Empirical Evidence for Hilberg's Conjecture in Single-Author Texts

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Institute of Computer Science
Polish Academy of Sciences
Warsaw, Poland

QUALICO 2012

1. What is Hilberg's conjecture?
2. Why is Hilberg's conjecture important?
3. Empirical evidence
4. Conclusion

Shannon's experiments

C. Shannon investigated predictability of a text in English.
As a measure of predictability he chose entropy.

Conditional entropy of letter $\mathbf{X_{n+1}}$ given $\mathbf{X_1^n := (X_1, X_2, ..., X_n)}$:

$$\mathbf{H(X_n|X_1^n) = -\sum_{x_1^{n+1}} P(x_1^{n+1}) \log P(x_{n+1}|x_1^n).}$$

1. The entropy was estimated using human subjects.
2. The obtained data points were $\mathbf{n = 0, 1, 2, ..., 15, 100}$.

Claude Shannon, (1951). *Prediction and entropy of printed English.* Bell System Technical Journal, 30:50–64

Hilberg's conjecture

W. Hilberg replotted Shannon's data in a log-log scale and observed a straightish line, i.e., a power-law relationship,

$$\mathbf{H(X_n|X_1^n) \propto n^{-1+\beta}}, \qquad \beta \approx \mathbf{0.5}.$$

Hilberg supposed that this relationship may be extrapolated for $\mathbf{n \gg 100}$.

Wolfgang Hilberg, (1990). *Der bekannte Grenzwert der redundanzfreien Information in Texten — eine Fehlinterpretation der Shannonschen Experimente?* Frequenz, 44:243–248.

Equivalent formulations

Entropy of a block of consecutive $\mathbf{n}$ letters:

$$\mathbf{H(n) = H(X_1^n) = -\sum_{x_1^n} P(x_1^n) \log P(x_1^n)}$$

$$\mathbf{= \sum_{m=1}^{n} H(X_m|X_1^{m-1}) \propto \int_0^n m^{-1+\beta} dm.}$$

Hence, the entropy of an $\mathbf{n}$-letter long text is:

$$\mathbf{H(n) \propto n^{\beta}, \qquad \beta \approx 0.5.}$$

The entropy rate of an $\mathbf{n}$-letter long text:

$$\mathbf{H(n)/n \propto n^{-1+\beta}.}$$

A more plausible formulation

Mutual information between two adjacent blocks of $\mathbf{n}$ letters:

$$\begin{aligned}\mathbf{E(n)} = \mathbf{I(X_1^n; X_{n+1}^{2n})} &= \mathbf{H(X_1^n)} + \mathbf{H(X_{n+1}^{2n})} - \mathbf{H(X_1^{2n})} \\ &= \mathbf{H(n)} + \mathbf{H(n)} - \mathbf{H(2n)} \quad \text{(by stationarity)} \\ &\propto \mathbf{2n^\beta} + \mathbf{(2n)^\beta} = \mathbf{(2 - 2^\beta)n^\beta}\end{aligned}$$

Hence the mutual information between the blocks is

$$\mathbf{E(n)} \propto \mathbf{n^\beta}, \qquad \beta \approx \mathbf{0.5}.$$

We would obtain the same for the entropy rate

$$\mathbf{H(n)/n} = \mathbf{Cn^{-1+\beta}} + \mathbf{h}, \qquad \mathbf{h > 0}.$$

In this formulation the entropy rate need not tend to $\mathbf{0}$.
(The latter implies asymptotic determinism of utterances.)
(Hilberg did not pay attention to this problem.)

Two classes of explanations of Zipf's and Herdan's laws

Mandelbrot (1954), Miller (1957):

texts are generated by independent sampling of single characters $\Longrightarrow$ the frequencies of space-to-space chunks in the text are distributed according to a power-law

Dębowski (2006):

texts repetitively convey certain information $\Longrightarrow$ the number of distinct set phrases (significantly often repeated chunks) in the text is not less than the amount of repeated information

Ł. Dębowski, (2006). *On Hilberg's Law and Its Links with Guiraud's Law.* Journal of Quantitative Linguistics, 13:81–109.

