# Language as a Meaningful Stochastic Process: Theorems about Facts and Words

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Institute of Computer Science
Polish Academy of Sciences

Zoom seminar at UC Irvine, May 6, 2020
(recycled from Szklarska Poręba, March 1–4, 2019)

# References

This talk is based mostly on the following sources:

1. Ł. Dębowski, (2011). *On the Vocabulary of Grammar-Based Codes and the Logical Consistency of Texts.* IEEE Transactions on Information Theory, vol. 57, pp. 4589–4599.
2. Ł. Dębowski, (2018). *Is Natural Language a Perigraphic Process? The Theorem about Facts and Words Revisited.* Entropy, vol. 20(2), pp. 85.
3. Ł. Dębowski, (2020). *Information Theory Meets Power Laws: Stochastic Processes and Language Models.* Wiley. (in press)
4. Ł. Dębowski, (2020). *On a Class of Markov Order Estimators Based on PPM and Other Universal Codes.* `https://arxiv.org/abs/2003.04754`

The aim of our research has been to make a few steps toward a probabilistic theory of meaningful texts, applying stochastic processes and information theory.

# "Animate" and "inanimate" stochastic processes

*If a Martian scientist sitting before his radio in Mars accidentally received from Earth the broadcast of an extensive speech [...], what criteria would he have to determine whether the reception represented the effect of animate process [...]? It seems that [...] the only clue to the animate origin would be this: the arrangement of the occurrences would be neither of rigidly fixed regularity such as frequently found in wave emissions of purely physical origin nor yet a completely random scattering of the same.*

— *George Kingsley Zipf (1965:187)*

# Information theory and meaning

*The concept of information [...] at first seems disappointing and bizarre—disappointing because it has nothing to do with meaning, and bizarre because [...] information and uncertainty find themselves to be partners.*

*[...] information and meaning may prove to be something like a pair of canonically conjugate variables in quantum theory, they being subject to some joint restriction that condemns a person to the sacrifice of the one as he insists on having much of the other.*

— *Warren Weaver (1949)*

# Meanings of meaningfulness

The aim of our research has been to make a few steps toward a probabilistic theory of meaningful texts.

- Meaningfulness of texts can be understood as:
    1. effective description of an external or imagined reality (descriptive meaningfulness);
    2. internal cohesion of the narration or the discourse (cohesive meaningfulness);
    3. effective control of an external reality toward some goal (telic meaningfulness).

  — A kind of Borges's classification of animals.

  — Is meaning of life ill-defined? ($\rightarrow$ Victor Frankl)
- We sought how to model these using stochastic processes.
- Thus, meaningful texts can be either natural or idealized:
    - natural texts $=$ texts created by humans;
    - idealized texts $=$ typical realizations of stochastic processes.

# Theorems about facts and words

**Question**

Is language structure a mathematical consequence of descriptive meaningfulness, i.e., effective reference of texts to some reality?

As for double articulation, YES since we will show this:

**Proposition** *(informally stated)*

The number of distinct words in a finite text is roughly greater than the number of independent facts described by the text.

When stated formally, the above proposition becomes a general result in information theory, which remains valid for random texts generated by any stationary stochastic process.

- words $\Longrightarrow$ grammar-based codes/PPM Markov order
- facts $\Longrightarrow$ algorithmic information theory/ergodic theory

## What are the words? (in our approach)

- In our approach, words are some particular substrings of symbols appearing in the text.
- We need an effective procedure for word segmentation which could approximately work both for natural language and stochastic processes (sequences of random letters).
- Two procedures are admissible:
  - Taking non-overlapping substrings that are repeated.
  - Taking all overlapping substrings of the optimal length.
- Both procedures are connected to universal codes, i.e., effective data compression procedures that approximate the entropy rate.

