation. The preprocessing methods for the other corpora were as follows:

- English: NLTK[2] was used to tokenize text into words.
- Japanese: Mecab[3] was used for tokenization, and KAKASI[4] was used for romanization.
- Chinese: ICTCLAS2013[5] was used for tokenization, and the pinyin Python library was used for pinyin romanization.
- Other European Languages: PunktWordTokenizer[6] was used for tokenization.

All the other natural language corpora were tokenized simply using spaces.

Following Golcher (2007), who first suggested testing constancy on programming languages, we also collected program sources from different languages (third block in Table 1). The programs were also considered solely in terms of words, not characters. C++ and Python were chosen to represent different abstraction levels, and Lisp was chosen because of its different ordering for function arguments. Source code was collected from language libraries. The programming language texts were preprocessed as follows. Comments in natural language were eliminated (although strings remained in the programs, where each was a literal token). Identical files and copies of sources in large chunks were carefully eliminated, although this process did not completely eliminate redundancy since most programs reuse some previous code. Finally, the programs were tokenized according to the language specifications.[7]

The last block of the table lists two corpora of unknown scripts. We consider these scripts at the end of this article in Section 4.3, through Figure 5, to show one possible application of the text constancy measures. The first unknown script is that of the Voynich manuscript, a famous text that is undeciphered but hypothesized to have been written in natural language. This corpus is considered in terms of both characters and words, where words were defined via the white space separation in the original text. Given the common understanding that the manuscript seems to have two different parts (Reddy and Knight 2011), we separated it into two parts according to the Currier annotation (identified as A and B, respectively). The second corpus of unknown text consists of the Rongorongo script of Easter Island (Daniels and Bright 1996, Section 13; Orliac 2005; Barthel 2013). This script's status as natural language is debatable, but if so, it is considered to possess characteristics of both phonographs and ideograms (Pozdniakov and Pozdniakov 2007). Because there are several ways to consider what constitutes a character in this script (Barthel 2013), we calculate values for the two most extreme cases as follows. For corpus RongoA-c, we consider a character inclusive of all adjoining parts (i.e., including accents and ornamental parts). On the other hand, for

[2] `http://nltk.org.`

[3] `http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html.`

[4] `http://kakasi.namazu.org.`

[5] `http://ictclas.nlpir.org.`

[6] `http://nltk.org.`

[7] With respect to the Lisp programming language, its culture favors long, hyphenated variable names that can be almost as long as a sentence. For this work, therefore, Lisp variable names were tokenized by splitting at the hyphens.

corpus RongoB-c, we separate parts as reasonably as possible, among multiple possible separation methods. Because the unit of word in this script is unknown, the Rongorongo script is only considered in terms of characters.

3.2 Random Data

The empirical verification of convergence for real data is controversial. We must first note that it does not conform with the standard approach to statistical testing. In the domain of statistics, it is a common understanding that "convergence" cannot be tested. A statistical test raises two contrasting hypotheses—called the null and alternative hypotheses—and calculates a p-value indicating the probability of the null hypothesis to occur. When this p-value is smaller than a certain value, the null hypothesis is considered unable to occur and is thus rejected. For convergence, the null hypothesis corresponds to "not converging," and the alternative hypothesis, to "converging." The problem here is that the null hypothesis is always related to the alternative hypothesis to a certain extent, because the difference between convergence and non-convergence is merely a matter of degree. In other words, the notion of convergence for a constancy measure does not conform with the philosophy of statistical testing. Convergence is therefore considered in terms of the distance from convergent values, or in terms of the error with respect to some parameter (such as data size). Such a distance cannot be calculated for real data, however, since the underlying mathematical model is unknown.

To sum up, verification of the convergence of real data must be considered by some other means. Our proposal is to consider convergence in comparison to a set of random data whose process is known. For this random data, we considered two kinds.

The first kind is used to examine data convergence in Section 4.1. This random data was generated from real data by shuffling the original text with respect to certain linguistic units. Tweedie and Baayen (1998) presented results by shuffling words, where the original texts were literary texts by single authors. Here, we generated random data by shuffling (1) words/characters, (2) sentences, or (3) documents. Because these options greatly increased the number of combinations of results, we mainly present the results with option (1) for large-scale data in this article. There are three reasons for this: Convergence must be verified especially at large scale; the most important convergence findings for randomized small-scale data were already reported in Tweedie and Baayen (1998); and the results for options (2) and (3) were situated within the range of option (1) and the original texts.

Randomization of the words and characters of original texts will destroy various linguistic characteristics, such as n-grams and long-range correlation. The convergence properties of the three measures V, h_1, and H_2 are as follows. The convergence of V is unknown, because it lacks a mathematical background. Even if the value of V did converge, the convergent value for randomized data would differ from that of the original text, since the measure is based on repeated n-grams in the text. h_1 converges to the entropy rate of the randomized text, if the data size suffices. This is supported by the mathematical background of the algorithm, which converges to the true entropy rate for stationary data. Even when h_1 converges for random data, the convergent value will be larger than that of the original text, because h_1 considers the probabilities of n-grams. Lastly, H_2 converges to the same point for a randomized text and the original text, because it is the approximated higher-order entropy, such that words and characters are considered to occur independently.

