

A Link Between the Number of Set Phrases in a Text and the Number of Described Facts

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Centrum Wiskunde & Informatica
Amsterdam, The Netherlands

QUALICO 2009

Zipf's ideas about statistical properties of texts

*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** on Earth, or merely the latest thunderstorm on Earth? It seems that the only criteria would be the **arrangement of occurrences** of the elements, and 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)

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

Mandelbrot (1954), Miller (1957):

texts are generated by independent sampling of single characters



the frequencies of **space-to-space chunks** in the text are distributed according to a power-law

Dębowski (2006):

texts repetitively convey certain information



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.

Recent developments in the new class of explanations

Theorem (an informal expression)

If an N -letter long text describes N^β independent facts in a repetitive fashion then the text contains at least $N^\beta / \log N$ different set phrases.

... for a certain mathematical model of facts, texts, and set phrases.

Ł. Dębowski, *On the vocabulary of grammar-based codes and the logical consistency of texts*, arxiv.org/abs/0810.3125

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.

$$\left\{ \begin{array}{l} \mathbf{A}_1 \mapsto \mathbf{A}_2 \mathbf{A}_2 \mathbf{A}_4 \mathbf{A}_5 \text{dear_children} \mathbf{A}_5 \mathbf{A}_3 \text{all.} \\ \mathbf{A}_2 \mapsto \mathbf{A}_3 \text{you} \mathbf{A}_5 \\ \mathbf{A}_3 \mapsto \mathbf{A}_4 \text{_to_} \\ \mathbf{A}_4 \mapsto \text{Good_morning} \\ \mathbf{A}_5 \mapsto \text{, -} \end{array} \right\}$$

*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 $i > 1$.
(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 , $k = 1, 2, 3, \dots$, 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 , $i = 1, 2, 3, \dots$, 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

$$\lim_{n \rightarrow \infty} \mathbf{P}(\mathbf{s}_k(\mathbf{X}_{i+1}, \mathbf{X}_{i+2}, \dots, \mathbf{X}_{i+n}) = \mathbf{Z}_k) = 1 \text{ for } i = 1, 2, 3, \dots$$

Two quantities and the claim

For the process as before, put $\mathbf{X}_1^n := (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n)$. Let

- $\mathbf{U}_\delta(\mathbf{n}) := \{\mathbf{k} \in \mathbb{N} : \mathbf{P}(\mathbf{s}_\mathbf{k}(\mathbf{X}_1^n) = \mathbf{Z}_\mathbf{k}) \geq \delta\}$
be the **set of sufficiently well predictable facts** and
- $\mathbb{V}[\Gamma(\mathbf{w})]$ be the **number of distinct nonterminal symbols** in the shortest grammar-based compression $\Gamma(\mathbf{w})$ of a string \mathbf{w} .

Theorem

For $\delta \in (1/2, 1)$, $\beta \in (0, 1)$, and $p > 1$,

$$\liminf_{n \rightarrow \infty} \frac{|\mathbf{U}_\delta(\mathbf{n})|}{n^\beta} > 0 \implies \limsup_{n \rightarrow \infty} \mathbb{E} \left(\frac{\mathbb{V}[\Gamma(\mathbf{X}_1^n)]}{n^\beta (\log n)^{-1}} \right)^p > 0.$$

This constitutes an explanation of Zipf's law if we can construct some **linguistically motivated** processes that satisfy the **red inequality**.

A process which almost satisfies the assumptions

Consider a process $(\mathbf{X}_i)_{i \in \mathbb{Z}}$ with variables

$$\mathbf{X}_i := (\mathbf{K}_i, \mathbf{Z}_{\mathbf{K}_i}), \quad \mathbf{P}(\mathbf{K}_i = \mathbf{k}) \propto \frac{1}{\mathbf{k}^\alpha}, \quad \alpha = \frac{1}{\beta},$$

where processes $(\mathbf{K}_i)_{i \in \mathbb{Z}}$ and $(\mathbf{Z}_k)_{k \in \mathbb{N}}$ are independent IID.

Process $(\mathbf{X}_i)_{i \in \mathbb{Z}}$ looks like a sequence of propositions.

- 1 Proposition $\mathbf{X}_i = (\mathbf{k}, \mathbf{z})$ asserts that the \mathbf{k} -th bit of $(\mathbf{Z}_k)_{k \in \mathbb{N}}$ has a value \mathbf{z} , in such way that both \mathbf{k} and \mathbf{z} can be **known**.
- 2 Probability of a bit being described follows a **power law**.
- 3 The description is **consistent**, i.e., $\mathbf{k} = \mathbf{k}' \implies \mathbf{z} = \mathbf{z}'$
for $\mathbf{X}_i = (\mathbf{k}, \mathbf{z})$ and $\mathbf{X}_j = (\mathbf{k}', \mathbf{z}')$.