# A context-free grammar that generates one text

$$\left\{ \begin{array}{lcl} A_1 & \rightarrow & A_2 A_2 A_4 A_5 \textbf{dear\_children} A_5 A_3 \textbf{all.} \\ A_2 & \rightarrow & A_3 \textbf{you} A_5 \\ A_3 & \rightarrow & A_4 \textbf{\_to\_} \\ A_4 & \rightarrow & \textbf{Good\_morning} \\ A_5 & \rightarrow & \textbf{,\_} \end{array} \right\}$$

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

# First approach: Minimal grammar-based codes

**Grammar-based coding:**

- a grammar transform $\boldsymbol{\Gamma} : \mathbb{X}^* \to \mathcal{G}$ which for each string $\boldsymbol{w} \in \mathbb{X}^*$ returns a grammar $\boldsymbol{\Gamma}(\boldsymbol{w})$ that generates this string.
- a grammar encoder $\boldsymbol{B} : \mathcal{G} \to \{0,1\}^*$ encodes the grammar as a binary string.

**Vocabulary of a grammar transform:**

- Let $\boldsymbol{V}(\boldsymbol{G})$ be the set of nonterminals in a grammar $\boldsymbol{G}$.
- For a grammar transform $\boldsymbol{\Gamma}$, let $\boldsymbol{V}_{\boldsymbol{\Gamma}}(\boldsymbol{w}) := \boldsymbol{V}(\boldsymbol{\Gamma}(\boldsymbol{w}))$.

**Minimal grammar transforms:**

- Grammar transform $\boldsymbol{\Gamma}$ is called minimal (w.r.t. $\boldsymbol{B}$ and $\mathcal{G}$) if $|\boldsymbol{B}(\boldsymbol{\Gamma}(\boldsymbol{w}))| \le |\boldsymbol{B}(\boldsymbol{G})|$ for any string $\boldsymbol{w}$ and any grammar $\boldsymbol{G} \in \mathcal{G}$ that generates $\boldsymbol{w}$.

Minimal grammar-based codes are NP-hard to compute.
But their approximations can be used
for rough word segmentation of texts in NLP.

# Second approach: Markov order estimators

- Strings: $x_m^n := (x_m, x_{m+1}, \ldots, x_n)$. Let $\inf \emptyset := \infty$.
- For a stationary measure $P$, the Markov order is

$$M := \inf \left\{ k \geq 0 : P(x_{k+1}^n | x_1^k) = \prod_{i=k+1}^{n} P(x_i | x_{i-k}^{i-1}) \text{ for all } x_1^n \right\}.$$

- Function $\mathbb{M} : \mathbb{X}^* \to \mathbb{N}$ is called a consistent estimator of $M$ if

$$\lim_{n \to \infty} \mathbb{M}(X_1^n) = M \text{ almost surely}$$

for any stationary ergodic probability measure $P$.

# PPM Markov order

Empirical frequency: $N(w_1^k|x_1^n) := \sum_{i=1}^{n-k+1} 1\left\{x_i^{i+k-1} = w_1^k\right\}$.

Empirical entropy and PPM measure:

$$h_k(x_1^n) := \frac{1}{n-k} \sum_{i=k+1}^{n} \log \frac{N(x_{i-k}^{i-1}|x_1^{n-1})}{N(x_{i-k}^{i}|x_1^{n})}, \quad k \geq 0,$$

$$\mathrm{PPM}_k(x_1^n) := D^{-k} \prod_{i=k+1}^{n} \frac{N(x_{i-k}^{i}|x_1^{i-1}) + 1}{N(x_{i-k}^{i-1}|x_1^{i-2}) + D}, \quad k \geq 0,$$

$$\Pi(x_1^n) := \frac{6^2}{\pi^4} \cdot \frac{1}{(n+1)^2} \sum_{k=0}^{\infty} \frac{\mathrm{PPM}_k(x_1^n)}{(k+1)^2}.$$

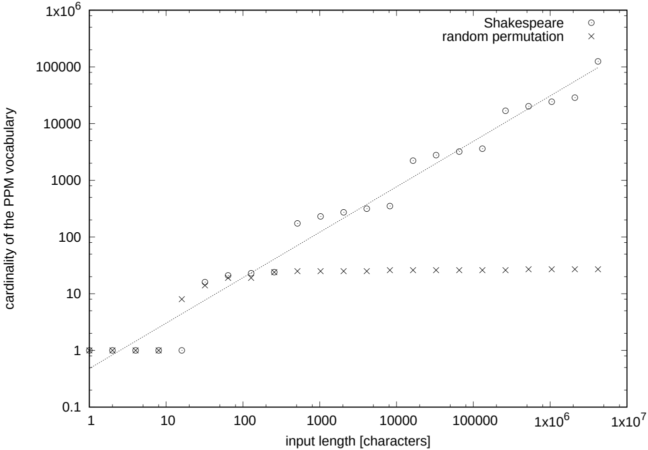
Some consistent estimator of $M$ is the PPM Markov order:

$$\mathbb{M}(x_1^n) := \min\left\{k \geq 0 : (n-k) h_k(x_1^n) \leq -\log \Pi(x_1^n)\right\}.$$

# Herdan-Heaps' power law for PPM vocabulary

Empirical vocabulary: $V_k(x_1^n) := \left\{ x_{t+1}^{t+k} : 0 \le t \le n-k \right\}$.

