Joachim von zur Gathen & Daniel Loebenberger (2017). Why one Cannot Estimate the Entropy of English by Sampling. *Journal of Quantitative Linguistics* 1–30.
URL https://dx.doi.org/10.1080/09296174.2017.1341724.

WHY ONE CANNOT ESTIMATE THE ENTROPY OF ENGLISH BY SAMPLING

Joachim von zur Gathen and Daniel Loebenberger

15 March 2017

Abstract. There have been attempts to approximate the entropy of English by frequency analysis of large corpora. Our original goal was to deduce more precise estimates by extensive calculations. This did not work well, thus confirming a widely held belief in linguistics. In order to put this belief on a firm basis, we used a simplified language model, closely related to others in the literature. This model exhibits an unexpected trichotomy: for very small n, say up to $n = 4$ in our case, n-gram counting is reasonably reliable; for medium n, up to 14, increasing statistical noise is added, and beyond that we see statistical noise only. The model is precise enough to yield explicit values for the thresholds given above in dependence of the corpus size. Even though a mathematically rigorous proof for English itself is out of reach, our model gives a strong indication that frequency counting in (large) corpora is a dead end for approximating the entropy of English, and different linguistic tools and insights are required.

As far as we know, this is the first rigorous quantifiable argument concerning the linguistic intuition that frequency counting of samples is insufficient for entropy determination. The reader only interested in these general conclusions may safely skip Section 2 and Section 3, which deal with the English language, but mainly serve as a case study for our general findings in the later sections.

1. Introduction

In the late 1940s, Claude Elwood Shannon defined the *entropy* of a probability distribution by an explicit formula involving the probabilities of the distribution and showed the importance of this concept in various areas. In his foundational paper on the subject, from 1948, he applied it to electronic communications and initiated the theory of error-correcting codes. The channel entropy determines ultimate limits of their efficiency, namely their transmission rate. In 1949, he used it in cryptography. Here, the entropies of cleartexts and keys determine ultimate limits on the ability to decipher encrypted messages without access to the secret key. In 1951, he studied the entropy of (printed) English and gave upper and lower bounds for it.

Our interest comes from the second subject, namely cryptography. A particular case is the cryptanalysis of Vigenère ciphers from Kasiski (1863). Here, the frequencies of individual plaintext letters, letter-digrams, etc., survive with a flattened distribution; see von zur Gathen (2015), Section C.1. Many classical and modern

encryption systems have been broken, but for short encrypted texts, the reliability of a decipherment depends critically on the entropy of the plaintext language, according to Shannon's *Unicity Bound*. In view of this, our original motivation was to find good bounds on the entropy of the distribution of letters and polygrams in English.

Since Shannon's 1951 work, linguists have tried to improve on his bounds; see for example Cover & King (1978) or Brown, Della Pietra, Mercer, Della Pietra & Lai (1992). All of them have in common that a fairly large corpus is used to reason about the true nature of English. We first follow this approach and describe in Sections 2 and 3 our calculations using the corpus of contemporary American English (COCA), see Davies (2008-2012), containing 450 million words of different categories of printed English, and determine some letter- and word-frequencies in it. Namely, for n between 1 and 30, we determine the frequency of letter-n-grams and the frequencies of word monograms. We can then apply Shannon's formula in two ways: either for a fixed n, we compute the entropy of, say, letter-n-grams, or we consider all letter-$(n-1)$-grams with the conditional frequency of the consecutive nth letter and compute the entropy of this distribution. The conditional entropy values calculated turn out to decrease with growing n.

However, although our corpus is presumably larger than those used earlier for this purpose, our numerical values from certain polygram lengths on were in conflict with known values and the intuition. Indeed, for the analysis, ideally one would have a corpus of all texts in the language under discussion, printed English in our case. Then for each $n \geq 1$ and each letter-n-gram, one would determine its frequency. This yields a distribution over finite sequences of letters, to which we can apply Shannon's formula. However, no such corpus is available. This issue is well-known in the linguistic community. During a fruitful discussion, Köhler (2016) expressed the following opinion on this:

(1.1) "Zudem ist die Schätzung der Entropie meines Wissens nur aufgrund eines unendlichen Strings möglich, wobei die Eigenschaften von Schätzungen mittels abgebrochenem String fraglich sind. Außerdem gibt es keine unendlich langen Texte. Andere Objekte wie Korpora sind künstlich zusammengestellt und entsprechen keiner linguistisch begründbaren Spracheinheit."[1]

Our experiments confirmed this intuition. However, we found that for very small n, our frequency analyses lead to consistent results, whichever way a reasonably representative corpus was selected. It thus seems that we are indeed able to determine specific frequencies by analyzing finite-sized corpora only, but we fail when considering larger n (and thus the entropy of English).

[1] Additionally, an estimate of the entropy is to my knowledge only possible when using an infinite string, whereas the properties of estimates employing truncated strings are questionable. Also, there are no infinitely long texts. Other objects, like corpora, are artificially assembled and do not correspond to linguistically justifiable language units.

Specifically, we performed the computations on various derivatives of the corpus, say, 26 letters plus space only, or only on extracts like texts labelled "fictional" in the corpus. Our results are presented in Section 3. One finding is that the choice of derivative or sub-corpus does not influence the count substantially, thus showing a certain robustness of the sampling method. It would be interesting to see how other corpora fare in this respect.

One can argue that the frequencies obtained from a corpus are not representative, since a corpus is always a collection of (linguistic) objects whose statistical properties may have little in common with the set of all such objects. This is, of course, a valid point of view. The purpose of this work is, however, to show that sampling from a language cannot be used to estimate the entropy of the language reasonably. It turns out that for the sampling method to provide reliable results, the required corpus size is completely out of reach.

Specifically, we noticed that from $n = 5$ on, there seems to be increasing statistical noise in our data on letter-polygrams, and for $n \geq 14$, the noise seems to dominate. This is, of course, a well-known behavior and consistent with (1.1).

The main contribution of this work, starting in Section 5 and not tied to a particular language, is a precise analysis of this phenomenon. Namely, we describe a stochastic model for the entropy that explains this observation. Even though the language model we use is well-known, we provide *explicit quantitative estimates* for the expected entropy in terms of the corpus and alphabet size, giving—at least for certain special cases—explicit bounds for those sizes which are necessary for good entropy approximations. As far as we know, such an explicit analysis has not been known before. Our results show that there is a trichotomy when analyzing n-grams this way in any representative corpus: reasonable approximations to the true value of the entropy for very small n, the truth with some statistical noise for medium sized n, and statistical noise only for large n. We conclude that the approach of bounding the entropy by analyzing polygrams in a fixed (large) corpus is a dead end and cannot be carried much further than the current work. In order to get a better hold of a numerical value for the entropy of English, more linguistic insights are needed.

Our observations on the difficulty of approximating language entropy are consistent with results from the theory of computational complexity on this question, namely, that determining the entropy of a distribution is hard for a certain complexity class, see Section 5.

We concentrate our analysis on letter frequencies due to our cryptographic interest, but also do some computations with word frequencies. Here, the limitations discussed above show up even earlier, since our corpus has fewer words than letters.

2. Description of the corpus

The corpus of contemporary American English (COCA), see Davies (2008-2012), consists of a large number of English texts from five different genres: Academic texts,

fictional texts, magazine and newspaper texts and excerpts of spoken English. For the analysis, we considered written English texts only, that is, did not analyze the part of the COCA that contained spoken English. One reason for this was that we did not succeed in removing from the transcripts of spoken English artificially introduced tags (such as names), which might skew our statistical analyses. The resulting corpus consists of $2 \cdot 10^9 \approx 2^{31}$ characters and contains more than 340 million English words and roughly 65 million punctuation marks.

The characters in the COCA are from the set of all 95 printable ASCII characters. These are classified as

- 26 lowercase Roman letters: `abcdefghijklmnopqrstuvwxyz`
- 26 uppercase Roman letters: `ABCDEFGHIJKLMNOPQRSTUVWXYZ`
- 10 Arabic numerals: `0123456789`
- 32 special symbols: `!"#$%&'()*+,-./:;<=>?@[\]^_‘{|}~`
- space: ␣

A first inspection shows that all but the special symbols `\^‘{|~` occur in the COCA. Special symbols are sometimes called *punctuation marks*.

Each of the above mentioned text genres is split into files containing corresponding texts from the years 1990 to 2012. Every file contains several articles which start with `##`, followed by a seven digit identifier. Each article is split into paragraphs that are separated using a special HTML-type tag. For copyright reasons, the corpus is split into blocks of roughly 200 words to be compliant with the US Fair Use Law, 17 US Code §107 through §118, on copyrighted material.

For our analysis, we purged the COCA from all article identifiers. We then replaced all sequences of Arabic numerals by the special symbol `#`. Furthermore, we substituted paragraph tags and block delimiters by one of the remaining unused symbols. These newly introduced special symbols are not directly used in our statistical analyses, but are used to only capture the properties of written English in a single block.

We chose to analyze the COCA over the full alphabet of all printable ASCII characters first, distinguishing uppercase from lowercase Roman letters and keeping space and punctuation marks. A second analysis was done using the lowercase Latin alphabet with space only, that is, changing every Roman letter to its lowercase analog while ignoring any punctuation but keeping space. Sequences of consecutive spaces were counted as a single space.

To distinguish the relevant cases, we use the following notions for the classification of certain types of ASCII characters:

- A *symbol* is any printable ASCII character.

- A *letter* is any lowercase Roman letter or space.
- A *string* is a sequence of symbols preceded and succeeded but not containing space.
- A *word* is a sequence of lowercase Roman letters preceded and succeeded by space.

A single occurrence of one of the above defined notions is also called a *monogram*. For a fixed $n \geq 1$, we call the *polygram* containing n consecutive monograms an *n-gram*.

The purged COCA thus resembles excerpts of written English containing a large number of string-monograms, separated by space, each containing an arbitrary number of symbols (excluding space). All punctuation marks and contractions like `n't`, `'re`, `'s` or the Saxon genitive `'s` are monograms.

