The Vocabulary of Grammar-Based Codes and the Logical Consistency of Texts

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Institute of Computer Science
Polish Academy of Sciences

Herdan's law (an integrated version of Zipf's law)

Consider texts in a natural language (such as English):

- **V** — the number of different words in the text,
- **n** — the length of the text.

We observe the relationship

$$\mathbf{V} \propto \mathbf{n}^{\boldsymbol{\beta}},$$

where β is between $\mathbf{0.5}$ a $\mathbf{1}$ depending on a text.

- *Władysław Kuraszkiewicz, Józef Łukaszewicz (1951),*
- *Pierre Guiraud (1954),*
- *Gustav Herdan (1964),*
- *H. S. Heaps (1978).*

Is Herdan's law evidence of "an animate process"?

> *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)*

The monkey-typing explanation

Zipf's and Herdan's law are observed if the letters and spaces in the text are obtained by pressing keys at random.

- Benoit B. Mandelbrot (1953),
- George A. Miller (1957).

The new explanation of Herdan's law

We will prove a theorem which can be stated informally in this way, for $\beta \in (0, 1)$:

> If a text of length $\mathbf{n}$ describes $\geq \mathbf{n}^{\beta}$ independent facts in a repetitive way then the text contains $\geq \mathbf{n}^{\beta} / \log \mathbf{n}$ distinct words.

For the formal statement, we shall adopt two postulates:

1. Words are understood as nonterminal symbols in the shortest grammar-based encoding of the text.
2. Texts are emitted by a finite-energy strongly nonergodic source.
3. Facts are independent binary variables which can be predicted from the text in a shift-invariant way.

A context-free grammar that generates one text

$$\left\{\begin{array}{l} \mathbf{A_1} \to \mathbf{A_2A_2A_4A_5}\textbf{dear\_children}\mathbf{A_5A_3}\textbf{all.} \\ \mathbf{A_2} \to \mathbf{A_3}\textbf{you}\mathbf{A_5} \\ \mathbf{A_3} \to \mathbf{A_4}\textbf{\_to\_} \\ \mathbf{A_4} \to \textbf{Good\_morning} \\ \mathbf{A_5} \to \textbf{,\_} \end{array}\right\}.$$

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

The vocabulary size and grammar-based codes

The vocabulary size of a grammar:

$$\mathbb{V}[\mathbf{G}] := \mathbf{n}, \quad \text{jeżeli} \quad \mathbf{G} = \left\{ \begin{array}{l} \mathbf{A_1} \to \alpha_1, \\ \mathbf{A_2} \to \alpha_2, \\ \ldots, \\ \mathbf{A_n} \to \alpha_n \end{array} \right\}.$$

A grammar-based code is a function of form $\mathbf{C} = \mathbf{B}(\mathbf{\Gamma}(\cdot))$, where

1. a grammar transform $\mathbf{\Gamma} : \mathbb{X}^+ \to \mathcal{G}$, for each string $\mathbf{w} \in \mathbb{X}^+$, returns a gramar $\mathbf{\Gamma(w)}$ that generates this string.
2. a grammar encoder $\mathbf{B} : \mathcal{G} \to \mathbb{X}^+$ codes the grammar as (another) string.

- John C. Kieffer, En-hui Yang (2000),
- Moses Charikar, Eric Lehman, ..., Abhi Shelat (2005).

Admissibly minimal codes

Let $\mathbb{X} = \{0, 1, ..., D - 1\}$. A grammar transform Γ and the code $B(\Gamma(\cdot))$ are called admissibly minimal if

1. $|B(\Gamma(w))| \le |B(G)|$ for each grammar G that generates w,
2. the encoder has the form $B(G) = B_S^*(B_N(G))$,
3. B_N encodes the grammar

$$G = \{A_1 \to \alpha_1, A_2 \to \alpha_2, ..., A_n \to \alpha_n\}$$

   as a string of integers

$$B_N(G) := F_1^*(\alpha_1) D F_2^*(\alpha_2) D ... D F_n^*(\alpha_n)(D + 1),$$

   using $F_i(x) := x$ for $x \in \mathbb{X}$ and $F_i(A_j) := D + 1 + j - i$,
4. $B_S : \{0\} \cup \mathbb{N} \to \mathbb{X}^+$ is an injection, the set $B_S(\{0\} \cup \mathbb{N})$ is prefix-free, $|B_S(\cdot)|$ is non-decreasing and

$$\limsup_{n \to \infty} |B_S(n)| / \log_D n = 1.$$

Two classes of stochastic processes

Let $(X_i)_{i\in\mathbb{Z}}$ be a stochastic process on the space $(\Omega, \mathfrak{J}, P)$, where $X_i : \Omega \to \mathbb{X}$ for a countable alphabet $\mathbb{X}$. Denote blocks as $X_{m:n} := (X_i)_{m\le k\le n}$.

