

Słownik kodów gramatykowych a spójność logiczna tekstów

Łukasz Dębowski
ldebowsk@ipipan.waw.pl



Instytut Podstaw Informatyki PAN

Prawo Herdana (scałkowana wersja prawa Zipfa)

Rozpatrujemy teksty w języku naturalnym (np. w j. polskim):

- **V** — liczba różnych słów w tekście,
- **n** — długość tekstu.

Obserwuje się empiryczną zależność

$$V \propto n^\beta,$$

gdzie β waha się między **0.5** a **1** w zależności od zbioru tekstów.

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

Czy prawo Herdana świadczy o „ożywionym procesie”?

*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)

Małpa przy maszynie do pisania



Prawa Zipfa i Herdana zaobserwujemy, jeżeli kolejne **litery** i **spacje** tekstu będziemy uzyskiwać wciskając klawisze **losowo**.

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

Nowe objaśnienie prawa Herdana

Udowodnię twierdzenie, które można nieformalnie wyrazić w sposób następujący, przyjmując $\beta \in (0, 1)$:

Jeżeli **tekst** długości n opisuje $\geq n^\beta$ niezależnych **faktów** w sposób powtarzalny, to tekst ten zawiera $\geq n^\beta / \log n$ różnych **słów**.

Do formalnego wyrażenia przyjmuję trzy postulaty:

- 1 **Słowa** będą rozumiane jako symbole nieterminalne w najkrótszym gramatycznym kodowaniu tekstu.
- 2 **Tekst** jest emitowany przez mocno nieergodyczne źródło informacji o skończonej energii.
- 3 **Fakty** są niezależnymi binarnymi zmiennymi losowymi przewidywalnymi na podstawie tekstu w sposób niezmienniczy ze względu na przesunięcia.

Gramatyka bezkontekstowa generująca jeden tekst

$$G = \left\{ \begin{array}{l} A_1 \rightarrow A_2 A_2 A_4 A_9 A_4 A_3 A_7 A_5 \\ A_2 \rightarrow A_6 A_9 A_6 A_3 \\ A_3 \rightarrow A_9 A_7 A_8, A_5 \\ A_4 \rightarrow \text{Jeszcze_raz} \\ A_5 \rightarrow A_8_nam_ \\ A_6 \rightarrow \text{Sto_lat} \\ A_7 \rightarrow \text{Niech} \\ A_8 \rightarrow _zyje \\ A_9 \rightarrow \!_ \end{array} \right.$$

Sto lat! Sto lat! Niech żyje, żyje nam.

Sto lat! Sto lat! Niech żyje, żyje nam.

Jeszcze raz! Jeszcze raz! Niech żyje, żyje nam.

Niech żyje nam.

Rozmiar słownika i kody gramatyczne

Rozmiar słownika gramatyki:

$$\mathbb{V}[\mathbf{G}] := n, \quad \text{jeżeli} \quad \mathbf{G} = \left\{ \begin{array}{l} \mathbf{A}_1 \rightarrow \alpha_1, \\ \mathbf{A}_2 \rightarrow \alpha_2, \\ \dots, \\ \mathbf{A}_n \rightarrow \alpha_n \end{array} \right\}.$$

Kod gramatyczny to funkcja postaci $\mathbf{C} = \mathbf{B}(\Gamma(\cdot))$, gdzie

- 1 transformacja gramatyczna $\Gamma : \mathbb{X}^+ \rightarrow \mathcal{G}$ dla każdego napisu $\mathbf{w} \in \mathbb{X}^+$ zwraca gramatykę $\Gamma(\mathbf{w})$ generującą ten napis.
- 2 koder $\mathbf{B} : \mathcal{G} \rightarrow \mathbb{X}^+$ koduje tę gramatykę jako (inny) napis.
 - John C. Kieffer, En-hui Yang (2000),
 - Moses Charikar, Eric Lehman, ..., Abhi Shelat (2005).