By ignoring all punctuation marks, replacing each uppercase Roman letter by its lowercase analogue and substituting any sequence of consecutive spaces by a single space, we obtain the corresponding corpus for word-monograms, where also the words are separated by space. Both corpora can then be used for the statistical analysis.

3. Elementary statistical analyses

After purging, we counted the occurrences of symbol-, letter-, string- and word-monograms. For any of these choices M of the set of monograms, we then computed the frequency distribution of n-grams. Such a frequency distribution simply counts how often a certain n-gram, say $g \in M^n$, occurs in the corpus. Dividing this count by the number of n-grams considered gives the *probability* $p_n(g)$ that the n-gram g occurs.

Definition 3.1. *(i) The* Shannon entropy *of the distribution p_n is*

$$H(p_n) = -\sum_{g \in M^n} p_n(g) \log_2 p_n(g).$$

(ii) The conditional entropy *of the distribution p_n given p_{n-1} is*

$$H(p_n : p_{n-1}) = H(p_n) - H(p_{n-1}).$$

The latter definition corresponds to the chain rule for the entropy and can be found in any textbook on information theory such as Cover & Thomas (2006).

Most of the following computational experiments were carried out on a 2,2 GHz Intel Core i7 with 8 GB RAM. For larger computations, we employed a small cluster with 8 dual-core 3.00GHz Intel Xeon CPUs and 64GB RAM. For all of the following numerical results, we provide in the text graphs only, see Appendix A for listings of the underlying numerical data.

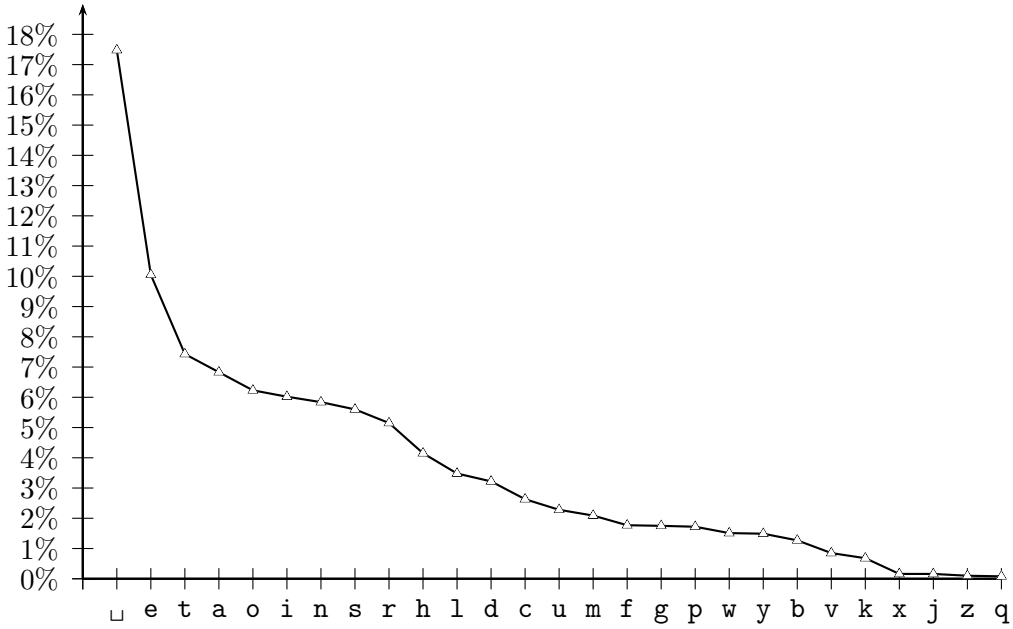

Figure 3.1: Frequencies of letter-monograms over the alphabet containing lowercase Latin letters and space only.

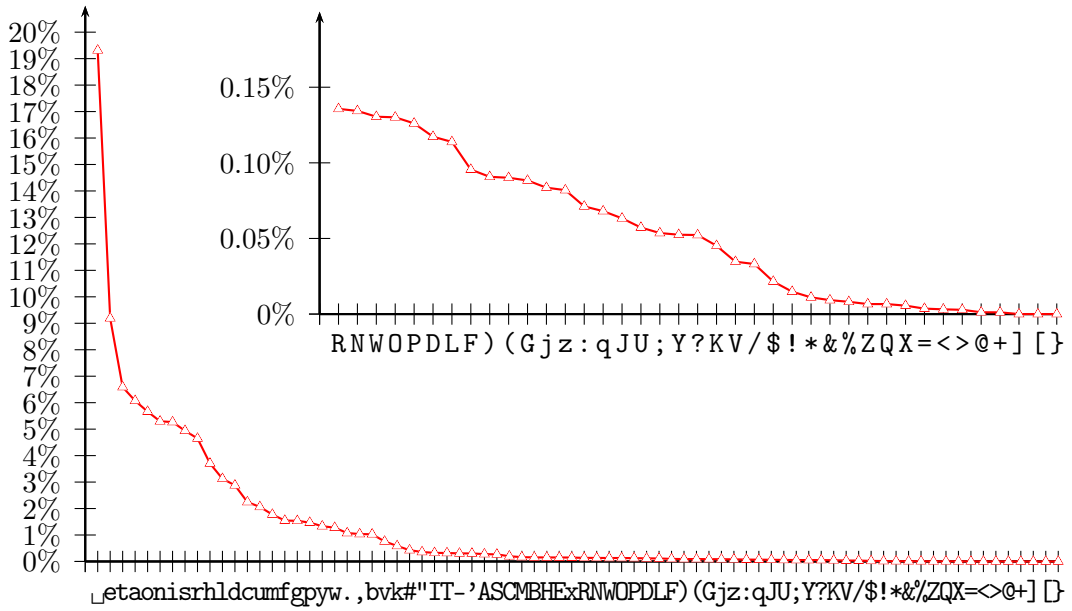

Figure 3.2: Frequencies of symbol-monograms over the alphabet of all printable ASCII characters. Characters which do not occur in the COCA are not plotted.

3.1. Monogram frequencies. The letter-monogram and the symbol-monogram frequencies of the purged COCA can be found in Figures 3.1 and 3.2, respectively. In both statistics, space ␣ is by far the most frequent character, followed by a number of lowercase Latin letters. From the figures, we directly observe that for both alphabets considered, the ten most frequent letters constitute more that 70% of all characters. From Definition 3.1 (i) we directly obtain the letter-entropy[2] of 4.12 bits and symbol-entropy of 4.46 bits. In the COCA, we have the most frequent word-monograms given in Figure 3.3. In this table the contraction `'s` and the Saxon genitive `'s` are counted

monogram	the	of	and	to	a	in	that	s	for	i
frequency (%)	5.72	2.74	2.73	2.52	2.34	1.89	1.10	0.97	0.89	0.85

Figure 3.3: The most frequent word-monograms.

as the single word-monogram `s`. For the frequency distribution, we obtain a word-entropy of 11.05 bits . Concerning string-monograms, we observe that the punctuation symbol `,` occurs most often, followed by `.` and some of the above most frequent word-monograms from Figure 3.3. As a side remark, we also list the most frequent nouns from the COCA in Figure 3.4.

monogram	time	people	years	way	year	world	day	life
frequency (‰)	1.55	1.30	1.12	0.96	0.82	0.74	0.74	0.69

Figure 3.4: The most frequent nouns in the COCA.

3.2. Polygram frequencies. Concerning polygram frequencies, we first analyzed the most frequent letter-digrams in the COCA:

digram	th	he	in	er	an	re	on	at	en	nd
frequency (%)	10.19	9.32	7.78	6.47	6.20	5.57	4.96	4.50	4.30	4.06

Figure 3.5: The most frequent letter-digrams.

For the letter-digram entropy, Definition 3.1 (i) yields 7.55 bits, and for the symbol-entropy, 8.01 bits per digram. This gives by Definition 3.1 (ii) a conditional letter-digram entropy of 3.43 bits and a conditional symbol-entropy of 3.56 bits .

For larger n, we obtain for the conditional letter-n-gram entropy the values given in Figure 3.6.

The figure indicates that the sampling errors introduced by considering the frequency distribution of successive n-grams grow as n gets larger. An intuitive explanation of this behavior might be the following: A fixed corpus of length ℓ is for growing n a decreasingly representative sample. Eventually we arrive at a value for n, for which

[2]The entropy of the distribution of lowercase Roman letters *without space* is 4.19.

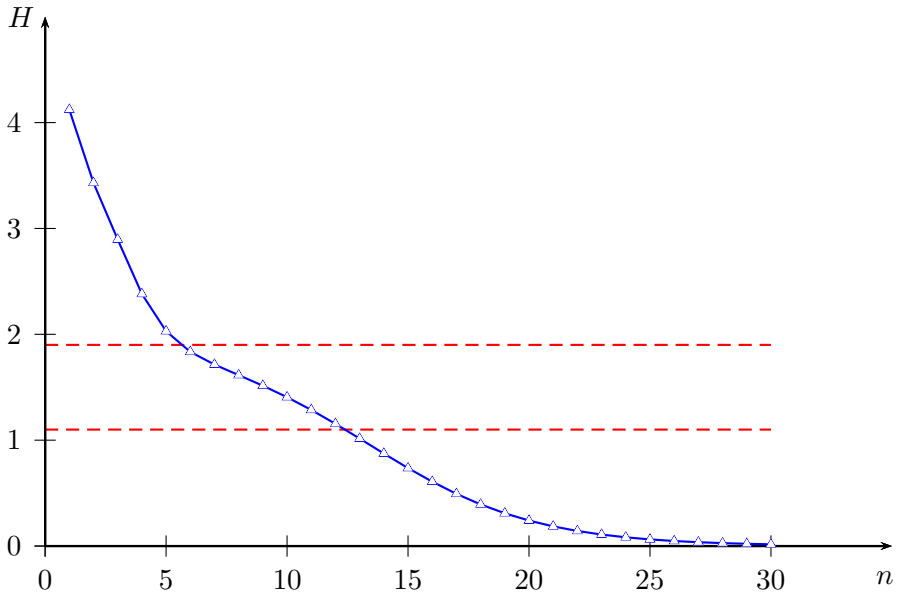

Figure 3.6: Conditional letter-n-gram entropy of the COCA for $n = 1, \ldots, 30$. The dashed red lines are Shannon's 1951 bounds for the entropy of English.

each of the $\ell - n + 1$ occurring n-gram appears exactly once in the corpus and we get absolute entropy $\log_2(\ell - n + 1)$. But then, each of the $\ell - n$ occurring $(n+1)$-gram also appears only once in the corpus and we get relative entropy $\log_2(1 + \frac{1}{\ell - n}) \approx \frac{1}{\ell - n}$, which is close to zero for the relevant choices for ℓ and n.[3]

The left-hand part of Figure 3.6 seems to indicate an English entropy around 1.5 bits, but this is pure guesswork. The only conclusion we can definitely draw is that the entropy is below 3—presumably a rather poor estimate. For a more detailed analysis of what happens, see Section 5.