Two very different causes lead to two similar effects.

Further developments in Dębowski's explanation

Theorem 1 (an informal expression)

If an **n**-letter long text describes $\mathbf{n}^{\beta}$ independent facts in a repetitive fashion then mutual information $\mathbf{E(n)}$ exceeds $\mathbf{n}^{\beta}$.

Theorem 2 (an informal expression)

If mutual information $\mathbf{E(n)}$ exceeds $\mathbf{n}^{\beta}$ then the text contains at least $\mathbf{n}^{\beta} / \mathbf{\log n}$ different set phrases.

Ł. Dębowski, (2011). *On the Vocabulary of Grammar-Based Codes and the Logical Consistency of Texts.* IEEE Transactions on Information Theory, 57:4589–4599. (presented also on QUALICO 2009)

How to estimate entropy?

There are two basic methods.

1. **Universal compression**:
   - can be performed by a computer (cheap!);
   - incurs a large systematic error because the computer has to learn the probability distribution from the data.
2. **Guessing or gambling**:
   - requires human subjects (costful!);
   - yields much smaller estimates of the entropy. (Human subjects know the language much better than a machine but they also learn from texts.)

Testing Hilberg's hypothesis requires multiple estimation of conditional entropy for growing contexts. This is extremely exhausting for human subjects. Therefore using universal compression remains the only available method.

The Lempel-Ziv (LZ) code

- For the experiments we will use the Lempel-Ziv (LZ) code.
- LZ code is an instance of a universal code.
- For stationary processes, the compression rate of universal codes asymptotically equals the entropy rate, i.e,

$$\lim_{n\to\infty} \frac{H_{LZ}(n)}{n} = \lim_{n\to\infty} \frac{H(n)}{n}. \tag{1}$$

- If (1) holds then the estimates of mutual information yielded by the code exceed the correct values, i.e.,

