The Puzzling Entropy Rate of Human Languages

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Institute of Computer Science
Polish Academy of Sciences

QUALICO 2018
Wrocław, 05–08.07.2018

Entropy rate of human languages

In this talk, we will look back into research
in the entropy rate of human languages.

- Entropy rate, introduced by Shannon (1948) is the limiting amount of information = unpredictability = randomness per single symbol of a stationary stochastic process.
- Shannon's information theory was motivated by an idea that the stream of utterances made by humans constitutes a typical realization of a stationary ergodic stochastic process.
- Under this assumption, there exists a unique entropy rate of human utterances and it can be effectively estimated.

Both stationarity and ergodicity of human utterances are some idealization, which impacts our estimates of the entropy rate.

1 Introduction

2 Definitions and ideas

3 Psychological methods

4 Computational methods

5 Around Hilberg's hypothesis

6 Conclusion

What is the Shannon entropy?

Shannon (1948):

- information = unpredictability = shortest encoding
- Shannon entropy:

$$\begin{aligned} H(X) &:= -\sum_x P(X=x)\log P(X=x) \\ &\approx \min_{C\in\mathcal{C}} \sum_x P(X=x)\,|C(x)| \end{aligned}$$

- Entropy is the amount of information in a random variable.
- Shannon conditional entropy:

$$H(X|Y) := -\sum_{x,y} P(X=x, Y=y)\log P(X=x|Y=y)$$

- We have $H(X) \geq H(X|Y) \geq H(X|Y,Z) \geq 0$.

What is the entropy rate?

Shannon (1948):

- Stationary stochastic process $\boldsymbol{(X_i)_{i=1}^{\infty} = (X_1, X_2, X_3, \ldots)}$.
- Blocks of random variables $\boldsymbol{X_j^k = (X_j, X_{j+1}, \ldots, X_k)}$.
- Shannon entropy rate:

$$h := \lim_{n\to\infty} \frac{H(X_1^n)}{n} = \lim_{n\to\infty} H(X_n|X_1^{n-1})$$

- Entropy rate is the limiting amount of information per symbol of a stationary process.
- Entropy rate can be **estimated** for an unknown process.

If the process is stationary and ergodic, then different samples of the process yield the same estimate of the entropy rate.

Stationary and ergodic processes

- Process $(X_i)_{i=1}^{\infty} = (X_1, X_2, X_3, \dots)$ is called stationary if probabilities $P(X_{i+1}^{i+k} = x_1^k)$ do not depend on i for all x_1^k.
- Frequency of a string $N(x_1^k|X_1^n) := \sum_{i=0}^{n-k} 1\left\{X_{i+1}^{i+k} = x_1^k\right\}$.

Birkhoff ergodic theorem

For any stationary process $(X_i)_{i=1}^{\infty}$ almost surely there exist limits:

$$\Phi(x_1^k|X_1^{\infty}) := \lim_{n\to\infty} \frac{N(x_1^k|X_1^n)}{n-k+1}$$

- Process $(X_i)_{i=1}^{\infty} = (X_1, X_2, X_3, \dots)$ is called ergodic if relative frequency $\Phi(x_1^k|X_1^{\infty})$ doesn't depend on X_1^{∞} for all x_1^k.

Some simple example

- Ergodic processes:

  A series of zeros:

  00000000000000000000000000000000000000000...

  entropy rate: $\boldsymbol{h = 0}$

- Non-ergodic processes:

  A series of zeros or a series of ones (first, we flip a coin):

  with probability **1/2**: **00000000000000000000000000...**
  with probability **1/2**: **11111111111111111111111111...**

  entropy rate: $\boldsymbol{h = 0}$

A more advanced example

- Ergodic processes:

  Bernoulli(θ) process: $x_i \in \{0, 1\}$, $\theta \in [0, 1]$,

$$P(X_1^n = x_1^n) := \prod_{i=1}^{n} p(x_i|\theta), \quad p(x_i|\theta) = \begin{cases} 1 - \theta, & x_i = 0, \\ \theta, & x_i = 1. \end{cases}$$

$$h = h(\theta) := -\theta \log \theta - (1 - \theta) \log(1 - \theta)$$

- Non-ergodic processes:

  Mixture Bernoulli process: $x_i \in \{0, 1\}$, $\theta_j \in [0, 1]$,

$$P(X_1^n = x_1^n) := \sum_{j=1}^{K} q_j \prod_{i=1}^{n} p(x_i|\theta_j), \quad q_j > 0, \quad \sum_{j=1}^{K} q_j = 1.$$

$h = \sum_{j=1}^{K} q_j h(\theta_j)$ but observed is $h(\theta_j)$ with probability q_j!