Kody dopuszczalnie minimalne

Niech $\mathbb{X} = \{0, 1, \dots, D - 1\}$. Transformację gramatykową Γ oraz kod $\mathbf{B}(\Gamma(\cdot))$ nazywam **dopuszczalnie minimalnymi**, jeżeli

- 1 $|\mathbf{B}(\Gamma(\mathbf{w}))| \leq |\mathbf{B}(\mathbf{G})|$ dla każdej gramatyki \mathbf{G} generującej \mathbf{w} ,
- 2 koder ma postać $\mathbf{B}(\mathbf{G}) = \mathbf{B}_S^*(\mathbf{B}_N(\mathbf{G}))$,
- 3 \mathbf{B}_N koduje gramatykę

$$\mathbf{G} = \{\mathbf{A}_1 \rightarrow \alpha_1, \mathbf{A}_2 \rightarrow \alpha_2, \dots, \mathbf{A}_n \rightarrow \alpha_n\}$$

jako ciąg liczb całkowitych

$$\mathbf{B}_N(\mathbf{G}) := \mathbf{F}_1^*(\alpha_1)\mathbf{D}\mathbf{F}_2^*(\alpha_2)\mathbf{D}\dots\mathbf{D}\mathbf{F}_n^*(\alpha_n)(\mathbf{D} + 1),$$

wykorzystując $\mathbf{F}_i(\mathbf{x}) := \mathbf{x}$, $\mathbf{x} \in \mathbb{X}$, $\mathbf{F}_i(\mathbf{A}_j) := \mathbf{D} + 1 + j - i$,

- 4 $\mathbf{B}_S : \{0\} \cup \mathbb{N} \rightarrow \mathbb{X}^+$ jest iniekcją, zbiór $\mathbf{B}_S(\{0\} \cup \mathbb{N})$ jest bezprzedrostkowy, $|\mathbf{B}_S(\cdot)|$ jest funkcją niemalejącą i spełnia

$$\limsup_{n \rightarrow \infty} |\mathbf{B}_S(n)| / \log_D n = 1.$$

Dwie klasy procesów stochastycznych

Niech $(\mathbf{X}_i)_{i \in \mathbb{Z}}$ będzie procesem stochastycznym na przestrzeni $(\Omega, \mathfrak{F}, \mathbf{P})$, gdzie $\mathbf{X}_i : \Omega \rightarrow \mathbb{X}$ zaś alfabet \mathbb{X} jest przeliczalny. Oznaczmy bloki $\mathbf{X}_{m:n} := (\mathbf{X}_i)_{m \leq k \leq n}$.

Proces $(\mathbf{X}_i)_{i \in \mathbb{Z}}$ nazywam **mocno nieergodycznym**, jeżeli istnieją zmienne $(\mathbf{Z}_k)_{k \in \mathbb{N}} \sim \text{IID}$, $\mathbf{P}(\mathbf{Z}_k = \mathbf{0}) = \mathbf{P}(\mathbf{Z}_k = \mathbf{1}) = \frac{1}{2}$, i funkcje $\mathbf{s}_k : \mathbb{X}^* \rightarrow \{\mathbf{0}, \mathbf{1}\}$, $\mathbf{k} \in \mathbb{N}$, takie, że

$$\lim_{n \rightarrow \infty} \mathbf{P}(\mathbf{s}_k(\mathbf{X}_{t+1:t+n}) = \mathbf{Z}_k) = 1, \quad \forall t \in \mathbb{Z}.$$

$\mathbf{Y} = \sum_{k \in \mathbb{N}} 2^{-k} \mathbf{Z}_k$ jest mierzalna wzgl. σ -ciała niezmienniczego.

Proces $(\mathbf{X}_i)_{i \in \mathbb{Z}}$ nazywam **o skończonej energii**, jeżeli

$$\mathbf{P}(\mathbf{X}_{t+1:t+m} = \mathbf{u} | \mathbf{X}_{t-n:t} = \mathbf{w}) \leq \mathbf{K}c^m, \quad \forall t \in \mathbb{Z}.$$

Główny wynik

$$\mathbf{U}_\delta(\mathbf{n}) := \{\mathbf{k} \in \mathbb{N} : \mathbf{P}(s_{\mathbf{k}}(\mathbf{X}_{1:n}) = \mathbf{Z}_{\mathbf{k}}) \geq \delta\}.$$

Twierdzenie 1

Niech $(\mathbf{X}_i)_{i \in \mathbb{Z}}$ będzie stacjonarnym procesem mocno nieergod. o skończ. energii nad skończ. alfabetem \mathbb{X} . Przypuśćmy, że

$$\liminf_{n \rightarrow \infty} \frac{\text{card } \mathbf{U}_\delta(\mathbf{n})}{n^\beta} > 0$$

dla pewnego $\beta \in (0, 1)$ i $\delta \in (\frac{1}{2}, 1)$. Wówczas

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

dla każdej dopuszczalnie minimalnej transformacji gramatykowej Γ .

