Kwartalnik Językoznawczy 2011/1 (5)

Agnieszka Kułacka

Matematyczny model dla prawa Kryłowa

1. Wstęp

Prawo Kryłowa zostało odkryte przez George'a K. Zipfa w 1949 roku podczas jego badań nad słownikiem języka angielskiego. Prawo to odnosi się do polisemii i opisuje relację pomiędzy należącą do danego słownika liczbą leksemów y a liczbą ich znaczeń x. Zipf zauważył, że wraz ze wrostem liczby znaczeń x liczba leksemów posiadających x znaczeń maleje i wyraził tę zależność następującą funkcją:

$$y = \frac{C}{x^2}, \tag{1}$$

gdzie C jest pewną stałą[1]. Ten model funkcyjny był później krytykowany m.in. przez Annę Wierzbicką[2].

W 1967 Ferenc Papp przeanalizował 60 tysięcy leksemów ze słownika węgierskiego. Zauważył, że dane empiryczne mogą być przybliżone przez fukcję malejącą i zaproponował następujące równanie funkcyjne:

$$y = \frac{W}{2^x}, \tag{2}$$

gdzie W jest liczbą leksemów w jednojęzycznym słowniku, y jest liczbą leksemów o x znaczeniach. Nie można ocenić, czy dane empiryczne są dobrze przybliżone tym modelem, gdyż Papp nie zawarł żadnych danych w swojej pracy[3].

W 1982 Jurij U. Kryłow zgromadził dane z dwóch słowników rosyjskich. Zauważył, że dane te wykazują podobne rozkłady statystyczne i wysunął wniosek, że różnice wynikają z niejasnych metod wyróżniania znaczeń wyrazów, a następnie przyznał zaobserowanej tendecji status prawa. Zaproponował inny model teoretyczny:

$$p_x = \frac{1}{2^x} \cdot \frac{1}{1 - p_1}, \tag{3}$$

[1] J. Sambor, *Słowa i liczby. Zagadnienia językoznawstwa statystycznego*,Wrocław 1972; R. Hammerl, J. Sambor, *O statystycznych prawach językowych*, Warszawa 1993.

[2] J. Sambor, *op.cit.*, s. 64.

[3] *Ibidem*, s. 75.

gdzie p_x jest prawdopodobieństwem wystąpienia leksemu mającego x znaczeń w danym słowniku, p_1 jest prawdopodobieństwem wylosowania leksemu o jednym znaczniu[4]. Ten model okazał się później niezadowalający i jako rozkład teoretyczny nie przybliża dobrze rozkładu empirycznego[5].

Jadwiga Sambor w 1990 roku opublikowała wyniki badań nad prawem Kryłowa, które przeprowadziła na słownikach języków polskiego, rosyjskiego i angielskiego, losując próbki o ustalonej liczbie leksemów. Badaczka porównała wyniki swoich badań nad danymi empirycznymi z wynikami otrzymanymi przez Kryłowa i dostrzegła podobieństwa między nimi. Uznała, że różnice powstały z powodu zastosowania różnych procedur wyróżniania znaczeń i właśności danego języka, np. słownik języka angielskiego zawiera więcej leksemów o dużej liczbie znaczeń niż słowniki języków polskiego i rosyjskiego[6].

Badania nad prawem Kryłowa nadal trwają. Gabriel Altmann wyjaśnia konieczność zachodzenia prawa Kryłowa regułami Zipfa, które leżą u podstaw zachowań ludzkich i stosują się do zasady najmniejszego wysiłku: rozmówcy dążą przy kodowaniu i dekodowaniu infromacji językowej do zużytkowania najmniejszej ilości energii. Wyniki działania tych sił podczas rozmowy są procesemi unifikacjii i różnicowania znaczeń leksemów w języku[7]. Ulepszony model teoretyczny opisujący to prawo został zaproponowany przez Rolfa Hammerla w 1991, do której odnoszą się autorzy podręcznika *O statystycznych prawach językowych*[8]. Inne modele zamieszczone są w atrykule G. Wimmera i G. Altmanna[9], jednakże żaden z nich nie został w pełni zweryfikowany[10].

W artykule Agnieszki Kułackiej[11] zaprezentowane zostały własności modelu matematycznego opisującego prawo Kryłowa. Przedstawiono w nim również jednolitą procedurę weryfikacji tego prawa, a także znaleziono wystarczającą wielkość prób, by dane empiryczne zbiegały się do pewnej funkcji, i wskazano sposób, w jaki można je gromadzić. W niniejszym artykule zgromadzono wyniki pierwszej próby znalezienia teoretycznego modelu, czyli funkcji, która będzie dobrze przybliżać dane.