PPM vocabulary: $V_{\mathbb{M}}(x_1^n) := V_{\mathbb{M}(x_1^n)}(x_1^n)$.



# Basic concepts in information theory

Shannon's approach: $X, Y$ — some random variables.

- Shannon entropy:
  $H(X) := -\sum_{x \in \mathbb{X}} P(X = x) \log P(X = x) \geq 0$.
- Shannon mutual information:
  $I(X; Y) := H(X) + H(Y) - H(X, Y) \geq 0$.

Algorithmic information theory: $x, y$ — some discrete objects.

- Kolmogorov complexity:
  $\mathbb{H}(x) \geq 0$ is the length of the shortest program to generate $x$.
- Algorithmic mutual information:
  $\mathbb{I}(x; y) := \mathbb{H}(x) + \mathbb{H}(y) - \mathbb{H}(x, y) \geq -c$.

Source coding:

- $H(X) \leq \mathbb{E}\, \mathbb{H}(X) \leq H(X) + \mathbb{H}(P) + c$.

# Application to stationary processes

- Let $\mathbb{X}$ be a finite alphabet of symbols.
- Let $(X_i)_{i=1}^{\infty}$ be a sequence of random variables $X_i : \Omega \to \mathbb{X}$.
- We will denote random strings $X_m^n := (X_m, X_{m+1}, ..., X_n)$.
- Process $(X_i)_{i=1}^{\infty}$ is stationary if probabilities $P(X_{t+1}^{t+n} = x_1^n)$ do not depend on positions $t$.
- For a stationary process, there exist two limits, entropy rate $h$ and excess entropy $E$:

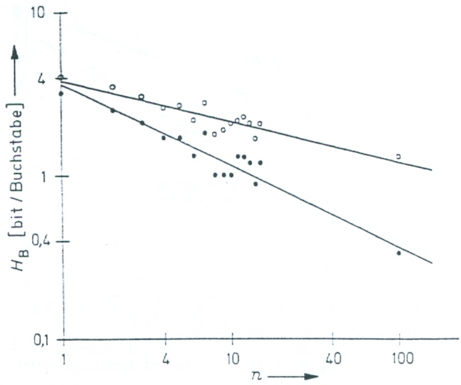
$$h := \lim_{n\to\infty} \frac{H(X_1^n)}{n} = \lim_{n\to\infty} \left[H(X_1^n) - H(X_1^{n-1})\right],$$

$$E := \lim_{n\to\infty} I(X_1^n; X_{n+1}^{2n}) = \lim_{n\to\infty} \left[H(X_1^n) - hn\right].$$

- $h$ is a measure of randomness, $E$ is a measure of structure.
- Finite-state hidden Markov processes satisfy $E < \infty$.

# Hilberg's hypothesis about conditional entropy of language

Hilberg's (1990) plot of Shannon's (1951) data for English:




$$H(X_n|X_1^{n-1}) = H(X_1^n) - H(X_1^{n-1}) \propto n^{-1/2}, \quad n \leq 100$$

# Hilberg's hypothesis for mutual information

We derive:

$$H(X_n|X_1^{n-1}) = H(X_1^n) - H(X_1^{n-1}) \propto n^{-1/2},$$

$$H(X_1^n) = \sum_{k=1}^{n} \left[ H(X_1^n) - H(X_1^{n-1}) \right] \propto n^{1/2},$$

$$I(X_1^n; X_{n+1}^{2n}) = H(X_1^n) + H(X_{n+1}^{2n}) - H(X_1^{2n}) \propto n^{1/2}.$$

The relaxed Hilberg hypothesis:

$$I(X_1^n; X_{n+1}^{2n}) \propto n^{\beta}, \quad \beta \in (0,1).$$

The above does not assume that the entropy rate is $h = 0$.