The proces $(X_i)_{i\in\mathbb{Z}}$ is called strongly nonergodic if there exist variables $(Z_k)_{k\in\mathbb{N}} \sim \text{IID}$, $P(Z_k = 0) = P(Z_k = 1) = \frac{1}{2}$, and functions $s_k : \mathbb{X}^* \to \{0,1\}$, $k \in \mathbb{N}$, such that

$$\lim_{n\to\infty} P(s_k(X_{t+1:t+n}) = Z_k) = 1, \qquad \forall t \in \mathbb{Z}.$$

$Y = \sum_{k\in\mathbb{N}} 2^{-k} Z_k$ is measurable against the shift-invariant σ-field.

The process $(X_i)_{i\in\mathbb{Z}}$ is called a finite-energy process if

$$P(X_{t+1:t+m} = u | X_{t-n:t} = w) \le Kc^m, \qquad \forall t \in \mathbb{Z}.$$

The main result

$$U_\delta(n) := \{k \in \mathbb{N} : P(s_k(X_{1:n}) = Z_k) \geq \delta\}.$$

Theorem 1

Let $(X_i)_{i\in\mathbb{Z}}$ be a stationary finite-energy strongly nonergodic process over a finite alphabet $\mathbb{X}$. Suppose that

$$\liminf_{n\to\infty} \frac{\text{card}\, U_\delta(n)}{n^\beta} > 0$$

for some $\beta \in (0,1)$ and $\delta \in (\frac{1}{2}, 1)$. Then

$$\limsup_{n\to\infty} \mathbb{E}\left(\frac{\mathbb{V}[\Gamma(X_{1:n})]}{n^\beta(\log n)^{-1}}\right)^p > 0, \quad p > 1,$$

for any admissibly minimal grammar transform Γ.

The first associated result

Denote the mutual information between **n**-blocks

$$E(n) := I(X_{1:n}; X_{n+1:2n}) = \mathbb{E} \log \frac{P(X_{1:2n})}{P(X_{1:n})P(X_{n+1:2n})}.$$

Theorem 2

Let $(X_i)_{i\in\mathbb{Z}}$ be a stationary strongly nonergodic process over a finite alphabet $\mathbb{X}$. Suppose that

$$\liminf_{n\to\infty} \frac{\operatorname{card} U_\delta(n)}{n^\beta} > 0$$

for some $\beta \in (0,1)$ and $\delta \in (\frac{1}{2}, 1)$. Then

$$\limsup_{n\to\infty} \frac{E(n)}{n^\beta} > 0.$$

The second associated result

Theorem 3

Let $(X_i)_{i\in\mathbb{Z}}$ be a stationary finite-energy process over a finite alphabet $\mathbb{X}$. Suppose that

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

for some $\beta \in (0,1)$. Then

$$\limsup_{n\to\infty} \mathbb{E}\left(\frac{\mathbb{V}[\Gamma(X_{1:n})]}{n^{\beta}(\log n)^{-1}}\right)^{p} > 0, \quad p > 1,$$

for any admissibly minimal grammar transform Γ.

Mutual information for natural language

Basing on Shannon's (1950) estimates of conditional entropy of printed English, Hilberg (1990) conjectured that mutual information between two **n**-blocks drawn from natural language satisfies

$$\mathbf{E(n)} \asymp \mathbf{n}^{\boldsymbol{\beta}}, \quad \boldsymbol{\beta} \approx \mathbf{1/2}.$$

- Theorem 2 – a rational motivation of Hilberg's conjecture
- Theorem 3 – Hilberg's conjecture implies Herdan's law

Besides that, Theorem 1 indicates

- why Herdan's law may be observed for the same text translated into different languages,
- why certain variation of the exponent in Herdan's law may be expected depending on a text.

Entropy, pseudoentropy, and code length

Denote the entropy of the **n**-block and entropy rate as:

$$\mathbf{H(n) := H(X_{1:n}) = -\mathbb{E} \log P(X_{1:n}), \quad h := \lim_{n\to\infty} H(n)/n.}$$

Define also "pseudoentropy"

$$\mathbf{H^U(n) := hn + [\log 2 - \eta(\delta)] \cdot card\, U_\delta(n).}$$

Let $\mathbb{X} = \{0, 1, ..., D-1\}$ and $\mathbf{C = B(\Gamma(\cdot))}$ be an admissibly minimal codes. Put its expected length

$$\mathbf{H^C(n) := \mathbb{E}\ |C(X_{1:n})| \log D.}$$

We have inequality

$$\mathbf{H^C(u) \geq H(n) \geq H^U(n)}$$

and equality of rates

$$\mathbf{\lim_{n\to\infty} H^C(n)/n = \lim_{n\to\infty} H(n)/n = \lim_{n\to\infty} H^U(n)/n = h.}$$

The lower bound for the excess code length $\mathbf{E^C(n)}$

Let a function $\mathbf{f} : \mathbb{N} \to \mathbb{R}$ satisfy $\lim_k \mathbf{f(k)/k = 0}$
and $\mathbf{f(n) \geq 0}$ for all but finitely many $\mathbf{n}$.
Then we have $\mathbf{2f(n) - f(2n) \geq 0}$ for infinitely many $\mathbf{n}$.