A process which satisfies the assumptions exactly

We need a similar process over a **finite** alphabet, such as $\{0, 1, \square\}$.

- $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \dots$ — the original sequence of $\mathbf{X}_i := (\mathbf{K}_i, \mathbf{Z}_{\mathbf{K}_i})$,
- $\mathbf{Y}_1 \mathbf{Y}_2 \mathbf{Y}_3 \dots := f(\mathbf{X}_1) f(\mathbf{X}_2) f(\mathbf{X}_3) \dots$ — **pair-to-string encoding**:

- $f(\mathbf{k}, z) := \mathbf{b}(\mathbf{k}) z \square$,
- $\mathbf{b}(\mathbf{k})$ is the binary representation of \mathbf{k} (*without the initial 1*),
- \square is an **overt space**,

- $\bar{\mathbf{Y}}_1, \bar{\mathbf{Y}}_2, \bar{\mathbf{Y}}_3, \dots$ — the **stationary mean** of $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \dots$.

The process $\bar{\mathbf{Y}}_1, \bar{\mathbf{Y}}_2, \bar{\mathbf{Y}}_3, \dots$ satisfies the assumptions of the main theorem for $\beta = \alpha^{-1} > \mathbf{0.7728\dots}$.

Ł. Dębowski, *Variable length coding of two-sided asymptotically mean stationary measures*, www.ipipan.waw.pl/~ldebowsk

How far are we from Mandelbrot's explanation?

Our example process accidentally seems to exhibit two properties:

the frequencies of
space-to-space chunks
obey a power law

the frequencies of
set phrases
obey a power law

Similar laws seem to hold also for natural language texts
— *and for 'monkey-typed' texts* —
but the **parameter values** and **interpretation of spaces** may differ.

Our '**overt spaces**' are called '**commas**' or '**dots**' in other domains.

Ideas for the future research

The experimental side:

- Can one confirm that the **Voynich manuscript** is a hoax because it contains too few set phrases?
- Distributions of space-to-space chunks and set phrases should be **systematically compared** for various texts and **idealized stochastic processes**.

To interpret the data well, we need more theory:

- The present result is a law for expectations.
- We expect a related probabilistic **strong law**, which applies algorithmic information theory.

Building blocks for the main theorem

Properties of the shortest grammar-based code:

$$\mathbb{V}[\Gamma(\mathbf{X}_1^{2^n})] \geq \frac{\|\Gamma(\mathbf{X}_1^n)\| + \|\Gamma(\mathbf{X}_{n+1}^{2^n})\| - \|\Gamma(\mathbf{X}_1^{2^n})\|}{M \cdot (1 + \mathbb{L}(\mathbf{X}_1^{2^n}))}. \quad (1)$$

A bound for the length of repetition in finite energy processes:

$$\sup_{n \in \mathbb{N}} \mathbb{E} (\mathbb{L}(\mathbf{X}_1^n) / \log n)^q < \infty, \quad q > 0. \quad (2)$$

Hölder's inequality:

$$(\mathbb{E} \mathbf{U}^p)^{1/p} (\mathbb{E} \mathbf{T}^q)^{1/q} \geq \mathbb{E} (\mathbf{U}\mathbf{T}), \quad (p-1)(q-1) = 1. \quad (3)$$

Excess bounding lemma:

$$\limsup_{n \rightarrow \infty} [2f(n) - f(2n)] \geq 0 \iff f(n) \geq 0 \text{ and } \lim_{n \rightarrow \infty} \frac{f(n)}{n} = 0. \quad (4)$$

Universal coding and ergodic decomposition of stationary processes:

$$\mathbb{E} \|\Gamma(\mathbf{X}_1^n)\| \geq \mathbf{H}(\mathbf{X}_1^n) \geq hn + C_\delta |\mathbf{U}_\delta(\mathbf{n})|, \quad (5)$$

$$\lim_{n \rightarrow \infty} \frac{\mathbb{E} \|\Gamma(\mathbf{X}_1^n)\|}{n} = \lim_{n \rightarrow \infty} \frac{\mathbf{H}(\mathbf{X}_1^n)}{n} = \lim_{n \rightarrow \infty} \frac{hn + C_\delta |\mathbf{U}_\delta(\mathbf{n})|}{n} = h. \quad (6)$$