Natural language may have excess entropy $E = \infty$.
Thus it cannot be a finite-state process (Chomsky vs. Skinner).

# What are the facts?

In our approach, facts are binary digits partly describing some (model of) unchangeable reality that is referred to by texts.

Imagine a long row of chairs randomly painted white or black:

...

The state of this row could be described by a collection of bits $(z_k)_{k=1}^{\infty}$, indexed by indices $k = 1, 2, 3, \ldots$, where

$$z_k := \begin{cases} 0 & \text{if } k\text{-th chair is white,} \\ 1 & \text{if } k\text{-th chair is black.} \end{cases}$$

Assume that chairs cannot be rearranged, repainted, or damaged.

—This is an abstract model of a complex eternal physical truth.

# Santa Fe process—a model of a random consistent text

Let $(K_i)_{i=1}^{\infty}$ be a sequence of random variables $K_i : \Omega \to \mathbb{N}$.

Let $(Z_k)_{k=1}^{\infty}$ be a sequence of random bits $Z_k : \Omega \to \{0, 1\}$.

The Santa Fe process $(X_i)_{i=1}^{\infty}$ is an infinite sequence of pairs

$$X_i := (K_i, Z_{K_i}).$$

## A semantic interpretation

Process $(X_i)_{i\in\mathbb{Z}}$ is a sequence of random propositions consistently describing the abstract reality, i.e., random chair colors $(Z_k)_{k=1}^{\infty}$:

- Proposition $X_i = (k, z)$ asserts that the $k$-th chair of the row has color $z$, in such way that one can determine both $k$ and $z$.
- For $X_i = (k, z)$ and $X_j = (k', z')$ we do not know in advance which chairs they describe but $k = k' \implies z = z'$.

# Zipfian Santa Fe processes

Let $(K_i)_{i=1}^{\infty}$ and $(Z_k)_{k=1}^{\infty}$ be independent IID processes, where

$$P(K_i = k) \propto k^{-\alpha}, \quad \alpha > 1,$$
$$P(Z_k = 0) = P(Z_k = 1) = \frac{1}{2}.$$

Consider the guessing function:

$$g(k, x_1^n) = \begin{cases} 0 & \text{if for } 1 \le i \le n,\ x_i = (k, z) \implies x_i = (k, 0), \\ 1 & \text{if for } 1 \le i \le n,\ x_i = (k, z) \implies x_i = (k, 1), \\ 2 & \text{else,} \end{cases}$$

and the set of effectively described facts, i.e., chairs:

$$\mathbb{U}_g(X_1^n | Z_1^\infty) := \left\{ l \in \mathbb{N} : g(k, X_1^n) = Z_k \text{ for all } k \le l \right\}.$$

We obtain Herdan-Heaps' power law for described facts

$$\lim_{n\to\infty} \frac{\mathbb{E}\ \#\,\mathbb{U}_g(X_1^n | Z_1^\infty)}{n^{1/\alpha}} \in (0, \infty).$$

# A general model of effective description of facts

- Let $(z_k)_{k=1}^{\infty}$ be a collection of facts $z_k \in \{0, 1\}$.
- We will denote finite texts $x_m^n := (x_m, x_{m+1}, \dots, x_n)$.
- Let $g : \mathbb{N} \times \mathbb{X}^* \to \{0, 1, 2\}$ be a (computable) function.
- We will say that text $x_m^n$ describes exactly $l$ facts if

$$g(k, x_m^n) = z_k \text{ for all } k \le l \text{ and } g(l+1, x_m^n) \ne z_{l+1}.$$

- In this case, we will write

$$\mathbb{U}_g(x_m^n | z_1^{\infty}) := \{1, 2, \dots, l\} .$$

# Independence of facts

To make the facts more abstract, we will assume that they are independent and maximally unpredictable.

We can do it in two ways:

1. Algorithmically independent facts:
   We keep individual bits $z_k \in \{0, 1\}$ intact but we assume Kolmogorov complexity $\mathbb{H}(z_1^k) \geq k - c$ for some $c > 0$.
   $\implies$ The text model is a perigraphic process.
2. Probabilistically independent facts:
   We replace individual bits $z_k \in \{0, 1\}$ by random variables $Z_k : \Omega \to \{0, 1\}$ and we assume $P(Z_1^k = z_1^k) = 2^{-k}$.
   $\implies$ The text model is a strongly nonergodic process.