One challenge in the statistical analysis of the COCA was to actually fit the frequency distributions of polygrams into computer memory: The corpus consists of roughly $2^{31} \approx 10^{21.49}$ characters. If we only *stored* all n-grams of the corpus in memory (to actually start the analysis), we would need approximately $n \cdot 2^{31}$ bytes, which starts getting impractical already for $n = 4$. Thus, one either needs to refrain from the idea of storing the data in memory (but use slow hard-drives instead) or develop an approach that uses a certain amount of memory which does not grow so fast with n. This can be achieved by considering the number of different n-grams in the corpus. Consider a given set of monograms M with $k = \#M$ letters. Then there are at most k^n different n-grams in the corpus. For our alphabets, we have $k = 27$ letters and $k = 95$ (more precisely, $k = 89$ as explained above) different symbols.

[3] In the purged COCA, we have $\ell \approx 2^{31}$. Even if only for $n > 2^{30}$ each letter n-gram occurs only once, we still get relative letter entropy $\log_2(1 + 2^{-30}) \approx 2^{-30} \approx 0$.

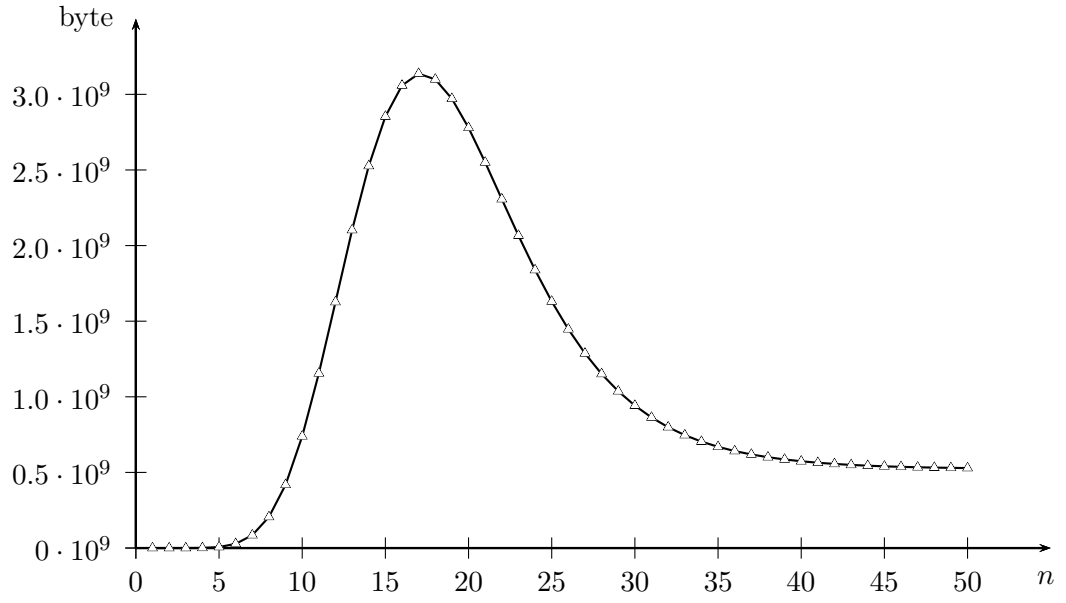

Figure 3.7: Storage requirements for the repeated n-grams over the alphabet of lowercase Latin letters and space in the COCA for successive n.

1. For small n, the number of different n-grams is comparatively small. Also, one expects for these choices of n that almost each n-gram occurs.

2. For very large n, we have an enormous amount of different n-grams. However, most of them either do not occur at all or just once.

Trading between these two extremes, the idea is to track *repeated* n-grams only, that is, n-grams that occur at least twice in the corpus. Then one expects for both small n and large n that the number of stored repeated n-grams is not too large. This in turn leads to the following algorithm: We read the corpus monogram by monogram and store for each monogram the position at which the monogram occurs. We then recursively use the frequency distribution of repeated $(n-1)$-grams to compute the frequency distribution of repeated n-grams. In this algorithm each position is stored using a 4 byte unsigned integer. We thus store for the full COCA with approximately $2 \cdot 2^{30}$ positions roughly $8 \cdot 2^{30} = 8\,\text{GB}$ in memory. Additionally, we have to store each occurring repeated n-gram. In Figure 3.7, the storage requirements for the occurring repeated n-grams are plotted.

From the figure, we see that the storage requirements grow monotonically for n from 1 to 17 and decrease afterwards monotonically. At max, we have to store roughly $3.13 \cdot 10^9 \approx 2.92 \cdot 2^{30}$ bytes. Thus, the whole algorithm requires at most 10.92 GB memory, which fits into decent hardware such as our small Intel Xeon cluster.

The occurrences of repeated n-grams deviate from what one would initially suspect. For growing n, the *number* of repeated n-grams tends to zero. However, one would expect that this happens much earlier than here. Indeed, there are whole parts of sentences that occur several times in the corpus and are thus counted repeatedly. For example, the 16-gram `indeed␣there␣are`, which is quite frequent in academic texts and incidentally also the beginning of the previous sentence, occurs 135 times in the COCA and is also counted 135 times.

A further problem in the statistical analysis of the COCA are inherent problems with the precision of the floating point arithmetic for computing the entropy. Since by Definition 3.1 (i), the entropy is computed a sum of k^n summands of the form $p \log_2 p$, tiny errors in the evaluation of the log-function may accumulate and lead to large errors in the evaluation of the entropy value. To circumvent this, we decided to use `mpfr`, a C library for multiple-precision floating-point computations with correct rounding, see Fousse, Hanrot, Lefèvre, Pélissier & Zimmermann (2007). Specifically, the use of arbitrary-precision arithmetic enabled us then to evaluate the entropy correctly.

4. Reproducibility of the results

We explain now how inclined readers of the current work can reproduce our results for written English, but also generalize the given methodologies for other written languages. The restriction to written languages is based on the definition of language as "a set (finite or infinite) of sentences, each finite in length and constructed out of a finite set of elements", see Chomsky (1957).

This is the basis of all our analyses: We start from the finite set of elements (as described in Section 2, these are Roman letters with some special punctuation marks in the case of written English) and perform statistical computations as described in Section 3.

To reproduce our results for written English, one needs a source of samples from the language. As explained in the introduction, we decided to use in our case Davies's COCA corpus, but any other source will do as well. From the selected source, all 1-grams, 2-grams, etc. are extracted and the corresponding probabilities (and possibly the positions) of the occurrences are stored. Then, one can use this as a basis for computing several relevant values, such as the (conditional) entropy as given in Definition 3.1. Of course, other metrics, such as for example the repetition rate can be computed as well. The results of the computation will exhibit the following behavior: when considering n-grams for growing n, the quality of the statistical results get worse and worse. This is intrinsic to the sampling methodology as we will show in the subsequent sections.

The techniques can be used for other (written) languages as well. For languages using a "small" alphabet, say with up to 100 characters and for which large corpora are available, the methods used for written English can be used as such and a comparison

with the results for English should be easily possible. It will be interesting to compare with languages having a large alphabet (such as Chinese), or a tiny corpus (such as Rongorongo), or a fairly large alphabet but fairly small corput (such as ancient Egyptian). This is out of scope of this treatise.

Similarly, one can extend our results to a phonetic representation of a language. There, the set of basic elements would be the phonetic characters and the computations would then be performed over concatenations thereof.

We will now show how we can model formally any kind of language using basic elements in the sense of Chomsky. Although our motivation comes from experiments with the English language, our arguments apply to a large class of languages.

5. A stochastic model for natural language entropy

It is well known that when we consider the language under consideration as a stationary random process

$$X = (\ldots, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots),$$

over a finite set M of $k \in \mathbb{N}_{\geq 1}$ monograms, the entropy of the process X is defined as

$$H(X) = \lim_{n\to\infty} H(X_n : X_0, \ldots, X_{n-1}).$$

If the language is an ergodic process, then for any n-gram $(x_0, \ldots, x_{n-1}) \in M^n$, we have by the Shannon-McMillan-Breiman theorem (see for example Algoet & Cover 1988) almost surely (over the choice of $(x_0, \ldots, x_{n-1})$)

$$H(X) = \lim_{n\to\infty} -\frac{1}{n} \log_2(\mathrm{prob}((X_0, \ldots, X_{n-1}) = (x_0, \ldots, x_{n-1}))), \tag{5.1}$$

Thus, one can compute the entropy of for such a process by looking at sufficiently long samples and computing the relative entropy of the distribution of successive n-grams. We have the following well-known result.

Fact 5.2. *For a stationary ergodic stochastic process X, $H(X_n : X_0, \ldots, X_{n-1})$ is non-increasing in n and has a limit $H'(X)$.*

Proof. By assumption

$$H(X_n : X_0, \ldots, X_{n-1}) \leq H(X_n : X_1, \ldots, X_{n-1}) = H(X_{n-1} : X_0, \ldots, X_{n-2}),$$

since X is stationary. Thus $H(X_{n-1} : X_0, \ldots, X_{n-2})$ is nonnegative and non-increasing and has a limit $H'(X)$. □

This is consistent with our observations depicted in Figure 3.6. By the chain rule we have $H(X_{n-1} : X_0, \ldots, X_{n-2}) = H(X_0, \ldots, X_{n-1}) - H(X_0, \ldots, X_{n-2})$ since X is stationary. Since X is ergodic, one might want to approximate it as in (5.1) by looking at sufficiently many examples.