---

[4] J.K. Kryłow, *Eine Untersuchung Statistischer Gesetzmässigkeiten auf der paradigmatischen Ebene der lexik natürlicher Sprachen*, „Studies on Zipf's law" 1982, s. 250.

[5] R. Hammerl, J. Sambor, *op.cit.*, s. 124.

[6] *Ibidem*, s. 120–123.

[7] Por. G. Altmann, *Diversification processes*, w: *Quantitative Linguistics. An International Handbook*, red. R. Köhler, G. Altmann, R.G. Piotrowski, Berlin–New York 2005, s. 97–113.

[8] R. Hammerl, J. Sambor, *op.cit.*, s. 125.

[9] G. Wimmer, G. Altmann, *Unified derivation of some linguistic laws*, w: *Quantitative Linguistics...*, s. 791–807.

[10] Badania były przeprowadzone na małych próbkach i ograniczone do jednego języka.

[11] A. Kułacka, *Procedura weryfikacji prawa Kryłowa*, „LingVaria" 2 (8), 2009, s. 9–20.

2. Równania krzywych aproksymujących

Jednym ze sposobów znalezienia równania, które opisywałoby badane zmienne, jest rozpatrywanie równań krzywych aproksymujących. Należy nanieść na układ współrzędnych punkty (X_i, Y_i) odpowiadające zmiennym X i Y – w ten sposób otrzymujemy wykres punktowy. Następnie omawia się pewne cechy modelu teoretycznego, który przybliżałby te dane, tak jak zrobiono to w artykule Agnieszki Kułackiej[12]. Dla danego wykresu puktowego można dobrać krzywą gładką, która by dobrze przybliżała dane empiryczne i która nosi nazwę krzywej aproksymującej. Ta krzywa, która najlepiej przybliża dane, nosi nazwę „najlepszej aproksymacji danych". Istnieje wiele równań krzywych, które można zastosować: równania funkcji wielomianowych, eksponencjalnych, logarytmicznych itd.

Przyjrzyjmy się następującemu przykładowi. Dane zostały wyekscerpowane ze *Słownika języka polskiego PWN*[13]:

Tabela 1. Częstości leksemów o x znaczeniach. Leksemy zaczynające się na literę „u"

Liczba znaczeń – x	1	2	3	4	5	6	7	8	9	10	Suma
Częstość – F	941	385	113	46	20	7	1	4	1	1	1519

Punkty, których pierwszą współrzędną stanowi liczba znaczeń x, a drugą – ich częstość, zostały zaznaczone na wykresie (Rysunek 1.) i połączone łamaną. Jak moża zauważyć, tylko 99,5% danych zostało przeanalizowanych i przedstawionych na wykresie. Ta procedura będzie stosowana w trakcie weryfikacji prawa oraz jego modelu. Warto wspomnieć, że jest to prawo statystyczne i danych o nikłej frekwencji nie rozważa się w stosunku do pozostałych wielkości.

Jak wspomniano we Wstępie, rozważano już wiele modeli przybliżających dane, ale nie znalezniono modelu zadowalającego. Celem niniejszego artykułu jest tę lukę wypełnić.

3. Model teoretyczny

Model, który opisuje prawo Kryłowa, oparty będzie na następującym wzorze:

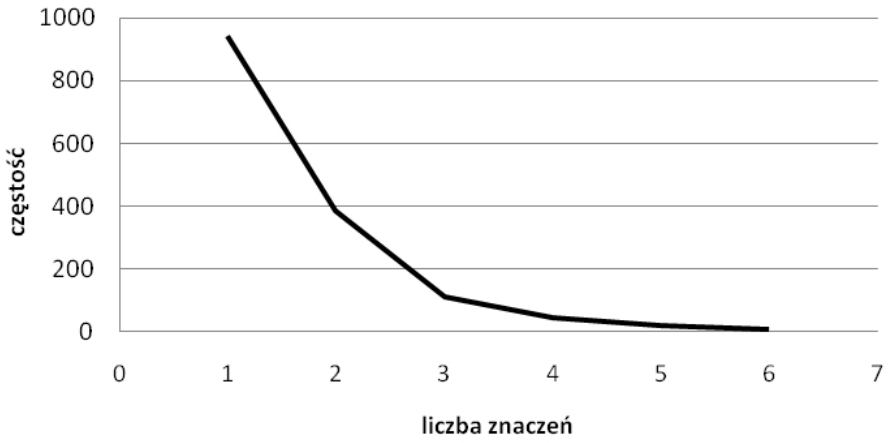
$$f(x) = \sum_{i=0}^{n} A_i x^i, \tag{4}$$