The second kind of random data is used to compare the convergent values of different texts for a constancy measure, as considered in Section 4.2. Random corpora were generated according to four different distributions: one uniform, and the other three following Zipf distributions with exponents of γ = 0.8, 1.0, and 1.3, respectively, for Equation (3). Because each set of real data consists of different numbers of distinct tokens, ranging from tens to billions, random data sets consisting of 2^n distinct tokens for every $n = 4 \ldots 19$, were randomly generated for each of the four distributions. We only consider the measures H_2 and H_0 for these data sets. Both of these measures have convergent values, given a sufficient data size.

4. Experimental Results

From the previous discussion, we applied the three measures V, h_1, and H_2 with five large-scale and eight small-scale natural language corpora, three programming language corpora, and two unknown script corpora, in terms of words and characters. Because there were many results for different combinations of measure, data, and token (word or character), this section is structured so that it best highlights our findings.

4.1 Empirical Constancy

Figures 1, 2, and 3 in this section can be examined in the following manner. The horizontal axis indicates the text size of each corpus, in terms of the number of tokens, on a log scale. Chunks of different text sizes were always taken from the head of the corpus.[8] The vertical axis indicates the values of the different measures: V, h_1, or H_2. Each figure contains multiple lines, each corresponding to a corpus, as indicated in the legends.

First, we consider the results for the large-scale data. Figure 1 shows the different measures for words (left three graphs) and characters (right three graphs). We can see that V increased for both words and characters (top two graphs). Golcher tested his measure on up to 30 megabytes of text in terms of characters (Golcher 2007). We also observed a stable tendency up to around 10^7 characters. The increase in V became apparent, however, for larger text sizes. Thus, it is difficult to consider V as a constancy measure.

As for the results for h_1 (middle graphs), both graphs show a gradual decrease. The tendency was clearer for words than for characters. For some corpora, especially for characters, it was possible to observe some values converging towards h^*. The overall tendency, however, could not be concluded as converging. This result suggests the difficulty in attaining convergence of the entropy rate, even with gigabyte-scale data. From the theoretical background of the Grassberger algorithm, the values would possibly converge with larger-scale data. The continued decrease could be due to multiple reasons, including the possibility of requiring far larger data than that used here, or a discrepancy between linguistic processes and the mathematical model assumed for the Grassberger algorithm.

We tried to estimate h^* by fitting the Equation (9). For the corpora with good fitting, all of the estimated values were larger than zero, but many of the results could not

[8] For real data, this was done without any randomization of the order of texts for all corpora besides Atext-w and Atext-c. The Watan corpus is distributed *not* in the chronological order of the publishing dates, but as a set of articles grouped into categories (i.e., all articles of one category, then all articles of another category, and so on). Because of this, there is a large skew in the vocabulary distribution, depending on the section of the corpus. We thus randomly reshuffled the articles by categories for the whole corpus before taking chunks of different sizes (always from the beginning) to generate our results. Apart from this, we avoided any arbitrary randomization with respect to the original data summarized in Table 1.