$$E_{LZ}(n) \geq E(n) \tag{2}$$

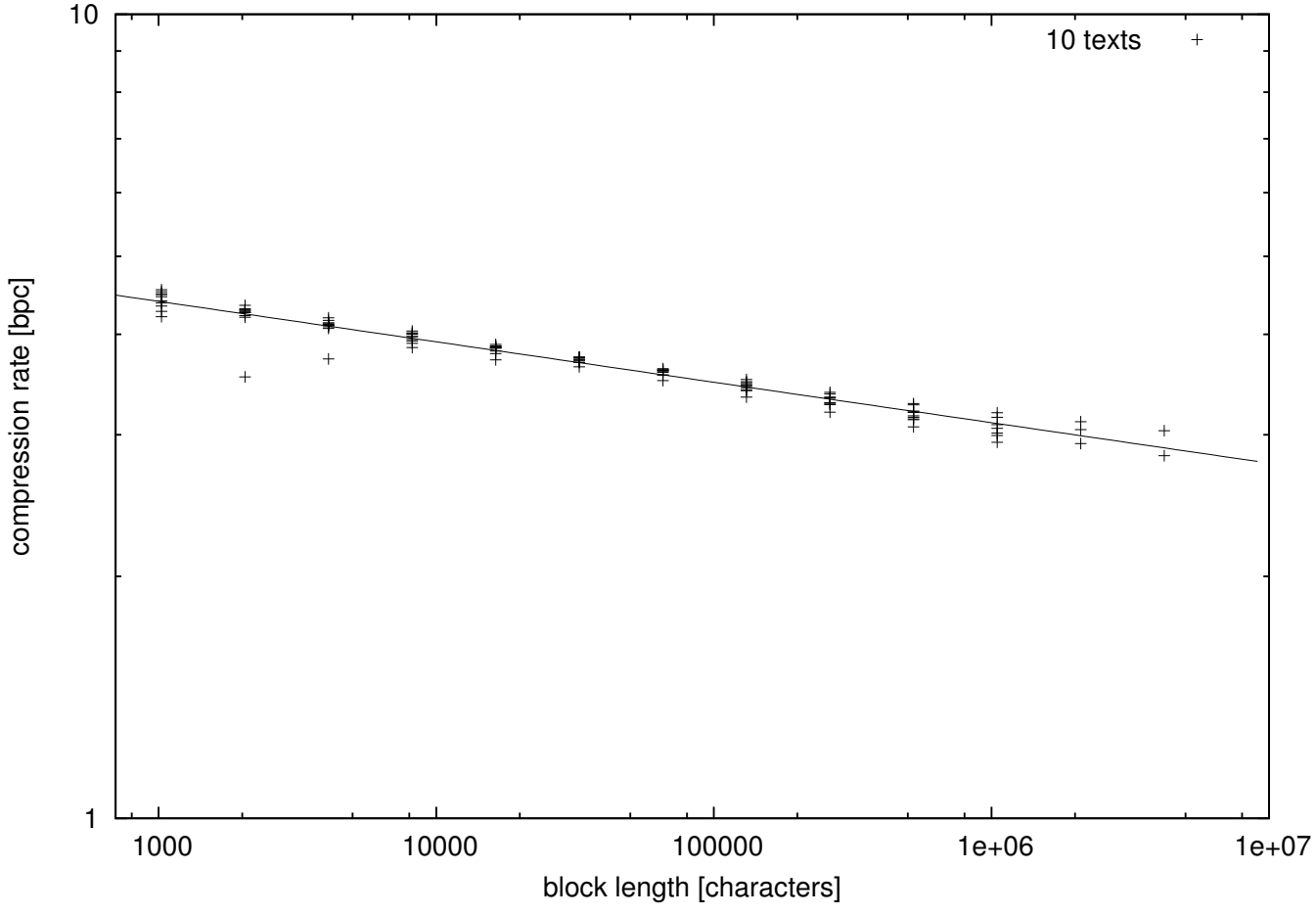
for infinitely many n, where $E_{LZ}(n) = 2H_{LZ}(n) - H_{LZ}(2n)$.

Texts compressed with the LZ code

Title	Author
First Folio/35 Plays	W. Shakespeare
Critical & Historical Essays	T. B. Macaulay
The Complete Memoirs	J. Casanova
Memoirs of Comtesse du Barry	E. Lamothe-Langon
The Descent of Man	C. Darwin
Gulliver's Travels	J. Swift
The Mysterious Island	J. Verne
Mark Twain, a Biography	A. B. Paine
The Journal to Stella	J. Swift
Life of William Carey	G. Smith