Pierwsze skojarzone stwierdzenie

Oznaczmy informację wzajemną między n -blokami

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

Twierdzenie 2

Niech $(\mathbf{X}_i)_{i \in \mathbb{Z}}$ będzie stacjonarnym procesem mocno nieergodycznym nad skończonym alfabetem \mathbb{X} . Przypuśćmy, że

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

dla pewnego $\beta \in (0, 1)$ i $\delta \in (\frac{1}{2}, 1)$. Wówczas

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

Drugie skojarzone stwierdzenie

Twierdzenie 3

Niech $(\mathbf{X}_i)_{i \in \mathbb{Z}}$ będzie stacjonarnym procesem o skończonej energii nad skończonym alfabetem \mathbb{X} . Przypuśćmy, że

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

dla pewnego $\beta \in (0, 1)$. Wówczas

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

dla każdej dopuszczalnie minimalnej transformacji gramatycznej Γ .

Informacja wzajemna dla języka naturalnego

W oparciu o pomiary entropii warunkowej Shannona (1950) dla tekstów w języku angielskim, Hilberg (1990) sformułował hipotezę, że dla odpowiedniego procesu stochastycznego modelującego teksty informacja wzajemna między n -blokami spełnia

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

- Twierdzenie 2 – racjonalna motywacja hipotezy Hilberga.
- Twierdzenie 3 – z hipotezy Hilberga wynika prawo Herdana.

Ponadto Twierdzenie 1 wskazuje,

- dlaczego prawo Herdana można zaobserwować dla tłumaczenia tego samego tekstu na różne języki,
- dlaczego wykładnik w prawie Herdana może do pewnego stopnia zależeć od tekstu.

- 1 Sformułowanie problemu
- 2 Zarys dowodu
- 3 Przykłady procesów
- 4 Podsumowanie

Entropia, pseudoentropia, długość kodu

Określmy entropię n -bloku oraz intensywność entropii

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

Oznaczmy także „pseudoentropię”

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

Niech $\mathbb{X} = \{0, 1, \dots, D - 1\}$ zaś $\mathbf{C} = \mathbf{B}(\Gamma(\cdot))$ będzie dopuszczalnie minimalnym kodem. Oznaczmy jego długość

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

Mamy nierówność

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

oraz równość intensywności

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

Dolne ograniczenie nadwyżki długości kodu $E^C(n)$

Niech funkcja $f : \mathbb{N} \rightarrow \mathbb{R}$ spełnia $\lim_k f(k)/k = 0$
 oraz $f(n) \geq 0$ dla wszystkich oprócz skończenie wielu n .
 Wówczas dla nieskończenie wielu n zachodzi $2f(n) - f(2n) \geq 0$.

Z równości i nierówności na poprzednim slajdzie otrzymujemy

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

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

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

dla $E(n) = 2H(n) - H(2n)$ oraz $E^C(n) = 2H^C(n) - H^C(2n)$.

Górne ograniczenie nadwyżki długości kodu $E^C(n)$

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

Dla dopuszczalnie minimalnego kodu $\mathbf{C} = \mathbf{B}(\Gamma(\cdot))$,

$$|C(\mathbf{u})| + |C(\mathbf{v})| - |C(\mathbf{w})| \leq \mathbf{W}_0 \mathbb{V}[\Gamma(\mathbf{w})](1 + \mathbb{L}(\mathbf{w})),$$

gdzie $\mathbf{w} = \mathbf{uv}$ dla $\mathbf{u}, \mathbf{v} \in \mathbb{X}^+$, $\mathbb{L}(\mathbf{w})$ oznacza maksymalną długość powtórzenia w napisie \mathbf{w} , zaś $\mathbf{W}_0 = |\mathbf{B}_S(\mathbf{D} + \mathbf{1})|$.

Dla procesu o skończonej energii $(\mathbf{X}_i)_{i \in \mathbb{Z}}$,

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

- 1 Sformułowanie problemu
- 2 Zarys dowodu
- 3 Przykłady procesów
- 4 Podsumowanie

Binarny proces wymierny

Rozważmy rodzinę binarnych rozkładów IID

$$\mathbf{P}(\mathbf{X}_{1:n} = \mathbf{x}_{1:n} | \theta) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i}.$$