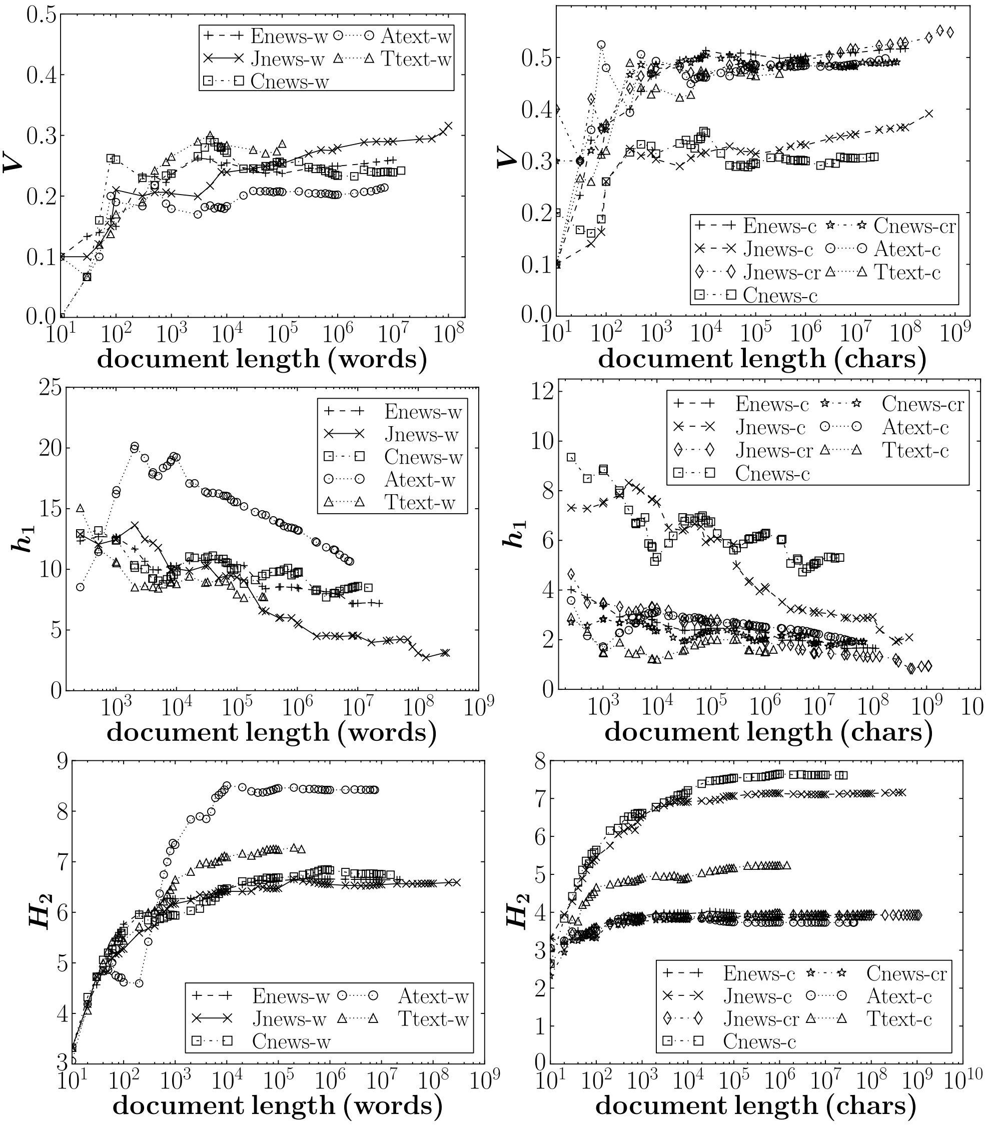

Figure 1
V, h_1, and H_2 values in terms of words and characters for the large-scale corpora.

be fitted easily, and the estimated values were unstable due to fluctuation of the lines. Whether a value for h^* is reached asymptotically and also whether $h^* > 0$ remain important questions requiring separate, more extensive mathematical and empirical studies.

In contrast, H_2 (or Yule's K, bottom graphs) showed convergence, already at the level of 10^5 tokens, for both words and characters. From the previous verification of Yule's K, we can conclude that H_2 is convergent. The final convergent values, however, differed for the various writing systems. We return to this issue in the next section.

To better understand the convergence, Figure 2 shows the results for the corresponding randomized data. As mentioned in Section 3.2, the original texts were