We will show in the following that the amount ℓ of text one needs for precise computations of the entropy of the language is too large to be feasible. Thus, any such approach can in principle only illuminate a limited part of the linguistic truth. To complete the picture, we also show how large a corpus needs to be *at most* for expectedly precise entropy computations.

Our observations are consistent with results from the theory of computational complexity on this question. Namely, Goldreich, Sahai & Vadhan (1999) show that determining the entropy of a distribution is hard for the complexity class NISZK of non-interactive statistical zero-knowledge. Here, a program tries to approximate the entropy of a distribution on about 2^n elements by just asking for samples according to the distribution; together these samples make up a corpus. Their result says that (under usual complexity-theoretic assumptions) it is infeasible to obtain good approximations to the entropy.

5.1. Description of the model. As above, we consider the language as a strongly stationary ergodic stochastic process $X = (\ldots, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots)$ over the set of $k \geq 1$ monograms $M = \{m_1, \ldots, m_k\}$. To simplify our analysis, we additionally assume in our model that for some $n \in \mathbb{N}_{\geq 1}$ the probability for the occurrence of a specific monogram only depends on the previous n letters. In other words, we model X as a homogeneous nth order Markov process. This is a frequently used stochastic model for English, a nice survey of other possible models can be found in Rosenfeld (2000).

For our analysis we are interested in the n-grams that come from X. Thus, we define a second process $X^{\bar{n}} = (\ldots, X^{\bar{n}}_{-2}, X^{\bar{n}}_{-1}, X^{\bar{n}}_0, X^{\bar{n}}_1, X^{\bar{n}}_2, \ldots)$ of n-grams, where for each $i \in \mathbb{Z}$, we define $X^{\bar{n}}_i = (X_i, \ldots, X_{i+n-1})$. The process $X^{\bar{n}}$ is now by construction a first order homogeneous Markov process. Thus, there are for any $x, y \in M^n$ (unknown) transition probabilities $T_n \colon M^n \times M^n \to \mathbb{R}$ for the process $X^{\bar{n}}_i$ induced by the language considered that specify the probability $T(x, y)$ of occurrence of a certain n-gram x given that the previous n-gram was y. Thus $T(x, y) = 0$ unless x and y overlap in all but one letter.

The stationary distribution $S_n(x)$ of the process $X^{\bar{n}}$ is the probability that a certain n-gram is observed, and defined as

$$\begin{aligned} S_n(x) &= \operatorname{prob}(X^{\bar{n}}_i = x) \\ &= \sum_{y \in M^n} \operatorname{prob}(X^{\bar{n}}_{i-1} = y) \operatorname{prob}(X^{\bar{n}}_i = x : X^{\bar{n}}_{i-1} = y) \\ &= \sum_{y \in M^n} \operatorname{prob}(X^{\bar{n}}_{i-1} = y) \cdot T_n(x, y) \end{aligned}$$

for $x \in M^n$. This distribution is well-defined if the underlying Markov process is *irreducible* and *recurrent*. This assumption seems to hold for English, and we will take it for granted in the following.

We define the *observed distribution* (in information theory also called the *type*) of the n-grams induced by the process X over a set of values $\mathrm{dom}(X) = M$ when observing $\ell \in \mathbb{N}_{\geq 0}$ consecutive outcomes by

$$p_n^\ell(X)\colon \begin{array}{rcl} \mathrm{dom}(X)^n & \longrightarrow & \frac{1}{\ell}\mathbb{Z}, \\ x & \longmapsto & \frac{1}{\ell}\#\left\{0 \leq i < \ell \,\middle|\, X_i^{\bar{n}} = x\right\}. \end{array} \tag{5.3}$$

Thus, for $x \in \mathrm{dom}(M^n)$, we have $p_n^\ell(X) = \frac{1}{\ell}\sum_{0\leq i<\ell} \mathbb{1}_{X_i^{\bar{n}}=x} =$, where $\mathbb{1}_{X_i^{\bar{n}}=x}$ is the indicator function of the predicate $X_i^{\bar{n}} = x$, that is, $\mathbb{1}_{X_i^{\bar{n}}=x} = 1$ if the ith n-gram in the process $X^{\bar{n}}$ is $x \in M^n$ and $\mathbb{1}_{X_i^{\bar{n}}=x} = 0$ otherwise. The observed distribution $p_n^\ell(X)$ is a random variable with values in the finite set

$$P_n^\ell(X) = \left\{p\colon \mathrm{dom}(X)^n \longrightarrow \frac{1}{\ell}\mathbb{Z} : p_n^\ell(X) = p\right\} \tag{5.4}$$

of all possible observable distributions induced by corpora of length ℓ.

The problem is now to estimate how far the conditional entropy $H(p_n^\ell(X) : p_{n-1}^\ell(X)) = H(p_n^\ell(X)) - H(p_{n-1}^\ell(X))$ of the observed distribution differs from the conditional entropy $H(S_n : S_{n-1})$ of the stationary distribution. Suppose we have $|H(p_{n-1}^\ell(X)) - H(S_{n-1})| \leq \varepsilon_{n-1}$ and $|H(p_n^\ell(X)) - H(S_n)| \leq \varepsilon_n$ for some $\varepsilon_{n-1}, \varepsilon_n > 0$. Then

$$|H(p_n^\ell(X) : p_{n-1}^\ell(X)) - H(S_n : S_{n-1})| \leq \varepsilon_{n-1} + \varepsilon_n$$

by the triangle inequality. In other words, it is sufficient to estimate when the observed entropies $H(p_{n-1}^\ell(X))$ and $H(p_n^\ell(X))$ differ only slightly from the true entropies $H(S_{n-1})$ and $H(S_n)$, respectively, in order to be able to deduce corresponding results for the conditional entropy. We will thus restrict our attention to the entropy only.

It is easy to establish an upper bound on the observed entropy $H(p_n^\ell(X))$. Because log is a concave function, the entropy in Definition 3.1(i) attains its maximum if $p_n^\ell(X)$ is the uniform distribution. Since there are in total $\#M^n = k^n$ possible n-grams and we consider exactly ℓ consecutive n-grams, we obtain the upper bound

$$H(p_n^\ell(X)) \leq \min(n\log_2(k), \log_2(\ell)). \tag{5.5}$$

We now analyze the behavior of the expectation $E(H(p_n^\ell(X)))$. Our primary goal is to establish a lower bound on ℓ for which the difference $|E(H(p_n^\ell(X))) - H(S_n)|$ is bounded from below. This in turn leads to the conclusion on how large we have to select the corpus size ℓ at least to be able to approximate the correct value of the

entropy with small error. The second goal is to provide an upper bound on ℓ giving an appropriate upper bound on $|E(H(p_n^\ell(X))) - H(S_n)|$. Having this allows us to conclude how large a corpus has to be *at most* for a useful entropy approximation.

By the definition of the expectation of a random variable, we have

$$E(H(p_n^\ell(X))) = \sum_{p \in P_n^\ell(X)} H(p) \operatorname{prob}(p_n^\ell(X) = p). \tag{5.6}$$

Before we consider this expression in full generality, we first discuss a special case which is easy to analyze. Afterwards we will argue that a similar reasoning also holds for arbitrary distributions.

5.2. Randomspeak. We now restrict ourself to the special case that S_n is the uniform distribution U_{M^n} on M^n, that is, for $x \in M^n$, we have $S_n(x) = U_{M^n}(x) = \frac{1}{k^n}$. We know that $H(S_n) = \log k^n$, but now ask how this value can be approximated by observations on corpora of some length ℓ. Indeed, in this case the desired bounds on ℓ can be derived. Consider the *necessary* size of ℓ first. By (5.5) we have $H(p_n^\ell(X)) \leq \min(n \log_2(k), \log_2(\ell))$ and by (5.6) also

$$E(H(p_n^\ell(X))) \leq \min(n \log_2(k), \log_2(\ell)). \tag{5.7}$$

Thus, $\log_2 k^n - E(H(p_n^\ell(X))) \geq \log_2 \frac{k^n}{\ell}$. In order to approximate the true value $\log_2 k^n$ with relative error at most $\alpha > 0$, we consider the inequality $\log_2 \frac{k^n}{\ell} \leq \alpha \log_2 k^n$ and solve for ℓ, giving

$$\ell \geq k^{(1-\alpha)n}. \tag{5.8}$$

This says, for example, that when we want to approximate the n-gram entropy with relative error at most $\alpha = 0.05$ over an alphabet with $k = 27$ letters, COCA does not provide sufficient data about the n-gram entropy for $n > 6$. If we wanted to say something about 10-grams, we would already need a corpus with at least 36 TB of text. A corpus of the storage size used by all of humankind[4] would let us look until $n = 15$, but even this does not provide enough data for $n > 15$.

We will now analyze which corpus size ℓ is *sufficient* for good entropy approximations by sampling only. Consider a distribution $p\colon M^n \to \mathbb{R}$ which is close to uniform. More specifically, assume that the statistical distance is bounded by $\delta \in \mathbb{R}_{>0}$ so that $\|p - U_{M^n}\|_\infty < \delta$. Then by definition of the max-norm, we have for all $g \in M^n$ that $|p(x) - k^{-n}| < \delta$, that is, $p(x) \in [k^{-n} - \delta, k^{-n} + \delta]$. Consequently, $-\log_2 p(x) \in [-\log(k^{-n} + \delta), -\log(k^{-n} - \delta)]$ and $-p(x) \log_2 p(x) \in [-(k^{-n} - \delta) \log_2(k^{-n} + \delta), -(k^{-n} + \delta) \log_2(k^{-n} - \delta)]$. We thus obtain the lower bound

$$\begin{aligned} H(p) &= -\sum_{x \in M^n} p(x) \log_2 p(x) \\ &\geq -(1 - \delta k^n) \log_2(k^{-n} + \delta). \end{aligned}$$

[4]Following Hilbert & López (2011) the storage used nowadays is estimated as 295 exabyte.