Skonstruujmy proces $(\mathbf{X}_i)_{i \in \mathbb{Z}}$ taki, że

$$\mathbf{P}(\mathbf{X}_{1:n} = \mathbf{x}_{1:n}) = \int_0^1 \mathbf{P}(\mathbf{X}_{1:n} = \mathbf{x}_{1:n} | \theta) \pi(\theta) d\theta$$

dla rozkładu a priori $\pi(\theta) > 0$. Dla $\mathbf{Y} = \lim_n n^{-1} \sum_{i=1}^n \mathbf{X}_i$ mamy

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

Proces $(\mathbf{X}_i)_{i \in \mathbb{Z}}$ jest **mocno nieergodyczny**, ponieważ \mathbf{Y} ma rozkład ciągły. Jednakże blok $\mathbf{X}_{1:n}$ jest warunkowo niezależny od $\mathbf{X}_{n+1:2n}$ względem sumy $\mathbf{S}_n := \sum_{i=1}^n \mathbf{X}_i$. Zatem

$$\mathbf{E}(n) = \mathbf{I}(\mathbf{X}_{1:n}; \mathbf{X}_{n+1:2n}) = \mathbf{I}(\mathbf{S}_n; \mathbf{X}_{n+1:2n}) \leq \mathbf{H}(\mathbf{S}_n) \leq \log(n + 1).$$

Proces, który wymyśliłem w Santa Fe Institute

Proces $(\mathbf{X}_i)_{i \in \mathbb{Z}}$ postaci

$$\mathbf{X}_i := (\mathbf{K}_i, \mathbf{Z}_{\mathbf{K}_i}),$$

gdzie $(\mathbf{K}_i)_{i \in \mathbb{Z}}$ i $(\mathbf{Z}_k)_{k \in \mathbb{N}}$ są niezależnymi procesami IID,

$$P(\mathbf{K}_i = k) = k^{-1/\beta} / \zeta(\beta^{-1}), \quad \beta \in (0, 1),$$

$$P(\mathbf{Z}_k = z) = \frac{1}{2}, \quad z \in \{0, 1\}.$$

Interpretacja lingwistyczna

Proces $(\mathbf{X}_i)_{i \in \mathbb{Z}}$ jest ciągiem losowych stwierdzeń **niesprzecznie** opisujących stan „wcześniej” wylosowanego obiektu $(\mathbf{Z}_k)_{k \in \mathbb{N}}$. $\mathbf{X}_i = (k, z)$ stwierdza, że k -ty bit $(\mathbf{Z}_k)_{k \in \mathbb{N}}$ ma wartość $\mathbf{Z}_k = z$.

- Mamy $\text{card } \mathbf{U}_\delta(\mathbf{n}) \geq \mathbf{A}n^\beta$.
- Niestety alfabet $\mathbb{X} = \mathbb{N} \times \{0, 1\}$ jest nieskończony.

Stacjonarne kodowanie tego procesu

Funkcję $f : \mathbb{X} \rightarrow \mathbb{Y}^+$ rozszerzamy do funkcji $f^{\mathbb{Z}} : \mathbb{X}^{\mathbb{Z}} \rightarrow \mathbb{Y}^{\mathbb{Z}}$,

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

Dla miary ν na $(\mathbb{Y}^{\mathbb{Z}}, \mathfrak{B}^{\mathbb{Z}})$ definiujemy średnią stacjonarną $\bar{\nu}$ jako

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

gdzie $T((y_i)_{i \in \mathbb{Z}}) := (y_{i+1})_{i \in \mathbb{Z}}$ jest przesunięciem.

Twierdzenie 4

Niech $\mu = \mathbf{P}((X_i)_{i \in \mathbb{Z}} \in \cdot)$ dla procesu z poprzedniego slajdu.

Weźmy $\mathbb{Y} = \{0, 1, 2\}$ oraz $f(k, z) := \mathbf{b}(k)z^2$, gdzie

$\mathbf{1b}(k) \in \{0, 1\}^+$ jest rozwinięciem binarnym liczby k . Proces o rozkładzie $\mu \circ (f^{\mathbb{Z}})^{-1}$ spełnia założenie Tw. 1 dla $\beta > 0.77$.