gdzie $n+1$ jest liczbą znaczeń leksemów, które zostały włączone do analizy, a x jest liczbą znaczeń leksemów wchodzących w skład zbioru o liczbie elementów zbióru $f(x)$. A_i są pewnymi współczynnikami.

[12] *Ibidem.*

[13] http://sjp.pwn.pl/

Rysunek 1. Wykres punktowy wraz z łamaną łączącą te punkty dla danych z Tabeli 1.



Dla danych w powyższym przykładzie będzie się poszukiwać funkcji, która przekształci:

$x = 1$ na $f(1) = 941$,
$x = 2$ na $f(2) = 345$,
$x = 3$ na $f(3) = 113$,
$x = 4$ na $f(4) = 46$,
$x = 5$ na $f(5) = 20$,
$x = 6$ na $f(6) = 7$.

Można pominąć wartości $x = 7, 8, 9, 10$, gdyż stanowią one mniej niż 0,5% wszystkich danych. Powodem ograniczenia zbioru danych do 99,5% liczby leksemów jest ujednolicenie procedury, która może być zastosowana do różnych słowników. Innym powodem jest uznanie 0,5% danych za te, które reprezentują elementy odstające.

Funkcję, której wykresem będzie krzywa aproksymująca, można znaleźć, rozwiązując układ równań:

(a) $941 = A_0 \times 1^0 + A_1 \times 1^1 + A_2 \times 1^2 + A_3 \times 1^3 + A_4 \times 1^4 + A_5 \times 1^5$,
(b) $345 = A_0 \times 2^0 + A_1 \times 2^1 + A_2 \times 2^2 + A_3 \times 2^3 + A_4 \times 2^4 + A_5 \times 2^5$,
(c) $113 = A_0 \times 3^0 + A_1 \times 3^1 + A_2 \times 3^2 + A_3 \times 3^3 + A_4 \times 3^4 + A_5 \times 3^5$,
(d) $46 = A_0 \times 4^0 + A_1 \times 4^1 + A_2 \times 4^2 + A_3 \times 4^3 + A_4 \times 4^4 + A_5 \times 4^5$,
(e) $20 = A_0 \times 5^0 + A_1 \times 5^1 + A_2 \times 5^2 + A_3 \times 5^3 + A_4 \times 5^4 + A_5 \times 5^5$,
(f) $7 = A_0 \times 6^0 + A_1 \times 6^1 + A_2 \times 6^2 + A_3 \times 6^3 + A_4 \times 6^4 + A_5 \times 6^5$.

Można je łatwo rozwiązać za pomocą metody macierzowej:

$$A = MX, \text{gdzie}$$

$$A = \begin{pmatrix} 941 \\ 34 \\ 113 \\ 46 \\ 20 \\ 7 \end{pmatrix}, \; M = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 & 16 & 32 \\ 1 & 3 & 9 & 27 & 81 & 243 \\ 1 & 4 & 16 & 64 & 256 & 1024 \\ 1 & 5 & 25 & 125 & 625 & 3125 \\ 1 & 6 & 36 & 216 & 1296 & 7776 \end{pmatrix} \text{ i } X = \begin{pmatrix} A_0 \\ A_1 \\ A_2 \\ A_3 \\ A_4 \\ A_5 \\ A_6 \end{pmatrix} \tag{5}$$

Rozwiązaniem powyższego równania jest:

$$X = \begin{pmatrix} 2154 \\ -1615 \\ 451 \\ -49.54 \\ 0.5 \\ 0.175 \end{pmatrix}. \tag{6}$$