Thus, we have for the expected value

$$\begin{aligned} E(H(p_n^\ell(X))) &= \sum_{p\in P_n^\ell(X)} H(p)\,\mathrm{prob}(p_n^\ell(X)=p) \\ &\geq \sum_{\substack{p\in P_n^\ell(X)\\ \|p-U_{M^n}\|_\infty<\delta}} H(p)\,\mathrm{prob}(p_n^\ell(X)=p) \\ &\geq -(1-\delta k^n)\log_2(k^{-n}+\delta) \sum_{\substack{p\in P_n^\ell(X)\\ \|p-U_{M^n}\|_\infty<\delta}} \mathrm{prob}(p_n^\ell(X)=p) \\ &= -(1-\delta k^n)\log_2(k^{-n}+\delta)\,\mathrm{prob}(\|p_n^\ell(X)-U_{M^n}\|_\infty<\delta). \end{aligned} \tag{5.9}$$

Without loss of generality, assume that n divides ℓ. Otherwise, pad the corpus accordingly. Since $p_n^\ell(X)(x)=\frac{1}{\ell}\sum_{0\leq i<\ell}\mathbb{1}_{X_i^{\bar{n}}=x}$, the idea is to split for $x\in M^n$ the relative counts $p_n^\ell(X)(x)$ into n independent parts, that is, consider for $0\leq j<n$ the relative counts

$$p_{n,j}^\ell(X)(x)=\frac{1}{\ell}\sum_{\substack{0\leq i<\ell\\ i=j \text{ in } \mathbb{Z}_n}}\mathbb{1}_{X_i^{\bar{n}}=x}=\frac{1}{\ell}\sum_{0\leq i<\frac{\ell}{n}}\mathbb{1}_{X_{in+j}^{\bar{n}}=x},$$

which are the occurrences of x at positions that fall into residue class j modulo n and apply then Hoeffding's inequality to get a bound for $\mathrm{prob}(\|p_n^\ell(X)-U_{M^n}\|_\infty<\delta)$.

Recall that Hoeffding's inequality states that when we have ℓ independent random variables $X_0,\ldots,X_{\ell-1}$ such that almost surely $a_i\leq X_i-E(X_i)\leq b_i$, then for all positive real constants $\varepsilon\in\mathbb{R}_{\geq 0}$ we have

$$\mathrm{prob}(|\sum_{0\leq i<\ell}(X_i-E(X_i))|\geq\varepsilon)\leq 2\exp(\frac{-2\varepsilon^2}{\sum_i(b_i-a_i)^2}).$$

Let $0\leq j<n$ and set $Y_i=\mathbb{1}_{X_{in+j}^{\bar{n}}=x}$ for $0\leq i<\frac{\ell}{n}$. Then the random variables $Y_0,\ldots,Y_{\frac{\ell}{n}-1}$ are independent and $E(Y_i)=k^{-n}$ for all i. Furthermore, we have $-k^{-n}\leq Y_i-E(Y_i)\leq 1-k^{-n}$ and Hoeffding's inequality gives for any $x\in M^n$ and $\varepsilon>0$

$$\begin{aligned} \mathrm{prob}(|p_{n,j}^\ell(X)(x)-\frac{1}{n}k^{-n}|\geq\frac{\varepsilon}{\ell}) &= \mathrm{prob}(|\frac{1}{\ell}\sum_{0\leq i<\frac{\ell}{n}}\mathbb{1}_{X_{in+j}^{\bar{n}}=x}-\frac{1}{n}k^{-n}|\geq\frac{\varepsilon}{\ell}) \\ &= \mathrm{prob}(|\sum_{0\leq i<\frac{\ell}{n}}Y_i-\sum_{0\leq i<\frac{\ell}{n}}E(Y_i)|\geq\varepsilon) \\ &= \mathrm{prob}(|\sum_{0\leq i<\frac{\ell}{n}}(Y_i-E(Y_i))|\geq\varepsilon)\leq 2\exp(\frac{-2n\varepsilon^2}{\ell}). \end{aligned}$$

Setting $\delta = \varepsilon/\ell$, we get

$$\text{prob}(|p^\ell_{n,j}(X)(x) - \frac{1}{n}k^{-n}| \geq \delta) \leq 2\exp(-2n\delta^2\ell).$$

Note that $p^\ell_n(X)(x) = \sum_{0\leq j<n} p^\ell_{n,j}(X)(x)$. By the triangle inequality we have

$$\begin{aligned}
\text{prob}(|p^\ell_n(X)(x) - k^{-n}| \geq n\delta) &\leq \text{prob}(\sum_{0\leq j<n} |p^\ell_{n,j}(X)(x) - \frac{1}{n}k^{-n}| \geq n\delta) \\
&\leq \text{prob}(\ \exists 0 \leq j < n : |p^\ell_{n,j}(X)(x) - \frac{1}{n}k^{-n}| \geq \delta) \\
&\leq n \cdot \text{prob}(\ |p^\ell_{n,0}(X)(x) - \frac{1}{n}k^{-n}| \geq \delta) \\
&\leq 2n\exp(-2n\delta^2\ell).
\end{aligned}$$

We obtain

$$\begin{aligned}
\text{prob}(\|p^\ell_n(X) - U_{M^n}\|_\infty \geq \delta) &= \text{prob}(\max_{x\in M^n} |p^\ell_n(X)(x) - k^{-n}| \geq \delta) \\
&\leq \sum_{x\in M^n} \text{prob}(|p^\ell_n(X)(x) - k^{-n}| \geq \delta) \\
&\leq 2k^n n\exp(-2n\delta^2\ell).
\end{aligned} \tag{5.10}$$

Plugging this into (5.9) gives

$$E(H(p^\ell_n(X))) \geq -(1-\delta k^n)\log_2(k^{-n}+\delta)(1-2k^n n\exp(-2n\delta^2\ell)). \tag{5.11}$$

This says, for example, that when we want to approximate the n-gram entropy with relative error $\alpha = 0.05$ and $\varepsilon = 0.05$ over an alphabet with $k = 27$ letters, COCA tells us only for sure a good approximation on the entropy for $n \leq 2$. A corpus of the storage size used by all of humankind, that is, 295 exabyte, definitely provides sufficient data for $n \leq 6$. We summarize our result in Figure 5.1.

Thus, we have satisfactory results in the case of randomspeak: First, it is infeasible to approximate the entropy by looking at increasingly long n-grams. Second, the amount of text we need to look at *at most* is enormous and a corpus of size of the COCA can serve as a basis for estimating the n-gram entropy for $n = 1$ and $n = 2$, since for these values the corpus size sufficient for good approximations lower than COCA's size. For $n = 3, \ldots, 6$ we do not know whether COCA is sufficiently large, but we know it is larger than what is necessary for a good entropy approximation by sampling. For $n > 6$, sampling cannot be used to estimate the entropy, since the necessary size of a corpus exceeds the one of COCA.

We are able to obtain the same result numerically, using the data given in Figure A.5: When we interpolate the data linearly by the equation

$$y = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1) + y_1,$$

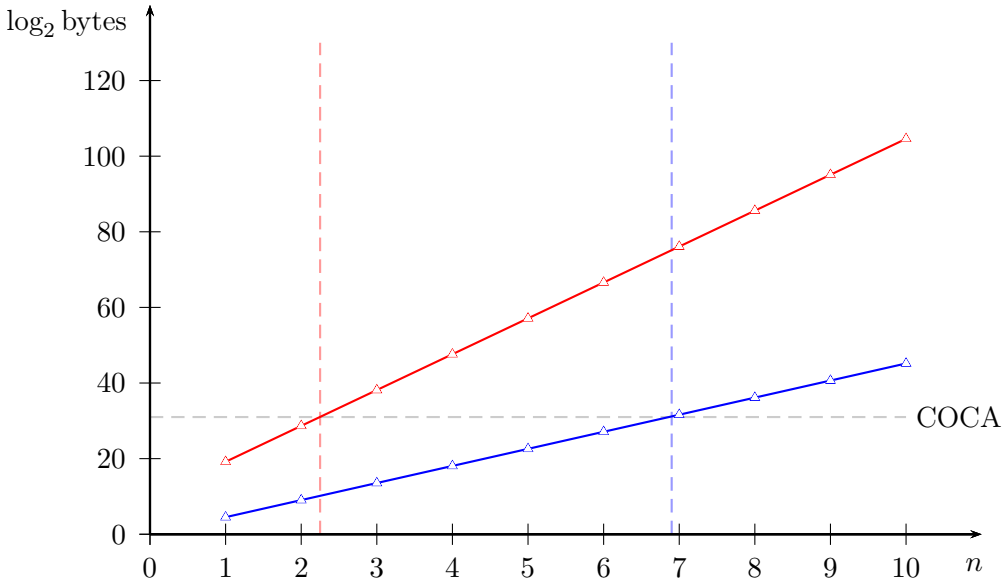

Figure 5.1: Logarithmic scale lower and upper bounds on the corpus size ℓ for approximating the entropy of randomspeak with relative error $\alpha = 0.05$ over an alphabet with 27 letters, instantiating (5.8) and (5.11). The horizontal gray dashes show the size of COCA and the vertical light red and light blue dashes pass through their intersections with the lines of corresponding color.

for points (x_1, y_1) and (x_2, y_2) on the line, we obtain that the lower bound is approximately given by the line equation

$$y = 4.517143127 \cdot x$$

and the upper bound is approximately given by

$$y = 9.49987258 \cdot x + 9.66950108$$

where we use for both extrapolations the corresponding points at $x_1 = 1$ and $x_2 = 2$, respectively. COCA's size is roughly 2^{31} bytes, which tells us that the upper bound is larger than the horizontal line at $y = 31$ for $x > 2.24$ and the lower bound for $x > 6.86$. This is consistent with our deduction based on (5.8) and (5.11).