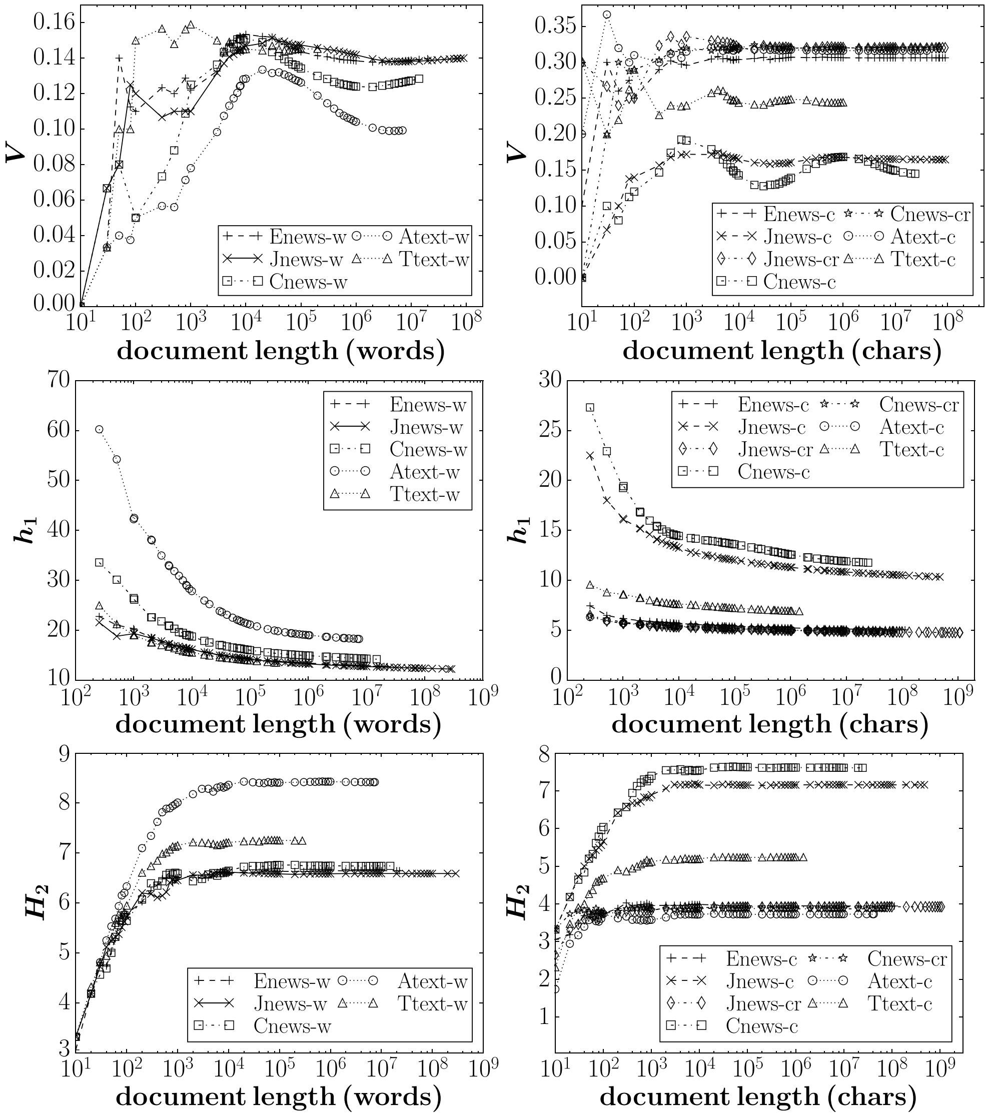

Figure 2
V, h_1, and H_2 values in terms of words and characters for the *randomized* large-scale corpora.

randomized by shuffling words and characters for the data examined by words and characters, respectively. Therefore, all n-gram characteristics existing in the text were destroyed, and what remained were the different words and characters appearing in a random order. Here, we see how the random data's behavior has some of the theoretical properties of convergence, as summarized in Section 3.2.

As mentioned previously, because V has no mathematical background, its behavior even for uniform random data is unknown, and even if it converged, the convergent value would be smaller than that of the original text. The top two graphs in Figure 2 exhibit some oscillation, especially for randomized Chinese (`Cnews-c,w`). Such peculiar oscillation was already reported by Golcher himself (Golcher 2007) for uniformly random data. This was easy to replicate, as reported in Kimura and Tanaka-Ishii (2014), for uniformly random data with the number of distinct tokens up to a hundred. Because the word distribution almost follows Zipf's law, the vocabulary is not uniformly distributed, yet oscillating results occur for some randomized data in the top left figure. Moreover, the values seem to increase for Japanese and English for words at a larger scale. Although the plots for some scripts seem convergent (top right graph), these convergent values are theoretically different from those of the original texts, if they exist, and this stability is not universal across the different data sets. Given this result, it is doubtful that V is convergent across languages.

In contrast, h_1 is mathematically proven to be convergent given infinite-length randomized data, but to larger values than those of the original texts, as mentioned in Section 3.2. The middle two graphs of Figure 2 show the results for h_1. The majority of the plots do not reach convergence even at the largest data sizes, but for certain results with characters, especially in the Roman alphabet, the plots seem to go to a convergent value (middle right). All the plots can be extrapolated to converge to a certain entropy rate above zero, although these values are larger than the convergent values—if they ever exist—of the real data. These results confirm the difficulty of judging whether the entropy rates of the original texts are convergent and whether they remain above zero.

Lastly, it is easy to see that H_2 is convergent for a randomized text (bottom two graphs), and the convergent values are the same for the cases with and without randomization. In fact, the plots converge to exactly the same points faster and more stably, which shows the effect of randomization.

As for the other randomization options, by sentences and documents, the findings—both the tendencies of the lines and the changes in the values—can be situated in the middle of what we have seen so far. The plots should increasingly fluctuate more like the real data because of the incomplete randomization, in the order of sentences and then documents.

Returning to inspection of the remaining real data, Figure 3 shows V, h_1, and H_2 in terms of words for the small-scale corpora (left column) and for the programming language texts (right column). For the small-scale corpora, in general, the plots are bunched together, and the results shared the tendencies noted previously for the large-scale corpora. V again showed an increase, while h_1 showed a tendency to decrease. H_2 converged rapidly and was already almost stable at 10^4 tokens. This again shows how H_2 exhibits stable constancy, especially with texts written by single authors.

As for the programming language results, the plots fluctuate more than for the natural language texts because of the redundancy within the program sources. Still, the global tendencies noted so far were just discernible. V had relatively larger values but h_1 and H_2 had smaller values for programs, as compared to the natural language texts. The differences in value indicate the larger degree of repetitiveness in programs.

Lastly, Figure 4 shows the H_α results for the *Wall Street Journal* in terms of words in unigrams (Enews-w). The horizontal axis indicates the corpus size, and the vertical axis indicates the approximated entropy value. The different lines represent the results for H_α with $\alpha = 1, 2, 3, 4$. The two H_1 plots represent calculations with and without Laplace smoothing (Manning and Schuetze 1999). We can see that without smoothing, H_1 increased, as Tweedie and Baayen (1998) reported, but in contrast to their conclusion, we observe a tendency of convergence for larger-scale data. The increase was due to the influence of low-frequency vocabulary pushing up the entropy. The opposite tendency to decrease was observed for the smoothed probabilities, with the plot eventually converging to the same point as that for the unsmoothed H_1 values. The convergence