Is natural language ergodic or not?

A process is ergodic when frequencies of strings of any length in sufficiently long samples converge to constants.

1. Suppose now, we choose at random a text in natural language.
2. Imagine counting the frequencies of a keyword, such as **lemma** for a math textbook and **love** for a romance. (example due to Yaglom and Yaglom, 1983)
3. We expect that the frequencies of keywords are random variables with values depending on the random text topic.
4. Since keywords are some strings, the stochastic process that models natural language should be not ergodic = non-ergodic.

This happens for finite texts though, whereas the mathematical definition of ergodic processes considers text lengths $\rightarrow \infty$.

Entropy rate(s) of human language(s)?

- Shannon's information theory was motivated by an idea that the stream of utterances made by humans constitutes a typical realization of a stationary ergodic stochastic process.
- But: We have a strong intuition that there is a huge variation of frequencies of keywords in natural texts.
- But: The actual estimates of the entropy rate for various texts do not depend so much on the text or a particular language.

Natural texts may be quite uniform with respect
to information measures in general, not only the entropy rate.
Information measures can be language universals.

1. Introduction
2. Definitions and ideas
3. Psychological methods
4. Computational methods
5. Around Hilberg's hypothesis
6. Conclusion

Shannon's method (1951)

Shannon's method (1951) of estimating entropy of language:

- Human subjects are asked to guess the next character of a text given n previous characters.
- Let q_{in} be the probability of guessing the character in i attempts given n characters in the optimal strategy.

Theorem

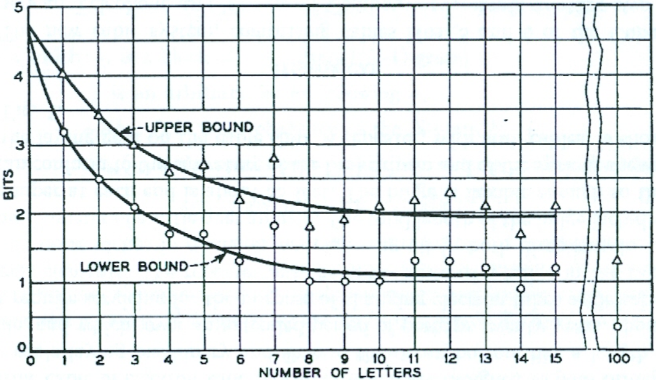
$$\sum_{i\ge 1} \left[q_{in} - q_{i+1,n}\right] i \log i \le H(X_{n+1}|X_1^n) \le -\sum_{i\ge 1} q_{in} \log q_{in}$$

- Let Q_{in} be the number of times that human subjects guessed the character in i attempts given n characters. If human guesses are close to the optimal strategy then we may estimate

$$q_{in} \approx \frac{Q_{in}}{\sum_{k\ge 1} Q_{kn}}.$$

Shannon's plot (1951)

The guessed text was *Jefferson the Virginian* by Dumas Malone.



Conclusion: The entropy rate of English ≈ **1** bit per character.

Cover and King's method (1978)

Cover and King's (1978) estimator of entropy of language:

- Human subjects are asked to bet on the next character of a text given $\boldsymbol{n}$ previous characters.
- Gambling system: $\boldsymbol{b(x_{n+1}|x_1^n) \geq 0}$, $\sum_{x_{n+1}} \boldsymbol{b(x_{n+1}|x_1^n) = 1}$.
- Accumulated capital: $\boldsymbol{S_0 := 1}$, $\boldsymbol{S_{n+1} := Ab(X_{n+1}|X_1^n)S_n}$.

Theorem

$$\liminf_{n\to\infty} \left[\log A - \frac{1}{n}\log S_n\right] \geq h \text{ almost surely}$$

The gambled text was *Jefferson the Virginian* again.

The estimate of the entropy rate was **1.25** bits per character (for the committee gambling).