5.3. Markov sampling. We will now argue that for non-uniform stationary distributions we also have the stated trichotomy. In fact, it seems that randomspeak is the worst case that can happen when sampling.

For the lower bound this is easy to see. In fact, we can proceed similarly as in the beginning of Section 5.2, but this time we do not assume anything about the Markov transition probabilities T_n (and thus the stationary distribution S_n). From (5.7), we know that we have for the expected entropy $E(H(p_n^\ell(X))) \leq \min(n \log_2(k), \log_2(\ell))$.

Thus, $H(S_n) - E(H(p_n^\ell(X))) \geq H(S_n) - \log_2 \ell$. In order to approximate the true value $H(S_n)$ with relative error at most $\alpha > 0$, we consider the inequality $H(S_n) - \log_2 \ell \leq \alpha H(S_n)$ and solve for ℓ, giving $\ell \geq 2^{(1-\alpha)H(S_n)}$. Since $H(S_n) \leq n \log_2 k$, this bound is indeed weaker than the corresponding bound in (5.8) and thus says that for non-uniform S_n, a smaller corpus is necessary for a good approximation of the entropy of English than for the case of randomspeak.

It remains to argue that this is also true for the *necessary* corpus size. Thus, one has to study difference between randomspeak and a Markov process with unknown (possibly non-uniform) transition probabilities T_n. In order to do so, we used successive approximations to English. Specifically, we computed for every $m \geq 1$ from the COCA the frequency of letter m given the previous $m-1$ letters and generated equally long texts randomly corresponding to the respective distributions. This well-known procedure gives for $m = 1$ exactly randomspeak, while for growing m the resulting random language from the $(m+1)$st order Markov process approaches English better and better.

Afterwards, we computed for all of the generated texts the n-gram entropy values for successive n and plotted the result, see Figure 5.2.

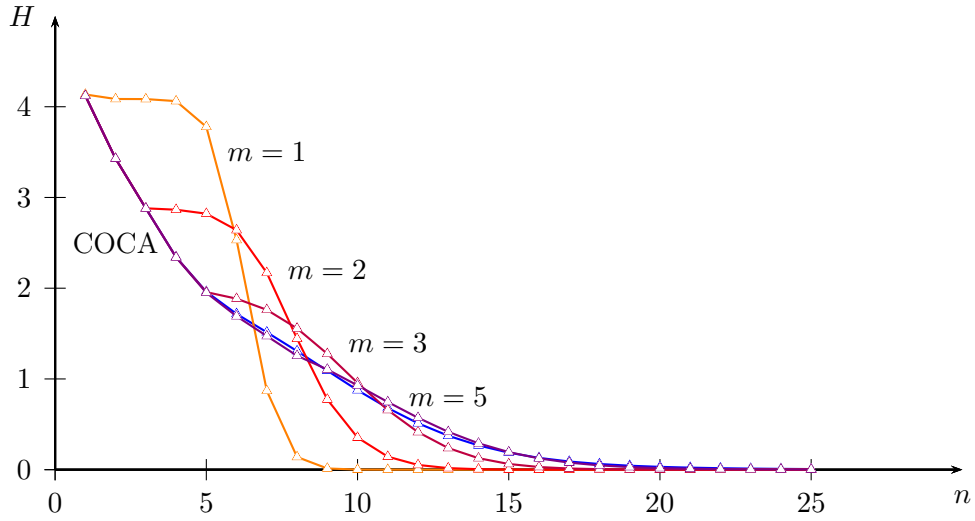

Figure 5.2: Letter entropy of a $\frac{1}{128}$-fraction of the COCA (blue) vs. letter entropy of zeroth order (orange), first order (red), second order (purple) and fourth order (violet) approximations.

The figure shows that in the case $m = 1$, that is, randomspeak, the conditional entropy value is maximal for very small n and decreases rapidly. When m grows the behavior gets more and more similar to the one of the COCA, where we have a much slower decrease in the entropy values for growing m, thus giving conditional entropy zero much later. This leads to the conclusion that statistical noise in case of English occurs somewhat later than in the case of uniform distributions.

6. The central conjecture

We have proven mathematically that there is a natural trichotomy in the case of randomspeak when analyzing n-grams by sampling: reasonable approximations to the true value of the entropy for very small $n \leq 2$, the truth with some statistical noise for medium sized $2 < n < 7$, and statistical noise only for large $n \geq 7$. We also argued that in the stochastic model a similar trichotomy holds in general and saw that the case of randomspeak is the worst case possible. The result is difficult to quantify, since the entropy of English and thus the specific bounds for the necessary and sufficient corpus size, respectively, are unknown. That this is also true for English (regardless of the model) leads to central conjecture of this article:

CONJECTURE 6.1. *The approximation of the n-gram entropy of English by sampling corpora leads to a natural trichotomy: The linguistic truth for very small n, the truth with some statistical noise for medium size n and statistical noise only for large n.*

If true, we further conjecture that it holds for other languages with "small" alphabet and "large" corpora, if they follow an irreducible recurrent Markov process.

We performed further experiments to underpin this conjecture. The idea was to analyze how the entropy of a representative fraction of the COCA differs from the entropy value of the full corpus. Our trichotomy conjecture implies that we expect almost no difference for very small and very large n, since in the former case we computed in both cases a good approximation to the true value and in the latter case we anyways have in both cases statistical noise only. The result of such an analysis are depicted in Figure 6.1.

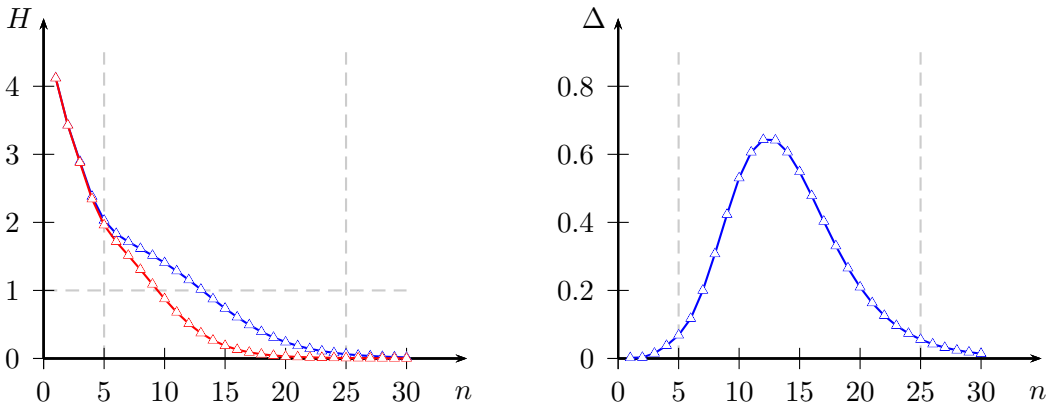

Figure 6.1: Illustration of the trichotomy. Left: Letter-entropy of the full COCA (blue) and a $\frac{1}{128}$-fraction thereof (red). Right: Absolute distance Δ between the two entropy values.

The figure indicates that in the case of English, we have a good approximation to the true value of the n-gram entropy for $n \leq 5$. For $n = 14$ the measured n-gram

entropy drops the first time below 1, which means that the statistical noise seems to dominate from this point in time on, resulting in noise only beyond $n \geq 25$. This observation is also consistent with our observation above that led us to the conclusion that randomspeak is in fact the worst case that can happen.

7. Bounding the expected entropy

We finish by giving bounds on the expectation (5.6) of the entropy in our Markov model described in Section 5.1. Recall that it is defined as

$$E(H(p_n^\ell(X))) = \sum_{p \in P_n^\ell(X)} H(p) \operatorname{prob}(p_n^\ell(X) = p).$$

First, let us compute the probability that an observed sequence $(X_0^{\bar{n}}, \ldots, X_\ell^{\bar{n}})$ of n-grams is equal to a fixed given one. To do so, we use the distribution $p_2^\ell(X^{\bar{n}})$ of consecutively occurring n-grams, that is, the bigram distribution of the process $X^{\bar{n}}$. We have for $(x_0^{\bar{n}}, \ldots, x_{\ell-1}^{\bar{n}}) \in (M^n)^\ell$ with $\operatorname{prob}(X_0^{\bar{n}} = x_0^{\bar{n}}) \neq 0$:

$$\begin{aligned}
&\frac{1}{\operatorname{prob}(X_0^{\bar{n}} = x_0^{\bar{n}})} \operatorname{prob}((X_0^{\bar{n}}, \ldots, X_{\ell-1}^{\bar{n}}) = (x_0^{\bar{n}}, \ldots, x_{\ell-1}^{\bar{n}})) \\
&\qquad = \prod_{1 \leq i < \ell} \operatorname{prob}(X_i^{\bar{n}} = x_i^{\bar{n}} : X_{i-1}^{\bar{n}} = x_{i-1}^{\bar{n}}) \\
&\qquad = \prod_{x,y \in M^n} T_n(x,y)^{\ell \cdot p_2^\ell(x^{\bar{n}})(x,y)} \qquad (7.1) \\
&\qquad = \prod_{x,y \in M^n} 2^{\ell \cdot p_2^\ell(x^{\bar{n}})(x,y) \log_2 T_n(x,y)} \\
&\qquad = 2^{-\ell \cdot (H\left(p_2^\ell(x^{\bar{n}}) \,\|\, T_n\right) + H(p_2^\ell(x^{\bar{n}})))},
\end{aligned}$$

where we write

$$H\left(p_2^\ell(x^{\bar{n}}) \,\|\, T_n\right) = \sum_{x,y \in M^n} p_2^\ell(x^{\bar{n}})(x,y) \log_2 \frac{p_2^\ell(x^{\bar{n}})(x,y)}{T_n(x,y)}$$

for the *conditional entropy* (also called *Kullback-Leibler divergence* or *information gain*) of $p_2^\ell(x^{\bar{n}})$ given T_n, see Kullback & Leibler (1951). The result of (7.1) is the Markov analog to a well-known result for independent draws, see for example Cover & Thomas (2006, section 11.1).