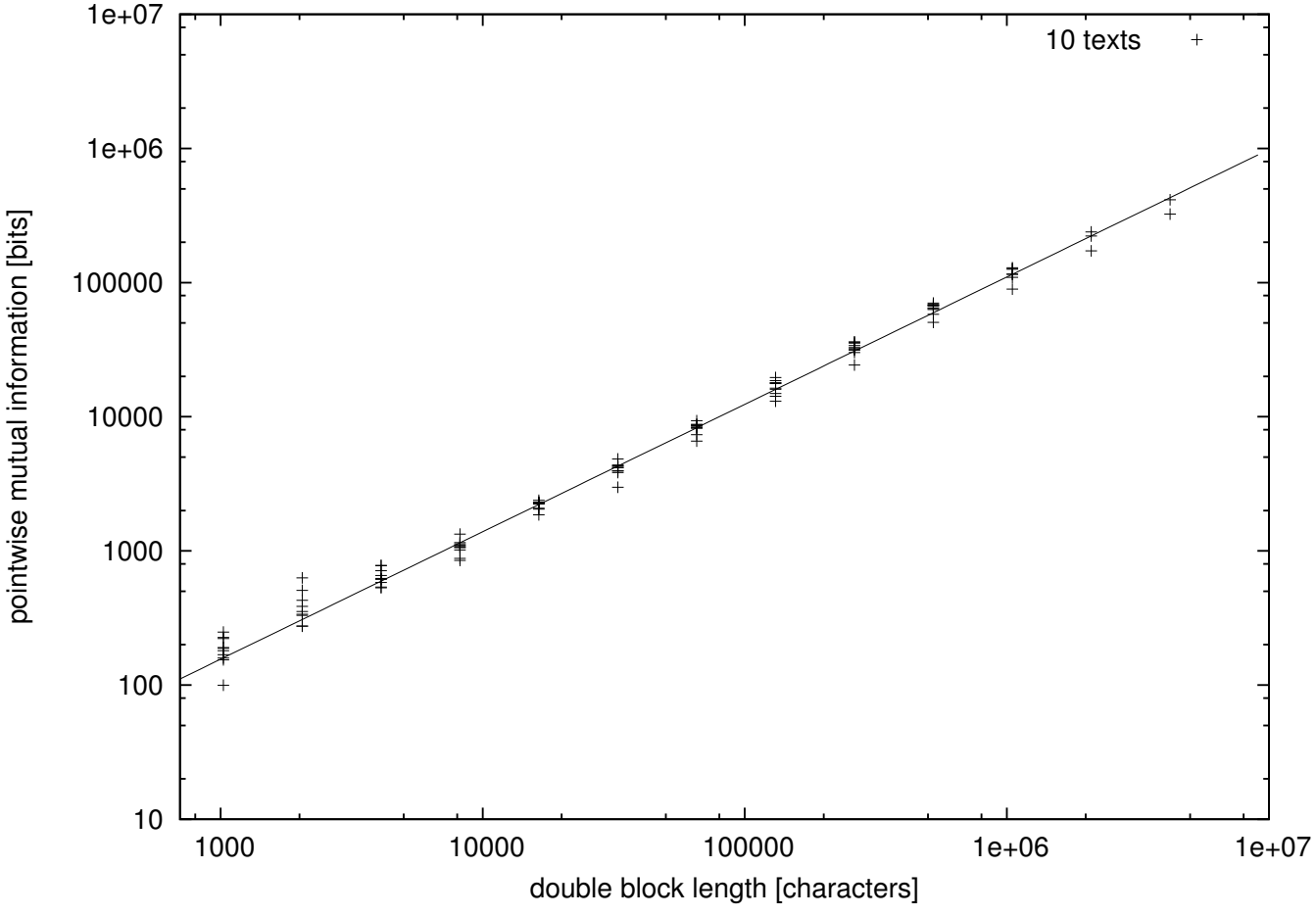
Downloaded from the Project Gutenberg.

Compression rate for LZ



$$\mathbf{H_{LZ}(n)/n \approx 6.22 n^{-1+0.949}}\ \text{[bpc]}$$

Mutual information between blocks for LZ



$$\mathbf{E_{LZ}(n) \propto n^{0.949}}$$

A crude model of human language competence

When human subjects are asked to gamble on the next letter, they use a prior knowledge of letter statistics for their language.

- We might model this phenomenon considering rather the conditional length of the LZ code

$$\mathbf{H}_{\mathrm{LZ}}(\mathbf{n}|\mathrm{Gulliver}) = \mathbf{H}_{\mathrm{LZ}}(\mathrm{Gulliver}, \mathbf{n}) - \mathbf{H}_{\mathrm{LZ}}(\mathrm{Gulliver}).$$