1. Introduction
2. Definitions and ideas
3. Psychological methods
4. Computational methods
5. Around Hilberg's hypothesis
6. Conclusion

Brown et al.'s method (1992)

Brown et al.'s method (1992) of estimating the entropy rate:

- Construct a statistical language model $\boldsymbol{M}$, i.e., a probability distribution over finite strings of characters.
- Collect a reasonably long test sample of text.
- With a high probability we have

$$\boldsymbol{h} \leq -\frac{\mathbf{1}}{\boldsymbol{n}} \log \boldsymbol{M}(\text{test sample}).$$

For a word-trigram statistical language model they obtained $\boldsymbol{h} \leq \mathbf{1.75}$ bits per character for the Brown Corpus of English.

Lempel and Ziv's method (1977)

Lempel and Ziv's method (1977) of estimating the entropy rate:

- The text is parsed into a sequence of shortest phrases that have not appeared before (except for the last phrase). For example, the sequence 001010010011100... is split into phrases 0, 01, 010, 0100, 1, 11, 00, ...
- Let C_n be the number of phrases in the compressed block X_1^n.
- Let $(X_i)_{i=1}^{\infty}$ be stationary ergodic over a finite alphabet.

Theorem

$$\lim_{n\to\infty} \frac{C_n \log C_n}{n} = h \text{ almost surely}$$

PPM — Ryabko (1984); Cleary and Witten (1984)

- For $k \in \{-1, 0, 1, \ldots\}$, we put

$$\mathrm{PPM}_k(x_i|x_1^{i-1}) := \begin{cases} D^{-1} & k = -1 \\ \dfrac{N(x_{i-k}^{i}|x_1^{i-1}) + 1}{N(x_{i-k}^{i-1}|x_1^{i-2}) + D} & k \geq 0 \end{cases}$$

$$\mathrm{PPM}_k(x_1^n) := \prod_{i=1}^{n} \mathrm{PPM}_k(x_i|x_1^{i-1})$$

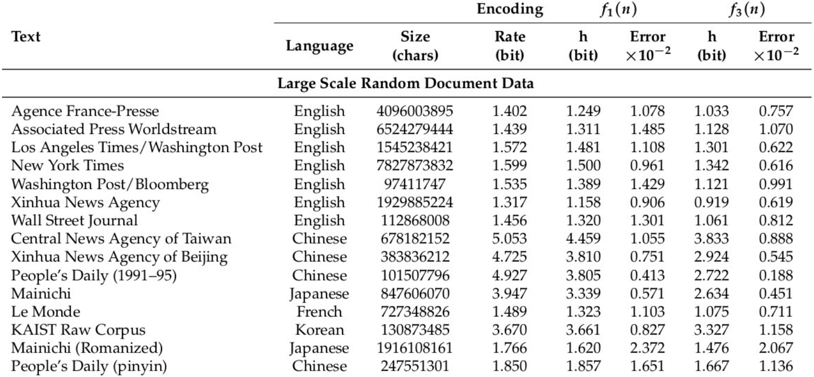
$$\mathrm{PPM}(x_1^n) := \frac{6}{\pi^2} \sum_{k=-1}^{\infty} \frac{\mathrm{PPM}_k(x_1^n)}{(k+2)^2}$$

- Let $(X_i)_{i=1}^{\infty}$ be stationary ergodic over a finite alphabet.

Theorem

$$\lim_{n\to\infty} \frac{1}{n} \left[-\log \mathrm{PPM}(X_1^n) \right] = h \textit{ almost surely}$$

Takahira et al.'s experiment (2016)



Text	Language	Size (chars)	Encoding Rate (bit)	$f_1(n)$ h (bit)	$f_1(n)$ Error $\times 10^{-2}$	$f_3(n)$ h (bit)	$f_3(n)$ Error $\times 10^{-2}$
Large Scale Random Document Data							
Agence France-Presse	English	4096003895	1.402	1.249	1.078	1.033	0.757
Associated Press Worldstream	English	6524279444	1.439	1.311	1.485	1.128	1.070
Los Angeles Times/Washington Post	English	1545238421	1.572	1.481	1.108	1.301	0.622
New York Times	English	7827873832	1.599	1.500	0.961	1.342	0.616
Washington Post/Bloomberg	English	97411747	1.535	1.389	1.429	1.121	0.991
Xinhua News Agency	English	1929885224	1.317	1.158	0.906	0.919	0.619
Wall Street Journal	English	112868008	1.456	1.320	1.301	1.061	0.812
Central News Agency of Taiwan	Chinese	678182152	5.053	4.459	1.055	3.833	0.888
Xinhua News Agency of Beijing	Chinese	383836212	4.725	3.810	0.751	2.924	0.545
People's Daily (1991–95)	Chinese	101507796	4.927	3.805	0.413	2.722	0.188
Mainichi	Japanese	847606070	3.947	3.339	0.571	2.634	0.451
Le Monde	French	727348826	1.489	1.323	1.103	1.075	0.711
KAIST Raw Corpus	Korean	130873485	3.670	3.661	0.827	3.327	1.158
Mainichi (Romanized)	Japanese	1916108161	1.766	1.620	2.372	1.476	2.067
People's Daily (pinyin)	Chinese	247551301	1.850	1.857	1.651	1.667	1.136