Using (7.1), we can compute the probability that an observed distribution of consecutive n-grams $p_2^\ell(X^{\bar{n}})$ equals a given distribution $q \in P_2^\ell(X^{\bar{n}})$. Assuming that the first n-gram of the corpus was drawn uniformly at random, that is, $\operatorname{prob}(X_0^{\bar{n}} = x_0) = \frac{1}{k^n}$, we have

$$\begin{aligned}\operatorname{prob}(p_2^\ell(X^{\bar n}) = q) &= \sum_{\substack{x\\ p_2^\ell(x^{\bar n})=q}} \operatorname{prob}((X_0^{\bar n},\dots,X_{\ell-1}^{\bar n}) = (x_0^{\bar n},\dots,x_{\ell-1}^{\bar n}))\\ &= \sum_{\substack{x\\ p_2^\ell(x^{\bar n})=q}} \operatorname{prob}(X_0^{\bar n} = x_0)\cdot 2^{-\ell\cdot(H\left(p_2^\ell(x^{\bar n})\,\|\,T_n\right)+H(p_2^\ell(x^{\bar n})))}\\ &= \frac{1}{k^n} c_q \cdot 2^{-\ell\cdot(H(q\,\|\,T_n)+H(q))},\end{aligned} \tag{7.2}$$

writing $x = (\dots, x_{-2}, x_{-1}, x_0, x_1, x_2, \dots)$ for a specific outcome of the process X and c_q for the number of such sequences with $p_2^\ell(x^{\bar n}) = q$. We have the following result in our context.

LEMMA 7.3. *Let* $c_q = \#\left\{x = (\dots, x_{-2}, x_{-1}, x_0, x_1, x_2, \dots)\,\middle|\, p_2^\ell(x^{\bar n}) = q\right\}$. *Then*

$$\frac{1}{(\ell+1)^{k^{n+1}}} k^n 2^{\ell H(q)} \le c_q \le k^n 2^{\ell H(q)}.$$

PROOF. This can be proved as in Cover & Thomas (2006, Theorem 11.1.4) by using (7.2) for the result of the probability of a certain distribution in the context of Markov processes, noting that we have at most k^{n+1} non-zero values in the distribution q. □

The lower bound in the lemma can be substantially improved. In fact, one can replace the constant $(\ell+1)^{k^{n+1}}$ by the much smaller $\#P_2^\ell(Y^{\bar n})$, where $Y^{\bar n}$ is the Markov process with transition probabilities given by $T_n = q$. Jacquet, Knessl & Szpankowski (2012) gave asymptotic estimates for this count when $n = 2$. They proved that up to a constant, we have asymptotically $\#P_2^\ell(Y^{\bar 2}) \approx \frac{\ell^{k^2-k}}{(k^2-k)!}$. The constant they give is dependent on the alphabet size k and expressed as a certain multi-integral. We are not aware of a similar result for arbitrary n. This might also be due to the fact that the behavior of overlapping strings is quite subtle. Guibas & Odlyzko (1981) analyzed this issue and gave fundamental results on the number of strings without a specified pattern. This should give better bounds on $\#P_2^\ell(Y^{\bar 2})$, but since we do not need this for the following, we stick to the lemma as stated above.

It remains to express the expected entropy (5.6) of $p_n^\ell(X)$ in terms of the entropy of the bigram distributions of its n-grams. By the chain rule and noting that the entropy of p differs from the entropy of the marginal distributions of q by at most a factor of 2 we have

$$\begin{aligned}E(H(p_n^\ell(X))) &= \sum_{p\in P_n^\ell(X)} H(p)\operatorname{prob}(p_n^\ell(X) = p)\\ &\in \left[\frac{1}{4}\dots 1\right] \sum_{q\in P_2^\ell(X^{\bar n})} H(q)\operatorname{prob}(p_2^\ell(X^{\bar n}) = q).\end{aligned} \tag{7.4}$$

Combining with (7.2) and Lemma 7.3, we get thus the following bounds on the expected entropy:

$$E(H(p_n^\ell(X))) \in \left[\frac{1}{4\cdot(\ell+1)^{k^{n+1}}}\ldots 1\right] k^n \sum_{q\in P_2^\ell(X^{\bar{n}})} H(q)\cdot 2^{-\ell H(q\,\|\,T_n)}. \tag{7.5}$$

This seems to be difficult to handle in its full generality due to the following reasons:

- Both the transition probabilities T_n of the Markov process and the corresponding stationary distribution S_n from which the samples are taken are unknown.
- Computing the exact number of sequences with a given bigram distribution of n-grams is out of reach at the moment.
- The conditional entropy is not a metric. Specifically, it does not satisfy the triangle inequality.

8. Conclusion

We performed a thorough analysis of Davies's corpus of contemporary American English and computed the entropy values for various alphabets and n-gram lengths. After observing that this gives results incompatible with known ones, we studied why sampling cannot be used for estimating the entropy of English in a satisfactory manner, since the size of the corpus necessary is beyond practical limits. To show this, we set up a simplified Markov model for a natural language like English and argued that sampling procedures for n-grams can only give reasonable approximations of the entropy for very small n and no result at all for large ones, leading to a natural trichotomy. Although our mathematical analysis applies to the artificial language randomspeak, we conjecture that regardless of the model, this trichotomy also applies to English and other languages with similar properties (sizes of alphabets and corpora, Markovian generation, as stated above), and give experimental results to validate this hypothesis.

The fundamental conclusion is that linguistic methods different from our style of computational analysis of orthographic representations are needed to understand the entropy of English.

Acknowledgements

The authors are grateful to Graeme Hirst for advice on the subject and to Reinhard Köhler for enlightening discussions. Their work was funded by the B-IT foundation and the state of Nordrhein-Westfalen.

References

Paul H. Algoet & Thomas M. Cover (1988). A Sandwich Proof of the Shannon-McMillan-Breiman Theorem. *The Annals of Probability* **16**(2), 899–909.

Peter F. Brown, Vincent J. Della Pietra, Robert L. Mercer, Stephen A. Della Pietra & Jennifer C. Lai (1992). An Estimate of an Upper Bound for the Entropy of English. *Computational Linguistics* **18**(1), 31–40. URL http://acl.ldc.upenn.edu/J/J92/J92-1002.pdf.

Noam Chomsky (1957). *Syntactic Structures*. Mouton Publishers, The Hague/Paris. ISBN 9027933855.

Thomas M. Cover & Roger C. King (1978). A Convergent Gambling Estimate of the Entropy of English. *IEEE Transactions on Information Theory* **24**(4), 413–421. URL http://www-isl.stanford.edu/~cover/papers/transIT/0413cove.pdf.

Thomas M. Cover & Joy A. Thomas (2006). *Elements of Information Theory 2nd Edition*. John Wiley & Sons, 2nd edition. ISBN 0471241954 (ISBN-10), 978-0471241959 (ISBN-13).

Mark Davies (2008-2012). The Corpus of Contemporary American English: 450 million words, 1990-present. URL http://corpus.byu.edu/coca/.

Laurent Fousse, Guillaume Hanrot, Vincent Lefèvre, Patrick Pélissier & Paul Zimmermann (2007). MPFR: A Multiple-Precision Binary Floating-Point Library with Correct Rounding. *ACM Transactions on Mathematical Software* **33**(2), 13:1–13:15. URL http://dx.doi.org/10.1145/1236463.1236468.

Joachim von zur Gathen (2015). *CryptoSchool*. Springer Verlag, Heidelberg. 888 pages.

Oded Goldreich, Amit Sahai & Salil Vadhan (1999). Can Statistical Zero Knowledge Be Made Non-interactive? or On the Relationship of $\mathcal{SZK}$ and $\mathcal{NISZK}$. In *Advances in Cryptology: Proceedings of CRYPTO 1999,* Santa Barbara, CA, M. Wiener, editor, volume 1666 of *Lecture Notes in Computer Science*, 467–484. Springer-Verlag, Berlin, Heidelberg. ISBN 978-3-540-48405-9 (Online) 978-3-540-66347-8 (Print). ISSN 0302-9743. URL http://dx.doi.org/10.1007/3-540-48405-1_30.

L. J. Guibas & A. M. Odlyzko (1981). String Overlaps, Pattern Matching, and Non-transitive Games. *Journal of Combinatorial Theory, Series A* **30**(2), 183–208. URL http://dx.doi.org/10.1016/0097-3165(81)90005-4.

Martin Hilbert & Priscila López (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science* **332**(6025). URL http://dx.doi.org/10.1126/science.1200970.

P. Jacquet, C. Knessl & W. Szpankowski (2012). Counting Markov Types, Balanced Matrices, and Eulerian Graphs. *IEEE Transactions on Information Theory* **58**(7), 4261–4272. URL http://dx.doi.org/10.1109/TIT.2012.2191476.

F. W. Kasiski (1863). *Die Geheimschriften und die Dechiffrir-Kunst*. E. S. Mittler und Sohn, Berlin. viii + 95 pp. + 6 tables.

Reinhard Köhler (2016). Private communication.

S. Kullback & R. A. Leibler (1951). On information and sufficiency. *Annals of Mathematical Statistics* **22**(1), 79–86.

R. Rosenfeld (2000). Two decades of statistical language modeling: where do we go from here? *Proceedings of the IEEE* **88**(8), 1270–1278. URL `http://dx.doi.org/10.1109/5.880083`.

C. E. Shannon, (1949). Communication Theory of Secrecy Systems. *Bell System Technical Journal* **28**, 656–715.

C. E. Shannon (1948). A Mathematical Theory of Communication. *Bell System Technical Journal* **27**, 379–423 and 623–656. Reprinted in Claude E. Shannon and Warren Weaver, *The Mathematical Theory Of Communication*, University of Illinois Press, Urbana IL, 1949.

C. E. Shannon (1951). Prediction and Entropy of Printed English. *Bell System Technical Journal* **30**, 50–64. URL `http://www.princeton.edu/~wbialek/rome/refs/shannon_51.pdf`.