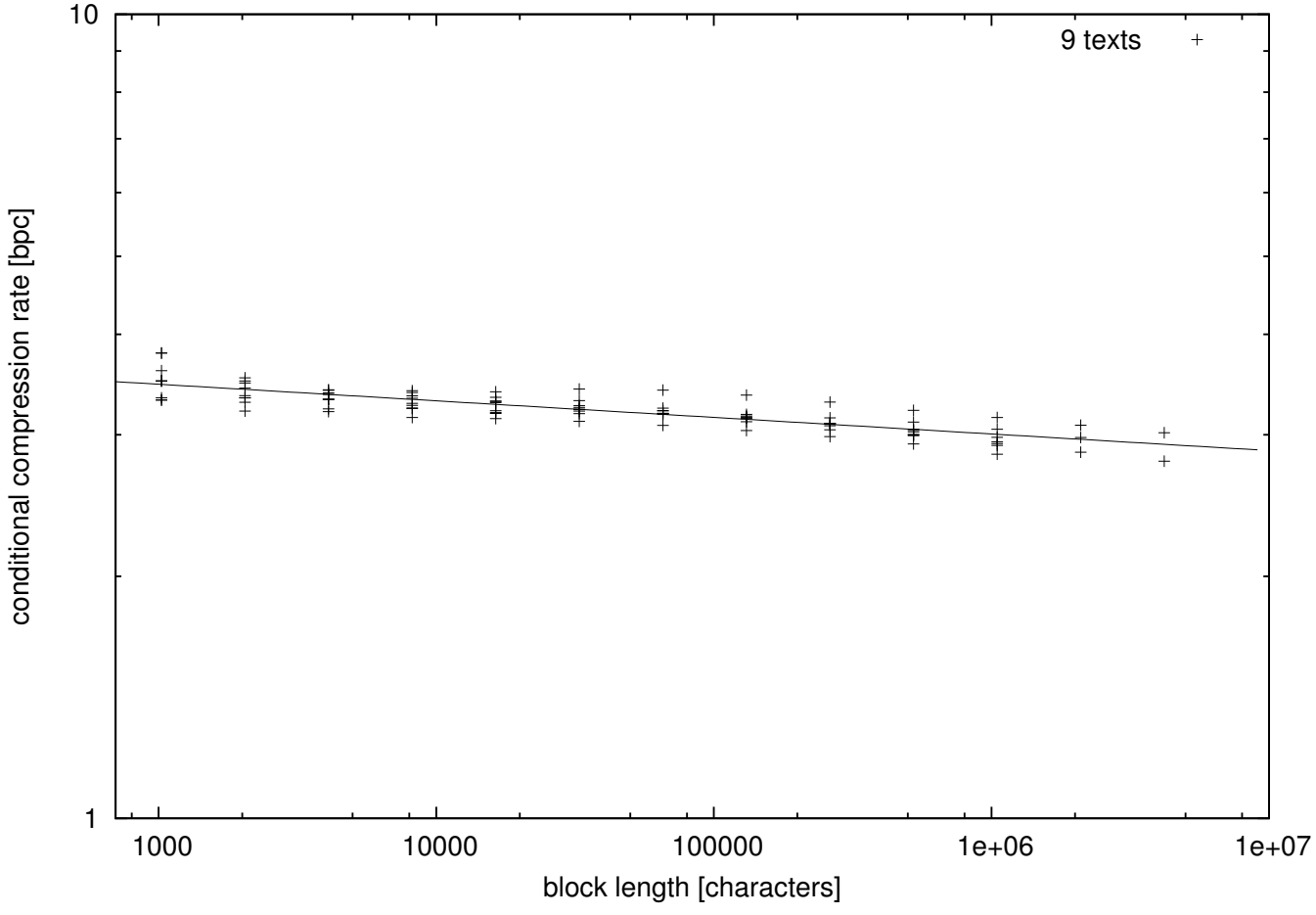
- Usually, we have

$$\mathbf{H}_{\mathrm{LZ}}(\mathbf{n}|\mathrm{Gulliver}) \leq \mathbf{H}_{\mathrm{LZ}}(\mathbf{n}).$$