The equalities and inequalities on the previous slide yield

$$\liminf_{n\to\infty} \frac{\operatorname{card} U_\delta(n)}{n^\beta} > 0 \implies \limsup_{n\to\infty} \frac{E^C(n)}{n^\beta} > 0, \quad \text{(Th. 1)}$$

$$\liminf_{n\to\infty} \frac{\operatorname{card} U_\delta(n)}{n^\beta} > 0 \implies \limsup_{n\to\infty} \frac{E(n)}{n^\beta} > 0, \quad \text{(Th. 2)}$$

$$\liminf_{n\to\infty} \frac{E(n)}{n^\beta} > 0 \implies \limsup_{n\to\infty} \frac{E^C(n)}{n^\beta} > 0. \quad \text{(Th. 3)}$$

for $\mathbf{E(n) = 2H(n) - H(2n)}$ and $\mathbf{E^C(n) = 2H^C(n) - H^C(2n)}$.

The upper bound for the excess code length $\mathbf{E}^C(n)$

$$\mathbf{E}^C(n) = \mathbb{E}\left[|C(X_{1:n})| + |C(X_{n+1:2n})| - |C(X_{1:2n})|\right] \log D.$$

For an admissibly minimal code $C = B(\Gamma(\cdot))$, we have

$$|C(u)| + |C(v)| - |C(w)| \le W_0 \mathbb{V}[\Gamma(w)](1 + \mathbb{L}(w)),$$

where $w = uv$ for $u, v \in \mathbb{X}^+$, $\mathbb{L}(w)$ denotes the maximal length of a repeat in w, and $W_0 = |B_S(D+1)|$.

For a finite-energy process $(X_i)_{i\in\mathbb{Z}}$,

$$\sup_{n\in\mathbb{N}} \mathbb{E}\left(\frac{\mathbb{L}(X_{1:n})}{\log n}\right)^q < \infty, \quad q > 0.$$

The binary exchangeable process

Consider a family of binary IID processes

$$P(X_{1:n} = x_{1:n}|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}.$$

Construct such a process $(X_i)_{i\in\mathbb{Z}}$ that

$$P(X_{1:n} = x_{1:n}) = \int_0^1 P(X_{1:n} = x_{1:n}|\theta)\pi(\theta)d\theta$$

for a prior $\pi(\theta) > 0$. For $Y = \lim_n n^{-1} \sum_{i=1}^{n} X_i$ we have

$$P(Y \leq y) = \int_0^y \pi(\theta)d\theta.$$

Process $(X_i)_{i\in\mathbb{Z}}$ is strongly nonergodic, because Y has a continuous distribution. However, block $X_{1:n}$ is conditionally independent from $X_{n+1:2n}$ given the sum $S_n := \sum_{i=1}^{n} X_i$. Thus

$$E(n) = I(X_{1:n}; X_{n+1:2n}) = I(S_n; X_{n+1:2n}) \leq H(S_n) \leq \log(n+1).$$

The process which I invented at Santa Fe Institute

Let a process $(X_i)_{i\in\mathbb{Z}}$ have the form

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

where $(K_i)_{i\in\mathbb{Z}}$ and $(Z_k)_{k\in\mathbb{N}}$ are independent IID processes,

$$P(K_i = k) = k^{-1/\beta}/\zeta(\beta^{-1}), \qquad \beta \in (0,1),$$
$$P(Z_k = z) = \tfrac{1}{2}, \qquad z \in \{0,1\}.$$

A linguistic interpretation

Process $(X_i)_{i\in\mathbb{Z}}$ is a sequence of random statements consistently describing the state of an "earlier drawn" random object $(Z_k)_{k\in\mathbb{N}}$. $X_i = (k, z)$ asserts that the k-th bit of $(Z_k)_{k\in\mathbb{N}}$ has value $Z_k = z$.

- We have $\mathbf{card}\, U_\delta(n) \geq A n^\beta$.
- Unfortunately, the alphabet $\mathbb{X} = \mathbb{N} \times \{0,1\}$ is infinite.