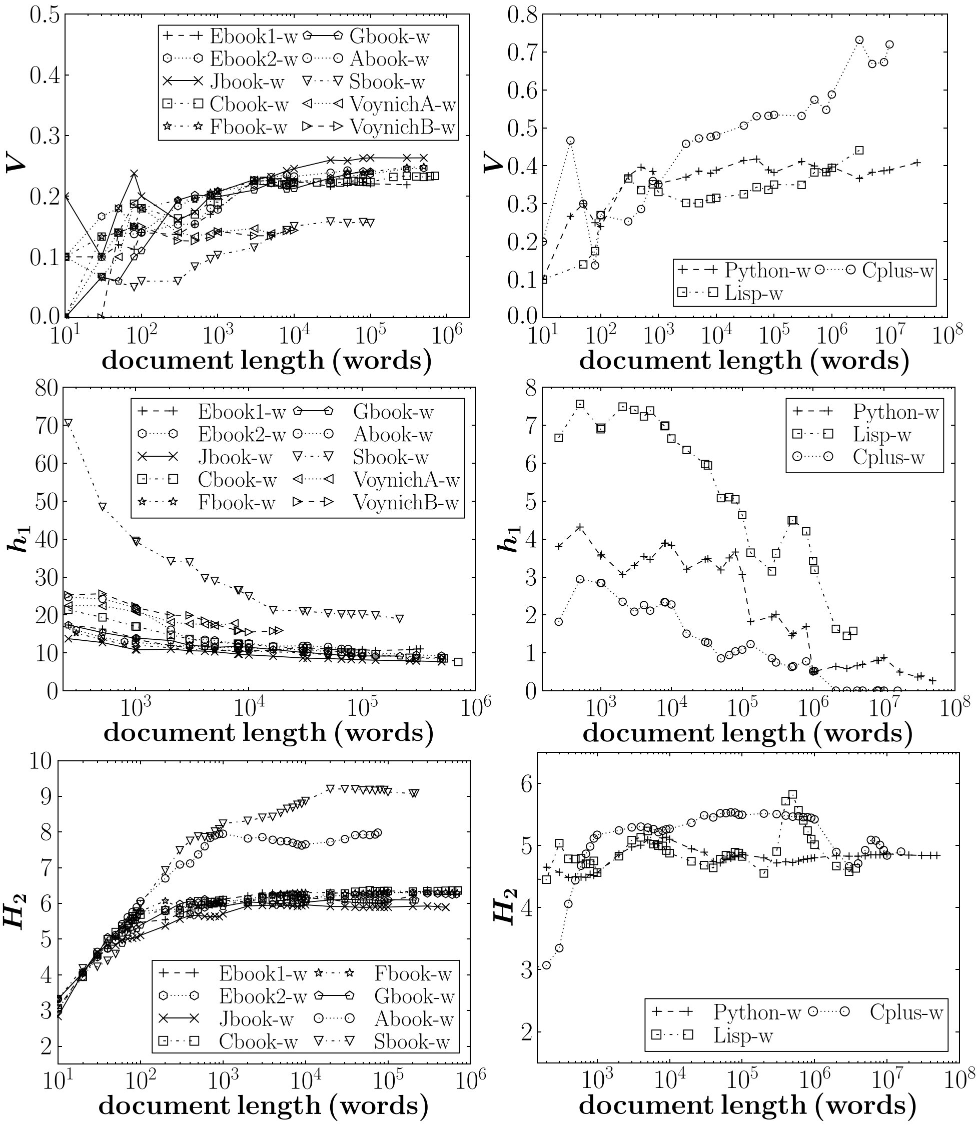

Figure 3
V, h_1, and H_2 values for the small-scale corpora and programming language texts in terms of words.

was by far slower for H_1 as compared with that for H_2, H_3, and H_4, which all had attained convergence already at 10^2 tokens. The convergence values naturally decreased for larger α, although the amount of decrease itself rapidly decreased with larger α.

In answer to Questions 1 and 2 raised in the Introduction—which measures show constancy, with sufficient convergence speed—the empirical conclusion from our data is that H_α with $\alpha > 1$ showed stable constancy when the values were approximated using relative frequencies. For H_1, the convergence was much slower because of the strong influence of low-frequency words. Consequently, the constancy of H_α with $\alpha > 1$ is attained by representing the gross complexity underlying a text.

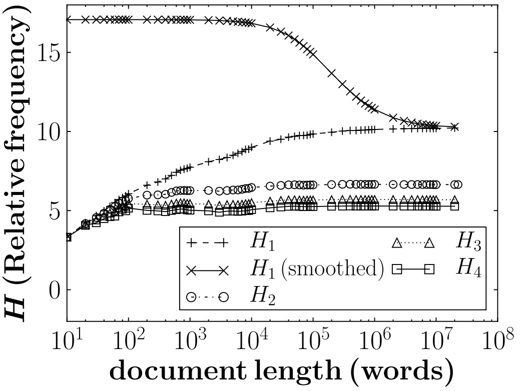

Figure 4
H_α with $\alpha = 1, \ldots, 4$ for *The Wall Street Journal* (Enews-w).