- The conditional mutual information is

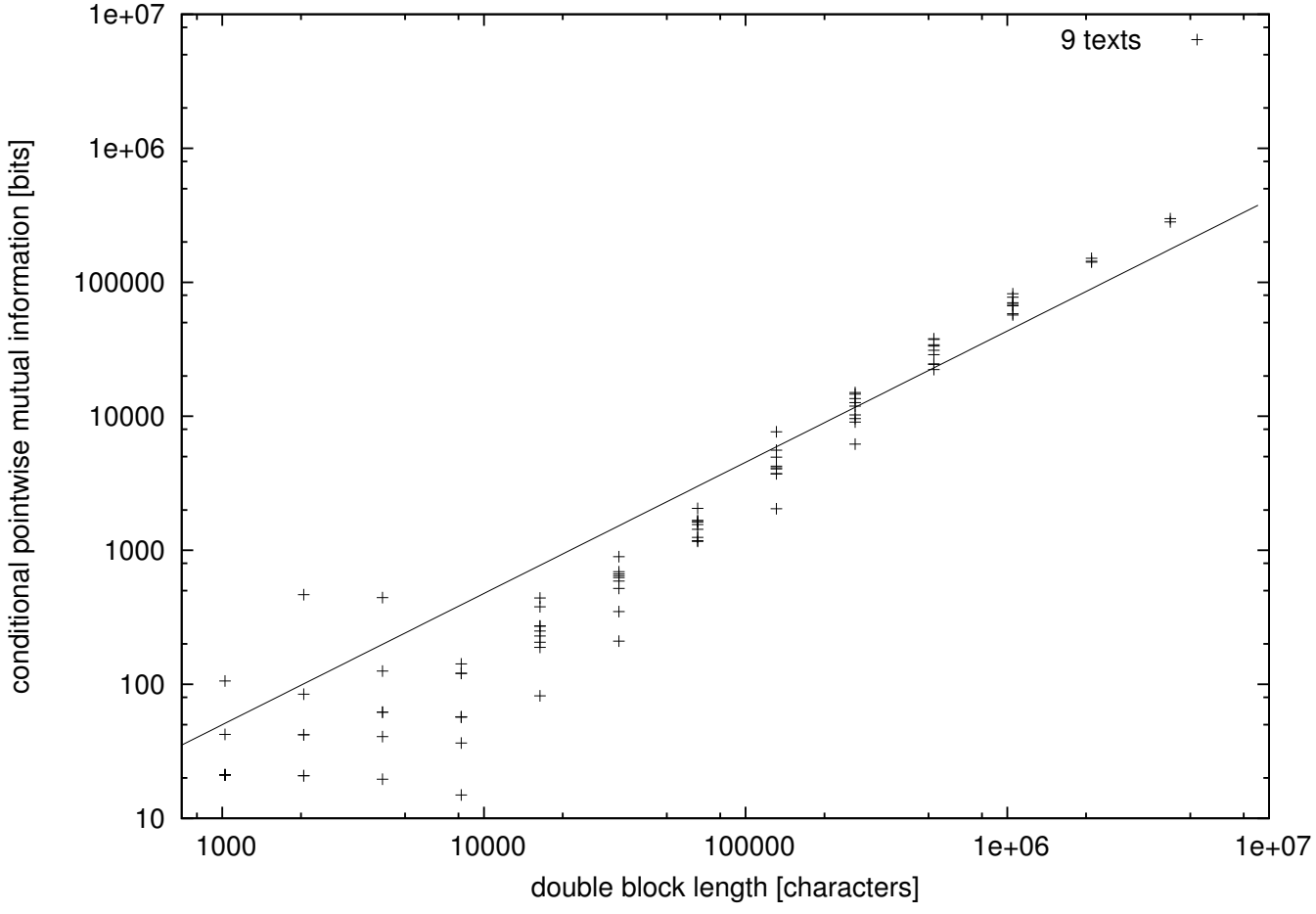
$$\mathbf{E}_{\mathrm{LZ}}(\mathbf{n}|\mathrm{Gulliver}) = 2\mathbf{H}_{\mathrm{LZ}}(\mathbf{n}|\mathrm{Gulliver}) - \mathbf{H}_{\mathrm{LZ}}(2\mathbf{n}|\mathrm{Gulliver}).$$

Conditional compression rate for LZ



$$\mathbf{H_{LZ}(n|\text{Gulliver})/n \approx 3.99 n^{-1+0.979}}\ \text{[bpc]}$$

Conditional information between blocks for LZ



$$\mathbf{E}_{\mathrm{LZ}}(\mathbf{n}|\mathrm{Gulliver}) \propto \mathbf{n}^{\mathbf{0.979}}$$