13 press corpora up to 8 GB; 5 languages; PPM algorithm

Gao et al.'s method (2008)

Gao et al.'s method (2008) of estimating the entropy rate:

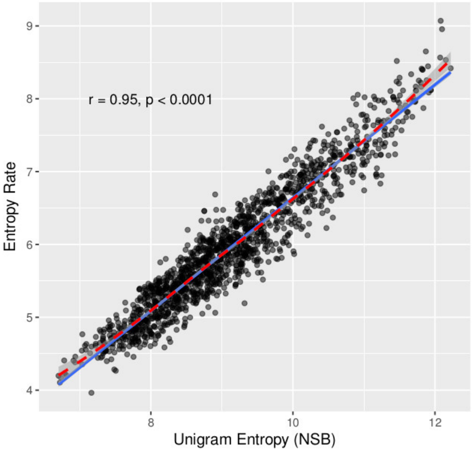
- Let $(X_i)_{i=1}^{\infty}$ be a stationary ergodic process. (Can be over a countably infinite alphabet.)
- The longest match length:

$$L_n := \max\left\{0 \le l \le n : X_{n+1}^{n+l} = X_{j+1}^{j+l} \text{ for some } 1 \le j \le n\right\}$$

Theorem

$$\lim_{n\to\infty} \frac{1}{n-1} \sum_{i=2}^{n} \frac{\log i}{1+L_i} = h \textit{ almost surely}$$

Bentz et al.'s experiment (2017) — entropy of words



3 parallel corpora; 450 million words; 1259 languages;
Gao et al. estimator for entropy rate $\boldsymbol{h}$;
plug-in estimator for unigram entropy $\boldsymbol{H(X_1)}$

Plug-in estimator for IID processes

- Let $(X_i)_{i=1}^{\infty}$ be a sequence of independent identically distributed random variables.
- Denote the discrete distribution $p(x) = P(X_i = x)$.
- Entropy of a discrete distribution:

$$H(p) := -\sum_{x} p(x) \log p(x)$$

- Empirical distributions:

$$\hat{p}_n(x) := \frac{1}{n} \sum_{i=1}^{n} 1\left\{X_i = x\right\}$$

- We have $\mathbb{E}\, \hat{p}_n(x) = p(x)$ and hence:

$$\mathbb{E}\, H(\hat{p}_n) \le \min\left\{\log n, H(p)\right\}$$

Dębowski's method (2016)

Dębowski's method (2016) of estimating the entropy rate:

- Let $(X_i)_{i=1}^{\infty}$ be stationary ergodic over a finite alphabet.
- Empirical distributions of blocks of length k:

$$\hat{p}_{k,n}(x_1^k) := \frac{1}{n} \sum_{i=1}^{n} 1 \left\{ X_{(i-1)k+1}^{ik} = x_1^k \right\}.$$

- We have $\mathbb{E}\, H(\hat{p}_{k,n}) \leq \min \left\{ \log n, H(X_1^k) \right\}$.

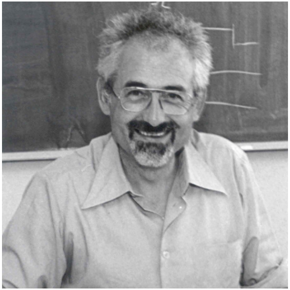
Theorem

For any $\epsilon > 0$ and $n(k) \geq 2^{k(h+\epsilon)}$, we have:

$$\liminf_{k \to \infty} \frac{1}{k} H(\hat{p}_{k,n(k)}) = h \text{ almost surely}$$

1. Introduction
2. Definitions and ideas
3. Psychological methods
4. Computational methods
5. Around Hilberg's hypothesis
6. Conclusion

Wolfgang Hilberg (1932–2015)