4.2 Discriminatory Power of H_2

Now we turn to Question 3 raised in the Introduction and examine the discriminatory power of H_2. As Yule intended, does H_2 identify authors? Given the influence of different writing systems, as seen previously in Figure 1, we examine the relation between H_2 and the number of **distinct tokens** (the alphabet/vocabulary size). Note that because this number corresponds to H_0 in Equation (11), this analysis effectively considers texts on the H_0-H_2 plane. Since H_0 grows according to the text size, unlike H_2, the same text size must be used for all corpora in order to meaningfully compare H_0 values.[9] Given that H_2 converges fast, we chose a size of 10^4 tokens to handle all of the small- and large-scale corpora.

For each of the corpora listed in Table 1 and the second kind of random corpora explained at the end of Section 3.2, Figure 5 plots the values of H_2 (vertical axis) and the number of distinct tokens H_0 (horizontal) measured for each corpus at a size of 10^4 tokens. The three large circles are groupings of points. The leftmost group represents news sources in alphabetic characters. All of the romanized Chinese, Japanese, and Arabic texts are located almost at the same vertical location as the English text. This indicates the difficulty for H_2 to distinguish natural languages if measured in terms of alphabetic characters. The middle group represents the programming language texts in terms of words. This group is located separately (vertically lower than the natural language corpora in terms of words), so H_2 is likely to distinguish between natural languages and programming languages. The rightmost group represents the small-scale corpora. Considering the proximity of these points despite the variety of the content, it is unlikely that H_2 can distinguish authors, in contrast to Yule's hope. Still, these points are located lower than those for news text. Therefore, H_2 has the potential to distinguish genre or maybe writing style.

[9] Because H_0 is not convergent, the horizontal locations remain unstable, unless the tokens are of a phonographic alphabet. In other words, for all word-based results and character-based results not based on a phonographic alphabet, the resulting horizontal locations are changed by increasing the corpus size. As for the random data, the H_0 values are convergent, because these data sets have a finite number of distinct tokens. Since H_0 is measured only for the first 10^4 tokens, however, the horizontal locations are underestimated, especially for random data following a Zipf distribution.

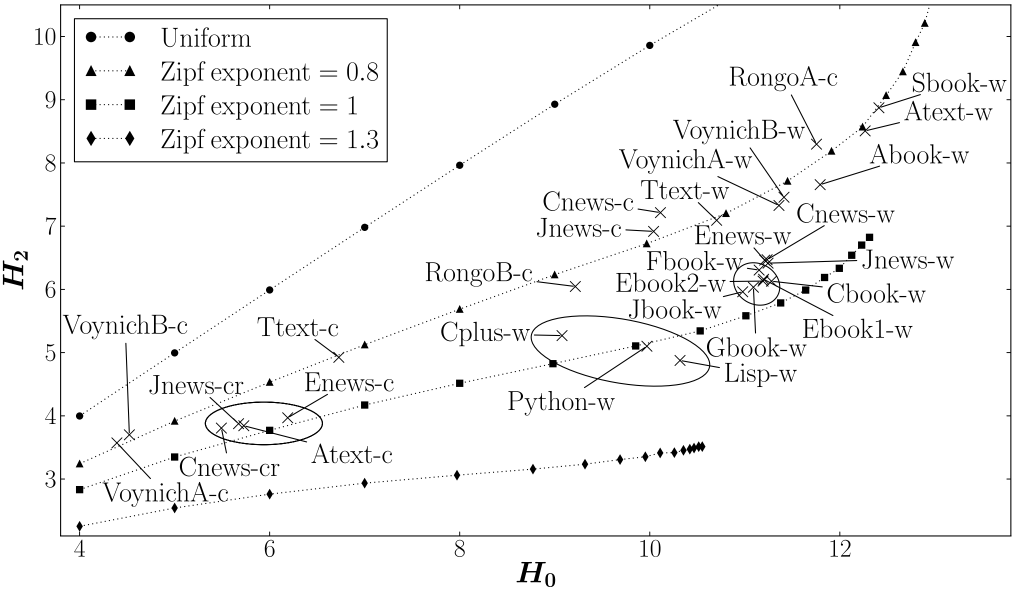

Figure 5
Convergent H_2 values with respect to the number of distinct tokens for each corpus.

The natural language texts located near the line for a Zipf exponent of 0.8 are those of the non-alphabetic writing systems.[10] Note that Chinese characters have morphological features, and the Arabic and Thai languages also have flexibility in terms of which units are considered words and morphemes. In other words, the plots closer to the random data with a smaller Zipf exponent are for language corpora of morphemic sequences. The group of plots measured for phonographic scripts is located near the line for a Zipf exponent of 1.0 (the grouping of points in the leftmost circle), which could suggest that morphemes are more randomized units than words.

4.3 Application to Unknown Scripts: Voynich Manuscript and Rongorongo Script

The nature of unknown scripts can also be considered through our understanding thus far. Figure 5 includes plots for the Voynich manuscript in terms of words and characters, and for the Rongorongo script in terms of characters. Like all the data seen in this figure, the points are placed at the H_2 values (vertically) for the number of distinct tokens (horizontally) at the specified size of 10^4 tokens, with the exception of Voynich-A in terms of words. Because this corpus consists of fewer than 10^4 words (refer to the data length by tokens listed for VoynichA-w in Table 1), its point is located horizontally at the vocabulary size corresponding to the corpus' maximum size.