Conclusion

1. We have confirmed a relaxed Hilberg's conjecture

$$\mathbf{E_{LZ}(n) \propto n^{0.949}} \quad \text{for} \quad \mathbf{n \in (1000, 10^7).} \tag{3}$$

2. The graph of conditional mutual information is more irregular.
3. We might conclude that

$$\mathbf{E(n) \leq Cn^{0.949}} \tag{4}$$

   if (3) held for $\mathbf{n \to \infty}$ and the compression rate of the LZ code converged asymptotically to the entropy rate.
4. The observed compression rate for the LZ code is much higher than the estimates of the entropy rate obtained by gambling (**3** bpc vs. **1.3** bpc). For this reason, it is too risky to infer (4).

One can repeat the experiment for other codes and longer texts.

Dębowski's theorems ...

The formal model of set phrases

We will identify set phrases in the text as nonterminal symbols of the shortest grammar-based compression of the text.

$$\left\{\begin{array}{l} \mathbf{A_1} \mapsto \mathbf{A_2}\mathbf{A_2}\mathbf{A_4}\mathbf{A_5}\texttt{dear\_children}\mathbf{A_5}\mathbf{A_3}\texttt{all.} \\ \mathbf{A_2} \mapsto \mathbf{A_3}\texttt{you}\mathbf{A_5} \\ \mathbf{A_3} \mapsto \mathbf{A_4}\texttt{\_to\_} \\ \mathbf{A_4} \mapsto \texttt{Good\_morning} \\ \mathbf{A_5} \mapsto \texttt{,\_} \end{array}\right\}$$

Good morning to you,
Good morning to you,
Good morning, dear children,
Good morning to all.

For longer texts, $\mathbf{A_i}$ often match the word boundaries, especially if $\mathbf{A_i}$ are defined using only terminal symbols for $\mathbf{i > 1}$.

(Wolff 1980, de Marcken 1996, Kit and Wilks 1999)

The considered probabilistic model

We assume that both a corpus of texts and a state of affairs, repetitively described in the corpus, are random variables.

facts Let $\mathbf{Z_k}$, $\mathbf{k = 1, 2, 3, ...}$, be the logical values (true or false), with respect to the random state of affairs, of certain systematically enumerated logically independent propositions.

Variables $\mathbf{Z_k}$ are equidistributed and probabilistically independent.

texts Let $\mathbf{X_i}$, $\mathbf{i = 1, 2, 3, ...}$, be the consecutive letters of the corpus. We assume that each $\mathbf{Z_k}$ can be inferred from the corpus if we start reading from an arbitrary position.

Variables $\mathbf{X_i}$ take a finite number of values and form a stationary finite-energy process and there exists such functions $\mathbf{s_k}$ that $\lim_{n\to\infty} \mathbf{P}\left(s_k\left(X_{i+1}, X_{i+2}, ..., X_{i+n}\right) = Z_k\right) = 1$ for $\mathbf{i = 1, 2, 3, ...}$.

Two quantities and the claim

For the process as before, put $\mathbf{X_1^n := (X_1, X_2, ..., X_n)}$.

- Let $\mathbf{U(n) := \left\{k : P\left(s_k\left(X_1^n\right) = Z_k\right) \geq \delta\right\}}$, $\mathbf{\delta \in (0.5, 1)}$, be the set of sufficiently well predictable facts.
- Let $\mathbf{V(n)}$ be the number of distinct nonterminal symbols in the shortest grammar-based compression of string $\mathbf{X_1^n}$.

Theorems 1 & 2

For $\beta \in (0, 1)$ and $p > 1$,

$$\liminf_{n\to\infty} \frac{|U(n)|}{n^\beta} > 0$$

$$\Downarrow$$

$$\liminf_{n\to\infty} \frac{E(n)}{n^\beta} > 0 \quad \Longrightarrow \quad \limsup_{n\to\infty} \mathbb{E}\left(\frac{V(n)}{n^\beta (\log n)^{-1}}\right)^p > 0.$$