Proces mieszający

Proces $(\mathbf{X}_i)_{i \in \mathbb{Z}}$ postaci

$$\mathbf{X}_i := (\mathbf{K}_i, \mathbf{Z}_{i, \mathbf{K}_i}),$$

gdzie $(\mathbf{K}_i)_{i \in \mathbb{Z}}$ i $(\mathbf{Z}_{ik})_{i \in \mathbb{Z}}$, $\mathbf{k} \in \mathbb{N}$, są procesami niezależnymi,

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

zaś $(\mathbf{Z}_{ik})_{i \in \mathbb{Z}}$ są łańcuchami Markowa o rozkładzie

$$\mathbf{P}(\mathbf{Z}_{ik} = \mathbf{z}) = \frac{1}{2},$$

$$\mathbf{P}(\mathbf{Z}_{ik} = \mathbf{z} | \mathbf{Z}_{i-1, \mathbf{k}} = \mathbf{z}) = 1 - \mathbf{p}_{\mathbf{k}}.$$

Obiekt $(\mathbf{Z}_{ik})_{\mathbf{k} \in \mathbb{N}}$ opisywany w tekście $(\mathbf{X}_i)_{i \in \mathbb{Z}}$ jest funkcją czasu i .

- Mamy $\liminf_{n \rightarrow \infty} \mathbf{E}(n) / n^\beta > 0$ dla $\mathbf{p}_{\mathbf{k}} \leq \mathbf{P}(\mathbf{K}_i = \mathbf{k})$.
- Stacjonarne kodowanie tego procesu jest procesem ergodycznym i spełnia $\liminf_{n \rightarrow \infty} \mathbf{E}(n) / (n \log^{-1} n)^\beta > 0$.

- 1 Sformułowanie problemu
- 2 Zarys dowodu
- 3 Przykłady procesów
- 4 Podsumowanie

Czy możemy sprawdzić, które objaśnienie jest lepsze?

Mała przy maszynie do pisania

Prawa Zipfa i Herdana zaobserwujemy, jeżeli kolejne **litery** i **spacje** tekstu będziemy uzyskiwać wciskając klawisze **losowo**.

vs.

Nowe objaśnienie

Jeżeli **tekst** długości n opisuje $\geq n^\beta$ niezależnych **faktów** w sposób powtarzalny, to tekst ten zawiera $\geq n^\beta / \log n$ różnych **słów**.

Czy można dobrze oszacować informację wzajemną?

- 1 Wmocnić Twierdzenia 1, 2 i 3:
 - Rozpatrzyć procesy **asymptotycznie średnio stacjonarne** (AMS).
 - Wyprowadzić wzrost słownika **prawie na pewno**.
 - Zastąpić $\limsup_{n \rightarrow \infty}$ przez $\liminf_{n \rightarrow \infty}$.
- 2 Niech $\mathbf{C}(\mathbf{u})$ będzie najkrótszym program generującym \mathbf{u} .
 $\mathbf{E}^{\mathbf{C}}(\mathbf{n})$ jest **informacją algorytmiczną** między blokami.
 - Niech $(\omega_k)_{k \in \mathbb{N}}$ będzie **algorytmicznie losową** liczbą z $(0, 1)$.
Zauważmy, że $\mathbf{E}(\mathbf{n}) = \mathbf{0}$ ale $\mathbf{E}^{\mathbf{C}}(\mathbf{n}) \asymp \mathbf{n}^\beta$ dla $\mathbf{X}_i := (\mathbf{K}_i, \omega_{\mathbf{K}_i})$.
 - Czy za pomocą pewnych kodów **uniwersalnych** można rozróżnić **pewne** źródła AMS o małej vs. dużej $\mathbf{E}^{\mathbf{C}}(\mathbf{n})$?
 - Czy za pomocą słownika kodów **gramatykowych** można rozróżnić **pewne** źródła AMS o małej vs. dużej $\mathbf{E}^{\mathbf{C}}(\mathbf{n})$?
- 3 Czy istnieją **kody dopuszczalnie minimalne**, które są obliczalne w **czasie wielomianowym**? (Lub dostatecznie podobne kody?)
 - Niech $(\mathbf{X}_i)_{i \in \mathbb{Z}}$ będzie binarnym procesem IID. Wówczas
 $\mathbb{V}[\Gamma(\mathbf{X}_{1:n})] = \Omega\left(\sqrt{\frac{hn}{\log n}}\right)$ dla transformacji **nieredukowalnych**.

Moje prace

- Ł. Dębowski, (2010). *On the Vocabulary of Grammar-Based Codes and the Logical Consistency of Texts*.
<http://arxiv.org/abs/0810.3125> [v4]
- Ł. Dębowski, (2010). *Variable-Length Coding of Two-Sided Asymptotically Mean Stationary Measures*. Journal of Theoretical Probability, 23:237–256.
- Ł. Dębowski, (2006). *On Hilberg's law and its links with Guiraud's law*. Journal of Quantitative Linguistics, 13:81–109.
- 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)