For the two Voynich manuscript parts, the plots in terms of words appear near the Arabic corpus for words (Abook-w). For characters, on the other hand, the plots are at the leftmost end of the figure. This was due to overestimation of the total number

[10] Note that here we use the values of the Zipf exponent for the *random* data, and *not* the estimated exponents for the real data. The rank-frequency distributions of characters, especially for phonetic alphabets, often do not follow a power law.

of characters for the alphabetic texts (e.g., both English and other, romanized language texts), since all ASCII characters, such as colons, periods, and question marks, are counted. Still, the H_2 values are located almost at the same position as for the other romanized texts, indicating that the Voinich manuscript has approximately similar complexity. These results suggest the possibility that the Voynich manuscript could have been generated from a source in natural language, possibly written in some script of the abjad type. This supports previous findings (Reddy and Knight 2011; Montemurro and Zanette 2013), which reported the possibility of the Voynich manuscript being in a natural language and the coincidence of its word length distribution with that of Arabic.

On the other hand, the plots for the Rongorongo script appear near the line for a Zipf exponent of 0.8, with RongoA near Arabic in terms of words but RongoB somewhat further down from Japanese in terms of characters. The status of Rongorongo as natural language has been controversial (Pozdniakov and Pozdniakov 2007). Both points in the graph, however, are near many other natural language texts (and not widely separated), making it reasonable to hypothesize that Rongorongo is indeed natural language. The characters can be deemed morphologically rich, because both plots are close to the line for a Zipf exponent of 0.8. In the case of RongoA, for which a character was considered inclusive of all parts (i.e., including accents and ornamental parts), the morphological richness is comparable to that of the words of an abjad script. On the other hand, when considering the different character parts as distinct (RongoB), the location drifts towards the plot for Thai, a phonographic script, in terms of characters. Therefore, the Rongorongo script could be considered basically morphemic, with some parts functioning phonographically. This conclusion again supports a previous hypothesis proposed by a domain specialist (Pozdniakov and Pozdniakov 2007).

This analysis of two unknown scripts supports previous conjectures. Our results, however, only add a small bit of evidence to those conjectures; clearly, reaching a reasonable conclusion would require further study. Moreover, the analysis of unknown scripts introduced here could provide another possible application of text constancy measures, from a broader context.

5. Conclusion

We have discussed text constancy measures, whose values are invariant across different sizes of text, for a given text. Such measures have a 70-year history, since Yule originally proposed K as a text characteristic, potentially with language engineering utility for problems such as author identification. We consider text constancy measures today to have scientific importance in understanding language universals from a computational view.

After overviewing measures proposed so far and previous studies on text constancy, we explained how K essentially has a mathematical equivalence to the Rényi higher-order entropy. We then empirically examined various measures across different languages and kinds of corpora. Our results showed that only the approximated higher-order Rényi entropy exhibits stable, rapid constancy. Examining the nature of the convergent values revealed that K does not possess the discriminatory power of author identification as Yule had hoped. We also applied our understanding to two unknown scripts, the Voynich manuscript and Rongorongo, and showed how our constancy results support previous hypotheses about each of these scripts.

Our future work will include application of K to other kinds of data besides natural language. There, too, we will consider the questions raised in the Introduction, of whether K converges and of how discriminatory it is. We are especially interested in considering the relation between the value of K and the meaningfulness of data.

Acknowledgments

This research was supported by JST's PRESTO program.

References

Barthel, T. 2013. The Rongorongo of Easter Island: Thomas Barthel's transliteration system. Available at `http://kohaumotu.org/rongorongo_org/corpus/codes.html`. Accessed June 2015.

Bell, T. C., J. G. Cleary, and I. H. Witten. 1990. *Text Compression*. Prentice Hall.

Bromiley, P. A., N. A. Thacker, and E. Bouhova-Thacker. 2010. Shannon entropy, Renyi entropy, and information. Available at `http://www.tina-vision.net/docs/memos/2004-004.pdf`. Accessed June 2015.

Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, J. C. Lai, and R. L. Mercer. 1992. An estimate of an upper bound for the entropy of English. *Computational Linguistics*, 18(1):31–40.

Brunet, E. 1978. *Vocabulaire de Jean Giraudoux: Structure et Evolution*. Slatkine.

Cover, T. and R. King. 1978. A convergent gambling estimate of the entropy of English. *IEEE Transactions on Information Theory*, 24(4):413–421.

Cover, T. M. and J. A. Thomas. 2006. *Elements of Information Theory*. Wiley-Interscience.

Daniels, P. T. and W. Bright, editors. 1996. *The World's Writing Systems*. Oxford University Press.

Dębowski, Ł. 2009. A general definition of conditional information and its application to ergodic decomposition. *Statistics and Probability Letters*, 79(9):1260–1268.

Dębowski, Ł. 2013. Empirical evidence for Hilberg's conjecture in single author texts. In *Methods and Applications of Quantitative Linguistics: Selected papers of the 8th International Conference on Quantitative Linguistics (Qualico)*, pages 143–151, Belgrade.