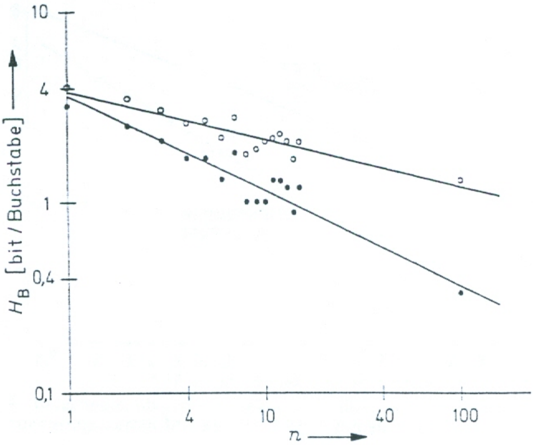


He was a German electrical engineer.

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

Hilberg's plot (1990)

Shannon's plot (1951) redrawn in the doubly logarithmic scale:



Hilberg's conclusion (1990): The entropy rate of English is **0**.

Hilberg's hypothesis (1990) and its relaxation

- Hilberg's power law (1990):

$$H(X_1^k) \approx Bk^{\beta}, \quad \beta \approx 1/2$$

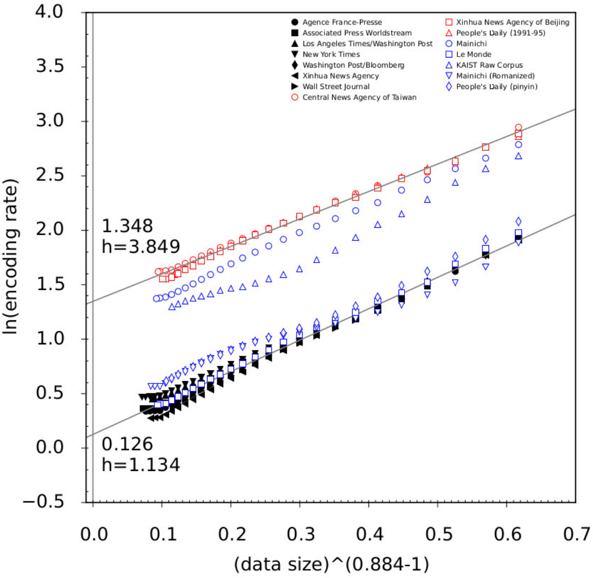
This implies an asymptotic determinism of natural texts.
(a dubious condition)

- Ebeling and Nicolis (1991), Crutchfield and Feldman (2003):

$$H(X_1^k) \approx Bk^{\beta} + hk, \quad \beta > 0$$

This implies that language is not a hidden Markov process of a finite order and the mutual information between the past and the future is unbounded. **(sounds reasonable)**

Takahira et al.'s experiment (2016): $\beta \leq \mathbf{0.884}$



Dębowski's experiment (2015)

- The (length of the) maximal repetition:

$$\boldsymbol{L(x_1^n) := \max\left\{k : N(w_1^k|x_1^n) \geq 2 \text{ for some } w_1^k\right\}}$$

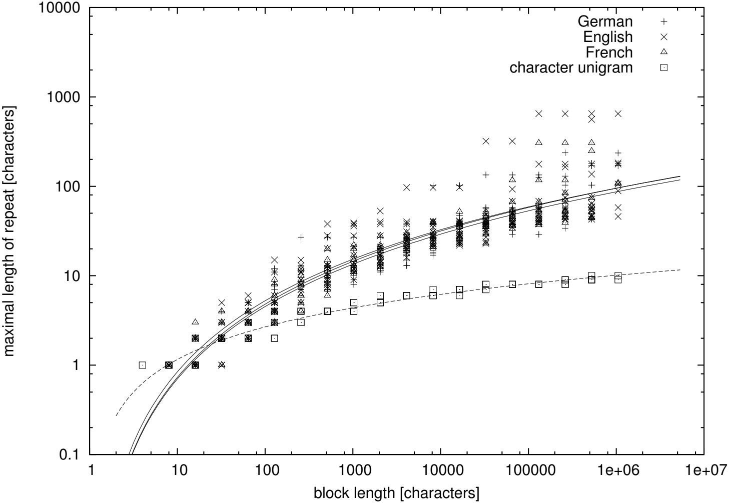
- It can be empirically observed that:

$$\boldsymbol{L(x_1^n) \approx A(\log n)^{\alpha}}$$

For typical IID and hidden Markov processes: $\boldsymbol{\alpha = 1}$.