Stationary (variable-length) coding of this process

A function $\mathbf{f} : \mathbb{X} \to \mathbb{Y}^+$ is extended to a function $\mathbf{f}^{\mathbb{Z}} : \mathbb{X}^{\mathbb{Z}} \to \mathbb{Y}^{\mathbb{Z}}$,

$$\mathbf{f}^{\mathbb{Z}}((x_i)_{i\in\mathbb{Z}}) := \ldots f(x_{-1})f(x_0).f(x_1)f(x_2)\ldots, \qquad x_i \in \mathbb{X}.$$

For a measure ν on $(\mathbb{Y}^{\mathbb{Z}}, \mathfrak{Y}^{\mathbb{Z}})$ we define its stationary mean $\bar{\nu}$ as

$$\bar{\nu}(A) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \nu \circ T^{-i}(A),$$

where $T((y_i)_{i\in\mathbb{Z}}) := (y_{i+1})_{i\in\mathbb{Z}}$ is the shift.

Theorem 4

Let $\mu = P((X_i)_{i\in\mathbb{Z}} \in \cdot)$ for the process from the previous slide. Put $\mathbb{Y} = \{0, 1, 2\}$ and $f(k, z) := b(k)z2$, where $\underline{1b(k) \in \{0, 1\}^+}$ is the binary expansion of $\mathbf{k}$. A process with measure $\mu \circ (f^{\mathbb{Z}})^{-1}$ satisfies the hypothesis of Th. 1 for $\beta > \mathbf{0.78}$.

A mixing process

Let a process $(X_i)_{i\in\mathbb{Z}}$ have the form

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

where $(K_i)_{i\in\mathbb{Z}}$ and $(Z_{ik})_{i\in\mathbb{Z}}$, $k \in \mathbb{N}$, are independent,

$$P(K_i = k) = k^{-1/\beta}/\zeta(\beta^{-1}), \qquad (K_i)_{i\in\mathbb{Z}} \sim \text{IID},$$

whereas $(Z_{ik})_{i\in\mathbb{Z}}$ are Markov chains with

$$P(Z_{ik} = z) = \frac{1}{2},$$
$$P(Z_{ik} = z | Z_{i-1,k} = z) = 1 - p_k.$$

Object $(Z_{ik})_{k\in\mathbb{N}}$ described by text $(X_i)_{i\in\mathbb{Z}}$ is a function of time i.

- We have $\liminf_{n\to\infty} E(n)/n^{\beta} > 0$ for $p_k \le P(K_i = k)$.
- The stationary coding of this process is an ergodic process and also satisfies $\liminf_{n\to\infty} E(n)/n^{\beta} > 0$.

Can we check which explanation is better?

Monkey-typing explanation

Zipf's and Herdan's law are observed if the letters and spaces in the text are obtained by pressing keys at random.

vs.

New explanation

If a text of length $\mathbf{n}$ describes $\geq \mathbf{n}^{\beta}$ independent facts in a repetitive way then the text contains $\geq \mathbf{n}^{\beta} / \log \mathbf{n}$ distinct words.

Can we estimate mutual information well?

1. Can we strengthen Theorems 1, 2, and 3?
   - Consider asymptotically mean stationary (AMS) processes.
   - Infer almost sure growth of vocabulary.
   - Replace $\limsup_{n\to\infty}$ with $\liminf_{n\to\infty}$.
2. Let $C(u)$ be the shortest program that generates u.
   Then $E^C(n)$ is the algorithmic information between blocks.
   - Let $(\omega_k)_{k\in\mathbb{N}}$ be an algorithmically random real in $(0,1)$.
     Mind that $E(n) = 0$ but $E^C(n) \asymp n^\beta$ for $X_i := (K_i, \omega_{K_i})$.
   - Can we use some universal codes to distinguish between some AMS sources with little vs. large $E^C(n)$?
   - Can we use vocabulary of grammar-based codes to distinguish between some AMS sources with little vs. large $E^C(n)$?
3. Do there exist admissibly minimal codes that are computable in polynomial time? (Or sufficiently similar codes?)
   - Let $(X_i)_{i\in\mathbb{Z}}$ be a binary IID process. Then
     $\mathbb{V}[\Gamma(X_{1:n}] = \Omega\left(\sqrt{\frac{hn}{\log n}}\right)$ for irredicible grammar transforms.

My work

- Ł. Dębowski, (2012). *Mixing, Ergodic, and Nonergodic Processes with Rapidly Growing Information between Blocks.* IEEE Transactions on Information Theory, vol. 58, pp. 3392-3401.
- 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.
- Ł. Dębowski, (2010). *Variable-Length Coding of Two-Sided Asymptotically Mean Stationary Measures.* Journal of Theoretical Probability, 23:237–256.
- Ł. Dębowski, (2007). *Menzerath's law for the smallest grammars.* In: P. Grzybek, R. Koehler, eds., Exact Methods in the Study of Language and Text. Mouton de Gruyter. (77–85)

`www.ipipan.waw.pl/~ldebowsk`