Thus we assume a compressed representation of described reality.

You might have heard of Chaitin's halting probability $\Omega$, which is a compressed representation of mathematical truth.

# Mutual information vs. common informations

The Shannon mutual information:

$$I(X;Y) := H(X) + H(Y) - H(X,Y).$$

The Gács-Körner and modified Wyner common information:

$$C^-(X;Y) := \sup_{W:W=f(X)=g(Y)} H(W),$$

$$C^+(X;Y) := \inf_{W:X\perp\!\!\!\perp Y|W} H(W).$$

We have

$$0 \le C^-(X;Y) \le I(X;Y) \le C^+(X;Y) \le H(X), H(Y).$$

These inequalities can be strict!

# Toward theorems about facts and words

The number of distinct words resembles $\boldsymbol{C^+(X;Y)}$.

**Conjecture I**

The power law growth of mutual information is slower than the power law growth of the number of distinct words in texts of an increasing length.

The number of independent facts resembles $\boldsymbol{C^-(X;Y)}$.

**Conjecture II**

The power law growth of the number of independent facts described in texts of an increasing length is slower than the power law growth of mutual information.

# Hilberg exponents for the power-law growth

To measure power-law growth, we introduce Hilberg exponent

$$\operatorname*{hilb}_{n\to\infty} s(n) := \limsup_{n\to\infty} \frac{\log s(n)}{\log n}.$$

We have $\operatorname*{hilb}_{n\to\infty} n^\beta = \beta$.

## Theorem 0 (excess bounds)

For a stationary process $(X_i)_{i=1}^\infty$ over a finite alphabet:

$$\begin{array}{ccc} \operatorname*{hilb}_{n\to\infty} \left[H(X_1^n) - hn\right] & = & \operatorname*{hilb}_{n\to\infty} I(X_1^n; X_{n+1}^{2n}) \in [0,1] \\ \text{\rotatebox{90}{$\geq$}} & & \\ \operatorname*{hilb}_{n\to\infty} \left[\mathbb{E}\,\mathbb{H}(X_1^n) - hn\right] & = & \operatorname*{hilb}_{n\to\infty} \mathbb{E}\,\mathbb{I}(X_1^n; X_{n+1}^{2n}) \in [0,1] \end{array}$$

# Theorems about mutual information and words

Consider a stationary process $(X_i)_{i=1}^{\infty}$ over a finite alphabet.

- Let $V(G)$ be the set of nonterminals in a grammar $G$.
- For a grammar transform $\Gamma$, let $V_\Gamma(w) := V(\Gamma(w))$.
- $L(w)$ is the length of the maximal repetition in string $w$.

**Theorem 1 (grammar-based codes)**

For a minimal grammar transform $\Gamma$, we have

$$\operatorname*{hilb}_{n\to\infty} \mathbb{E}\, \mathbb{I}(X_1^n; X_{n+1}^{2n}) \le \operatorname*{hilb}_{n\to\infty} \mathbb{E}\, L(X_1^n)\, \#\, V_\Gamma(X_1^n).$$

- Let $V_k(w)$ be the set of substrings of length $k$ of string $w$.
- For a function $\mathbb{M} : \mathbb{X}^* \to \mathbb{N}$, let $V_{\mathbb{M}}(w) := V_{\mathbb{M}(w)}(w)$.

**Theorem 2 (PPM vocabulary)**

For the PPM Markov order $\mathbb{M} : \mathbb{X}^* \to \mathbb{N}$, we have

$$\operatorname*{hilb}_{n\to\infty} \mathbb{E}\, \mathbb{I}(X_1^n; X_{n+1}^{2n}) \le \operatorname*{hilb}_{n\to\infty} \mathbb{E}\, \left[\mathbb{M}(X_1^n) + \#\, V_{\mathbb{M}}(X_1^n)\right].$$