Dębowski, Ł. 2014. The relaxed Hilberg conjecture: A review and new experimental support. Available at `http://www.ipipan.waw.pl/ldebowsk/`. Accessed June 2015.

Dugast, D. 1979. *Vocabulaire et Stylistique. I Théâtre et Dialogue*. Slatkine-Champion. Travaux de Linguistique Quantitative.

Farach, M., M. Noordewier, S. Savari, L. Shepp, A. Wyner, and J. Ziv. 1995. On the entropy of DNA: Algorithms and measurements based on memory and rapid convergence. In *Proceedings of the Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 48–57, San Francisco, CA.

Genzel, D. and E. Charniak. 2002. Entropy rate constancy in text. In *Annual Meeting of the Association for the ACL*, pages 199–206, Philadelphia, PA.

Golcher, F. 2007. A stable statistical constant specific for human language texts. In *Recent Advances in Natural Language Processing*, Borovets.

Good, I. J. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3–4):237–264.

Grassberger, P. 1989. Estimating the information content of symbol sequences and efficient codes. *IEEE Transactions on Information Theory*, 35:669–675.

Guiraud, H. 1954. *Les Charactères Statistique du Vocabulaire*. Universitaires de France Press.

Gusfield, D. 1997. *Algorithms on Strings, and Sequences: Computer Science and Computational Biology*. Cambridge University Press.

Herdan, G. 1964. *Quantitative Linguistics*. Butterworths.

Hilberg, W. 1990. Der bekannte grenzwert der redundanzfreien information in texten eine fehlinterpretation der shannonschen experimente? *Frequenz*, 44(9–10):243–248.

Honoré, A. 1979. Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing Bulletin*, 7:172–177.

Kimura, D. and K. Tanaka-Ishii. 2011. A study on constants of natural language texts. *Journal of Natural Language Processing*, 18(2):119–137.

Kimura, D. and K. Tanaka-Ishii. 2014. A study on constants of natural language texts. *Journal of Natural Language Processing*, 21:877–895. Special issue of awarded papers. [The English translated version of the article appeared in 2011 in Japanese].

Kitchens, B. 1998. *Symbolic Dynamics: One-sided, Two-sided and Countable State Markov Shifts*. Springer.

Levy, R. and T. F. Jaeger. 2007. Speakers optimize information density through information density through syntactic reduction. In *Annual Conference on Neural Information Processing Systems*, pages 1–8, Vancouver.

Maas, H. D. 1972. Zusammenhang zwischen wortschatzumfang und länge eines textes [Relationship between vocabulary and text length]. *Zeitschrift für Literaturwissenschaft und Linguistik*, 8:73–70.

Mandelbrot, B. 1953. An informational theory of the statistical structure of language. *Communication Theory*, 486–500.

Manning, C. and H. Schuetze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

Montemurro, M. and D. Zanette. 2013. Keywords and co-occurrence patterns in the Voynich Manuscript: An information-theoretic analysis. *PLOS One*. doi: 10.1371/journal.pone.0066344.

Orliac, C. 2005. The Rongorongo tablets from Easter Island: Botanical identification and 14c dating. *Archaeology in Oceania*, 40(3):115–119.

Orlov, J. K. and R. Y. Chitashvili. 1983. Generalized z-distribution generating the well-known ‘rank-distributions’. *Bulletin of the Academy of Sciences of Georgia*, 110:269–272.

Pozdniakov, K. and I. Pozdniakov. 2007. Rapanui writing and the Rapanui language: Preliminary results of a statistical analysis. *Forum for Anthropology and Culture*, 3:3–36.

Reddy, S. and K. Knight. 2011. What we know about the Voynich Manuscript. In *ACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Portland, OR.

Rényi, A. 1960. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematics, Statistics and Probability*, pages 547–561, Berkeley, CA.

Rényi, A. 1970. *Foundations of Probability*. Dover Publications.

Schümann, T. and P. Grassberger. 1996. Entropy estimation of symbol sequences. *Chaos*, 6(3):414–427.

Shannon, C. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656.

Shields, P. C. 1992. Entropy and prefixes. *Annals of Probability*, 20(1):403–409.

Sichel, H. S. 1975. On a distribution law for word frequencies. *Journal of the American Statistical Association*, 70(351):542–547.

Simpson, E. H. 1949. Measurement of diversity. *Nature*, 163:688.

Stamatatos, E., N. Fakotakis, and G. Kokkinakis. 2001. Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4):471–495.

Stein, B., N. Lipka, and P. Prettenhofer. 2010. Intrinsic plagiarism analysis. *Language Resources and Evaluation*, 45(1):63–82.

Teh, Y. W. 2006. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference On Computational Linguistics and 44th Annual Meeting of the ACL*, pages 985–992, Sydney.

Tuldava, J. 1977. Quantitative relations between the size of the text and lexical richness. *SMIL Quarterly, Journal of Linguistic Calculus*, 4:28–35.

Tweedie, F. J. and Baayen, R. H. 1998. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, 32:323–352.

Yule, G. U. 1944. *The Statistical Study of Literary Vocabulary*. Cambridge University Press.

Zipf, G. K. 1965. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Hafner, New York.