A. Numerical results

For better reproducibility of our results, we list here numerical results of our findings. Specifically, we give for each statistical plot of the current work the numerical data underlying it.

Letter	Percentage	Letter	Percentage	Letter	Percentage
␣	17.48	e	10.05	t	7.43
a	6.83	o	6.23	i	6.02
n	5.84	s	5.60	r	5.15
h	4.15	l	3.48	d	3.22
c	2.63	u	2.28	m	2.09
f	1.77	g	1.75	p	1.72
v	0.85	k	0.68	x	0.16
j	0.16	z	0.10	q	$8.00 \cdot 10^{-2}$

Figure A.1: Numerical data for Figure 3.1.

Joachim von zur Gathen
B-IT
Universität Bonn
Dahlmannstr. 2
53113 Bonn
Germany
gathen@bit.uni-bonn.de
http://cosec.bit.uni-bonn.de/

Daniel Loebenberger
B-IT
Universität Bonn
Dahlmannstr. 2
53113 Bonn
Germany
daniel@bit.uni-bonn.de
http://cosec.bit.uni-bonn.de/

Symbol	Percentage	Symbol	Percentage	Symbol	Percentage
␣	19.31	e	9.19	t	6.59
a	6.07	o	5.66	n	5.30
i	5.27	s	4.94	r	4.65
h	3.70	l	3.12	d	2.88
c	2.24	u	2.06	m	1.77
f	1.55	g	1.54	p	1.47
y	1.33	w	1.28	.	1.07
,	1.04	b	1.03	v	0.76
k	0.59	#	0.43	"	0.36
I	0.33	I	0.32	-	0.30
’	0.30	A	0.28	S	0.27
C	0.20	M	0.16	B	0.15
H	0.15	E	0.15	x	0.15
R	0.14	N	0.13	W	0.13
O	0.13	P	0.13	D	0.12
L	0.11	F	$9.55 \cdot 10^{-2}$	)	$9.08 \cdot 10^{-2}$
(	$9.01 \cdot 10^{-2}$	G	$8.84 \cdot 10^{-2}$	j	$8.36 \cdot 10^{-2}$
z	$8.20 \cdot 10^{-2}$	:	$7.12 \cdot 10^{-2}$	q	$6.81 \cdot 10^{-2}$
J	$6.33 \cdot 10^{-2}$	U	$5.72 \cdot 10^{-2}$	;	$5.36 \cdot 10^{-2}$
Y	$5.25 \cdot 10^{-2}$	?	$5.23 \cdot 10^{-2}$	K	$4.52 \cdot 10^{-2}$
V	$3.46 \cdot 10^{-2}$	/	$3.31 \cdot 10^{-2}$	$	$2.15 \cdot 10^{-2}$
!	$1.48 \cdot 10^{-2}$	*	$1.10 \cdot 10^{-2}$	&	$9.22 \cdot 10^{-3}$
%	$8.10 \cdot 10^{-3}$	Z	$6.62 \cdot 10^{-3}$	Q	$6.61 \cdot 10^{-3}$
X	$5.58 \cdot 10^{-3}$	=	$3.61 \cdot 10^{-3}$	<	$3.12 \cdot 10^{-3}$
>	$2.83 \cdot 10^{-3}$	@	$1.20 \cdot 10^{-3}$	+	$1.14 \cdot 10^{-3}$
]	$1.80 \cdot 10^{-6}$	[	$6.16 \cdot 10^{-7}$	}	$9.48 \cdot 10^{-8}$

Figure A.2: Numerical data for Figure 3.2.

n	Entropy	n	Entropy	n	Entropy
1	4.122 3	2	3.430 8	3	2.895 4
4	2.381 4	5	2.029 5	6	1.834 3
7	1.714 5	8	1.614 5	9	1.514 3
10	1.406 2	11	1.285 8	12	1.154 4
13	1.015 3	14	0.873 8	15	0.736 2
16	0.608 1	17	0.492 9	18	0.393 1
19	0.309 4	20	0.240 7	21	0.185 6
22	0.142 3	23	0.108 5	24	0.082 4
25	0.062 5	26	0.047 5	27	0.036 0
28	0.027 5	29	0.021 0	30	0.016 2

Figure A.3: Numerical data for Figure 3.6.

n	Byte	n	Byte	n	Byte
1	27	2	1456	3	53787
4	861416	5	6442885	6	27780660
7	85127511	8	205537080	9	418009311
10	738908360	11	1155084634	12	1626795432
13	2102850516	14	2526503210	15	2853373740
16	3058867728	17	3136688632	18	3099008322
19	2970835060	20	2778861960	21	2549987811
22	2307096836	23	2066215902	24	1838391504
25	1630647350	26	1446641768	27	1287056007
28	1151009720	29	1036822268	30	941963970
31	863460794	32	799113120	33	746578932
34	703882144	35	669317215	36	641373948
37	618802763	38	600526426	39	585757302
40	573754120	41	564009202	42	556096002
43	549656057	44	544449532	45	540244035
46	536847646	47	534095357	48	531887664
49	530129775	50	528773400		

Figure A.4: Numerical data for Figure 3.7.

n	lower bound	upper bound
1	4.517 1	19.169 4
2	9.034 3	28.669 2
3	13.551 4	38.140 8
4	18.068 6	47.620 0
5	22.585 7	57.106 2
6	27.102 9	66.597 4
7	31.620 0	76.092 2
8	36.137 1	85.589 6
9	40.654 3	95.089 0
10	45.171 4	104.589 9

Figure A.5: Numerical data for Figure 5.1. Lower and upper bound are given in $\log_2$ bytes.

n	COCA	$m=1$	$m=2$	$m=3$	$m=5$	$m=8$
1	4.120 4	4.135 1	4.119 3	4.120 3	4.120 3	4.120 6
2	3.427 8	4.085 3	3.426 1	3.427 2	3.428 0	3.427 9
3	2.880 1	4.084 4	3.420 9	2.880 4	2.880 1	2.879 1
4	2.343 4	4.061 9	3.410 7	2.864 9	2.341 3	2.339 5
5	1.960 9	3.778 6	3.325 1	2.821 4	1.956 8	1.950 1
6	1.716 9	2.531 4	2.913 7	2.640 6	1.884 7	1.690 9
7	1.514 4	0.869 0	1.941 7	2.169 2	1.761 9	1.472 6
8	1.307 1	0.139 9	0.853 4	1.443 3	1.557 3	1.255 4
9	1.091 1	0.013 2	0.244 9	0.773 3	1.277 0	1.104 7
10	0.875 7	0.001 0	0.050 3	0.353 6	0.957 3	0.929 1
11	0.679 5	$7.665\,4 \cdot 10^{-05}$	0.008 2	0.143 5	0.657 2	0.746 2
12	0.511 3	$6.792\,3 \cdot 10^{-06}$	0.001 2	0.052 1	0.413 9	0.572 1
13	0.373 6	0.000 0	0.000 2	0.017 1	0.238 2	0.416 9
14	0.267 3	0.000 0	$2.210\,7 \cdot 10^{-05}$	0.005 2	0.126 5	0.290 5
15	0.187 6	0.000 0	0.000 0	0.001 4	0.062 8	0.193 8
16	0.130 0	0.000 0	0.000 0	0.000 4	0.029 2	0.124 2
17	0.090 0	0.000 0	0.000 0	0.000 1	0.012 8	0.077 7
18	0.062 4	0.000 0	0.000 0	$2.612\,2 \cdot 10^{-05}$	0.005 5	0.047 3
19	0.043 6	0.000 0	0.000 0	$4.510\,4 \cdot 10^{-06}$	0.002 2	0.028 4
20	0.031 0	0.000 0	0.000 0	$1.258\,1 \cdot 10^{-06}$	0.000 9	0.017 1
21	0.022 3	0.000 0	0.000 0	0.000 0	0.000 3	0.010 4
22	0.016 2	0.000 0	0.000 0	0.000 0	0.000 1	0.006 3
23	0.011 9	0.000 0	0.000 0	0.000 0	$5.130\,6 \cdot 10^{-05}$	0.004 0
24	0.008 9	0.000 0	0.000 0	0.000 0	$2.310\,3 \cdot 10^{-05}$	0.002 6
25	0.006 7	0.000 0	0.000 0	0.000 0	$8.432\,8 \cdot 10^{-06}$	0.001 7

Figure A.6: Numerical data for Figure 5.2.

n	COCA	$\frac{1}{128}$ COCA	Distance
1	4.122 3	4.120 4	0.001 9
2	3.430 8	3.427 8	0.003 0
3	2.895 4	2.880 1	0.015 3
4	2.381 4	2.343 4	0.038 0
5	2.029 5	1.960 9	0.068 6
6	1.834 3	1.716 9	0.117 4
7	1.714 5	1.514 4	0.200 1
8	1.614 5	1.307 1	0.307 4
9	1.514 3	1.091 1	0.423 2
10	1.406 2	0.875 7	0.530 5
11	1.285 8	0.679 5	0.606 3
12	1.154 4	0.511 3	0.643 1
13	1.015 3	0.373 6	0.641 7
14	0.873 8	0.267 3	0.606 5
15	0.736 2	0.187 6	0.548 6
16	0.608 1	0.130 0	0.478 1
17	0.492 9	0.090 0	0.402 9
18	0.393 1	0.062 4	0.330 7
19	0.309 4	0.043 6	0.265 7
20	0.240 7	0.031 0	0.209 7
21	0.185 6	0.022 3	0.163 3
22	0.142 3	0.016 2	0.126 1
23	0.108 5	0.011 9	0.096 6
24	0.082 4	0.008 9	0.073 5
25	0.062 5	0.006 7	0.055 8
26	0.047 5	0.005 1	0.042 3
27	0.036 0	0.004 0	0.032 0
28	0.027 5	0.003 2	0.024 3
29	0.021 0	0.002 5	0.018 5
30	0.016 2	0.002 0	0.014 2

Figure A.7: Numerical data for Figure 6.1.