For texts in German, English, and French: $\boldsymbol{\alpha \approx 3}$.

Dębowski's plot (2015)



Dębowski's result (2018)

For a stationary process $(X_i)_{i=-\infty}^{\infty}$ over a finite alphabet, let us write the conditional Rényi entropy:

$$H_2(X_1^k|X_{-m}^0) := -\log \mathbb{E}\, P(X_1^k|X_{-m}^0) \leq \mathbb{E}\left[-\log P(X_1^k|X_{-m}^0)\right] \leq H(X_1^k)$$

Theorem

$$\limsup_{n\to\infty} \frac{L(X_1^n)}{A(\log n)^\alpha} \geq 1 \implies \liminf_{k\to\infty} \frac{H_2(X_1^k|X_{-D^k}^0)}{Bk^{1/\alpha}} \leq 1$$

Hence for natural language the conditional Rényi entropy rate is:

$$\liminf_{k\to\infty} \frac{H_2(X_1^k|X_{-D^k}^0)}{k} = 0$$

Some open mathematical problems

- First problem:

Constructing mathematical models of stochastic processes with Shannon entropy rate $> \mathbf{0}$ and conditional Rényi entropy rate $= \mathbf{0}$ is an open problem with possible applications to linguistics.

Shields (1992) constructed an example of such a process but it does not have a clear linguistic interpretation.

- Second problem:

Do Shannon entropy rate $> \mathbf{0}$ and cond. Rényi entropy rate $= \mathbf{0}$ imply mutual information between past and future $= \infty$?

If yes, we could confirm the relaxed Hilberg hypothesis.

1 Introduction

2 Definitions and ideas

3 Psychological methods

4 Computational methods

5 Around Hilberg's hypothesis

6 Conclusion

Conclusion

- Entropy rate is an important parameter of natural language, measuring the limiting amount of unpredictability of a text per single character.
- An interesting open research issue is the rate of convergence of the block entropy to the entropy rate.
- This rate of convergence is another parameter of natural language connected to its being non-hidden Markovian and its long-range dependence.
- Analyzing the respective phenomena on a mathematical level can contribute not only to quantitative linguistics but also to better statistical language models in computational linguistics.

In spite of possible non-ergodicity of natural language, some information measures can be language universals!

References (1/2)

Bentz, C., Alikaniotis, D., Cysouw, M., and Ferrer-i-Cancho, R. (2017). The entropy of words—learnability and expressivity across more than 1000 languages. *Entropy*, 19(6):275.

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

Cover, T. M. and King, R. C. (1978). A convergent gambling estimate of the entropy of English. *IEEE Trans. Inform. Theory*, 24:413–421.

Crutchfield, J. P. and Feldman, D. P. (2003). Regularities unseen, randomness observed: The entropy convergence hierarchy. *Chaos*, 15:25–54.

Dębowski, Ł. (2015). Maximal repetitions in written texts: Finite energy hypothesis vs. strong Hilberg conjecture. *Entropy*, 17:5903–5919.

Dębowski, Ł. (2016). Consistency of the plug-in estimator of the entropy rate for ergodic processes. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 1651–1655.

Dębowski, Ł. (2018). Maximal repetition and zero entropy rate. *IEEE Trans. Inform. Theory*, 64(4):2212–2219.

Ebeling, W. and Nicolis, G. (1991). Entropy of symbolic sequences: the role of correlations. *Europhys. Lett.*, 14:191–196.

References (2/2)

Gao, Y., Kontoyiannis, I., and Bienenstock, E. (2008). Estimating the entropy of binary time series: Methodology, some theory and a simulation study. *Entropy*, 10:71–99.

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

Shannon, C. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.*, 30:379–423,623–656.

Shannon, C. (1951). Prediction and entropy of printed English. *Bell Syst. Tech. J.*, 30:50–64.

Shields, P. C. (1992). Entropy and prefixes. *Ann. Probab.*, 20:403–409.

Takahira, R., Tanaka-Ishii, K., and Dębowski, Ł. (2016). Entropy rate estimates for natural language—a new extrapolation of compressed large-scale corpora. *Entropy*, 18(10):364.

Yaglom, A. M. and Yaglom, I. M. (1983). *Probability and Information*. Theory and Decision Library. Springer.

Ziv, J. and Lempel, A. (1977). A universal algorithm for sequential data compression. *IEEE Trans. Inform. Theory*, 23:337–343.