# Theorems about facts and mutual information

**Theorem 3 (ergodic perigraphic processes)**

Let $(X_i)_{i=1}^{\infty}$ be a stationary process over a finite alphabet, let $(z_k)_{k=1}^{\infty}$ be a collection of algorithmically independent facts, and let $g : \mathbb{N} \times \mathbb{X}^* \to \{0, 1, 2\}$ be computable. We have

$$\underset{n\to\infty}{\mathrm{hilb}}\ \mathbb{E}\,\#\,\mathbb{U}_g(X_1^n|z_1^{\infty}) \le \underset{n\to\infty}{\mathrm{hilb}}\ \mathbb{E}\,\mathbb{I}(X_1^n; X_{n+1}^{2n}).$$

**Theorem 4 (strongly nonergodic processes)**

Let $(X_i)_{i=1}^{\infty}$ be a stationary process over a finite alphabet, let $(Z_k)_{k=1}^{\infty}$ be a collection of probabilistically independent facts measurable with respect to the shift invariant $\sigma$-field of $(X_i)_{i=1}^{\infty}$, and let $g : \mathbb{N} \times \mathbb{X}^* \to \{0, 1, 2\}$ be any function. We have

$$\underset{n\to\infty}{\mathrm{hilb}}\ \mathbb{E}\,\#\,\mathbb{U}_g(X_1^n|Z_1^{\infty}) \le \underset{n\to\infty}{\mathrm{hilb}}\ I(X_1^n; X_{n+1}^{2n}).$$

# The main result of this talk

Is language structure a mathematical consequence of descriptive meaningfulness, i.e., effective reference of texts to some reality?

As for double articulation, YES since we have shown that:

The number of distinct words in a finite text is roughly greater than the number of independent facts described by the text.

The above proposition is a general result in information theory connected to Hilberg's hypothesis and Herdan-Heaps' law.

**Applications to natural language:**

- The number of words grows like a power of the text length.
- Can we lower-bound the number of described facts?
- Can we make the formal concept of a fact less static?

# An account of descriptive meaningfulness

- Meaningfulness of texts can be understood as:
  1. effective description of an external or imagined reality (descriptive meaningfulness);
  2. internal cohesion of the narration or the discourse (cohesive meaningfulness);
  3. effective control of an external reality toward some goal (telic meaningfulness).
- The theorems about facts and words concern only descriptive meaningfulness.
- Realities are both described and created by texts.
- Realities evolve in time, which may cause $\boldsymbol{E} < \infty$.
- Complexity of realities is extended by technical tools created by humans over ages (like script or internet).

# Toward cohesive and telic meaningfulness

- Here our understanding and modeling is less advanced.
- Random hierarchical association (RHA) processes: selection and replication of hierarchical memes.
- Cohesive meaningfulness:
  - power-law logarithmic growth of maximal repetition, power-law growth of conditional Rényi entropy;
  - power-law decay of letterwise mutual information, large scale context-free structures.
- Telic meaningfulness:
  - arrow of time, (un)bounded accumulation of knowledge, (no) point Omega (singularity), AMS processes;
  - control of a (non)random environment, (non)deterministic interpretation of texts, positive entropy rate.
- Does cohesive m-fulness imply descriptive & telic m-fulness?
- Can animal communication, music, mathematical vernacular, and programming languages shed light onto meaningfulness?
- Natural meaningful texts vs. idealized meaningful texts.

# Idealization in statistical language models

- Stochastic processes = idealized models of possible texts.
  - This idealization becomes clear upon a closer scrutiny of these models, which takes effort, time, and imagination.
  - Imagination is a skill constructed through examples.
  - Linguistic and math intuitions can help each other.

- Sorts of idealization in stochastic processes:
  - actual or potential infinities (unbounded texts),
  - unbounded sources of (algorithmic) randomness,
  - infinite precision,
  - infinite recursion,
  - (conditional) computability of distributions,
  - rigid structure of mathematical definitions,
  - plethora of processes that cannot be effectively defined...
  - ... but these processes can be theorized about.

# It's time for a synthesis!

*Entropy not only speaks the language of arithmetic;*
*it also speaks the language of language.*

— *Warren Weaver (1949)*

*It is an irony of 20th century linguistics that Shannon's theory of information, though explicitly linked to semantics, was deemed irrelevant by linguists, while Chomsky's formal syntax, though explicitly dissociated from semantics, was adopted as the default theory of natural language.*

— *Christian Bentz (2018)*

Thank you!