Liczby będące rozwiązaniem powyższego układu równań podane są z dokładnością do 4 cyfr znaczących. Podstawiając je do równania funkcji, otrzymujemy:

$$f(x) = 2154 - 1615x + 451x^2 - 49.54x^3 + 0.5x^4 + 0.175x^5. \tag{7}$$

Ten wzór funkcyjny jest zależny od liczby badanych leksemów. Dlatego też, by porównać dane pochodzące z różnych słowników i różnych próbek, należy go ulepszyć poprzez podzielenie obu stron równania (7) przez całkowitą częstość, otrzymując nową funkcję – *g*:

$$g(x) = \frac{f(x)}{\textit{całkowita częstość}} \tag{8}$$

W przypadku danych zaprezentowanych wyżej częstość całkowita wynosi 1519, a więc nowa fukncja ma następujący wzór (liczby podane są z dokładnością do 5 cyfr znaczących):

$$g(x) = 1.4180 - 1.0632x + 0.29691x^2 - 0.032614x^3 + 0.00032916x^4 + 0.00011521x^5. \tag{9}$$

Przed omówieniem wartości współczynników należy sprawdzić, czy dane teoretyczne (częstości oczekiwane) dobrze przybliżają dane empiryczne (zaobserwowane częstości). Porównano częstości oczekiwane i zaobserwowane dla każdej wartości x (liczby znaczeń lesemów), stosując test chi-kwadrat

polegający na porównaniu częstości empirycznych i danych teoretycznych wyliczonych za pomocą wzoru funkcji f. Ponadto jedyny powód, dla którego funkcja g będzie preferowana, to możliwość porównania danych z różnych słowników.

Tabela 2. Częstości zaobserwowane i oczekiwane dla wyrazów o x znaczeniach

x	Częstość obserwowana	Częstość oczekiwana
1	941	941.14
2	345	345.28
3	113	113.45
4	46	46.64
5	20	20.875
6	7	8.16

Dla danych w Tabeli 2. $\chi^2 = 0.2124$ z $v = 5$ stopniom swobody i jest mniejszy niż $\chi^2_{0.95} = 11.1$, co oznacza, że model dobrze przybliża zaobserwowane częstości. Może wydawać się, że zbyt dobrze przybliża dane, jednakże warto zauważyć, że małą wartość χ^2 zawdzięczamy znalezieniu współczynników rozwiązując układ równań. Wskazuje też na to, że liczba cyfr znaczących jest wystarczająca, by dobrze przybliżać dane.

4. Słowniki języka polskiego

Stosując procedurę opisaną we wspomnianym artykule[14], zweryfikowano prawo Kryłowa na materiale uzyskanym z następujących słowników: (1) *Słownik języka polskiego*, pod red. W. Doroszewskiego, t. 6, Warszawa 1964, z 120 tysiącami leksemów, (2) *Słownik języka polskiego*, pod red. M. Szymczaka, t. 1, Warszawa 1978, z 80 tysiącami leksemów i (3) *Słownik współczesnego języka polskiego*, pod red. B. Dunaja, Warszawa 2000, z 62 tysiącami leksemów. Dane surowe zebrane są w Tabelach 3–5.

Tabela 3. Częstości leksemów o x znaczeniach w słowniku Doroszewskiego

x	1	2	3	4	5	6	7	8	9	10	14	15	16	19	Suma
F	14486	2121	570	197	87	35	23	6	5	7	1	1	1	1	17541

Tabela 4. Częstości leksemów o x znaczeniach w słowniku Szymczaka

x	1	2	3	4	5	6	7	8	9	10	11	12	Suma
F	14183	2180	557	193	73	35	12	8	3	4	3	2	17253

Tabela 5. Częstości leksemów o x znaczeniach w słowniku Dunaja

x	1	2	3	4	5	6	7	8	9	10	11	12	13	15	16
F	14606	2975	553	268	104	49	26	17	7	3	4	4	2	1	1

[14] *Ibidem.*

(cd. Tab. 5.)

x	18	19	26	Suma
F	1	1	1	18719

Punkty, których pierwsza współrzędna to liczba znaczeń, a druga to liczba leksemów o danej liczbie znaczeń, układają się na krzywej ściśle malejącej, jeśli weźmie się pod uwagę 99.5%. A to oznacza, że prawo zachodzi w analizowanym materiale słownikowym.

Dla każdego z badanych słowników obliczono wartości współczynników funkcji $g(x)$ w sposób opisany w części 3. niniejszego artykułu. Dla każdego ze słowników analizowano 99.5% danych, czyli leksemy o co najmniej 6 znaczeniach:

– dla słownika Doroszewskiego

$$g_1(x) = 3.6375 - 4.6965x + 2.4285x^2 - 0.61941x^3 + 0.077442x^4 - 0.0037878x^5 \tag{10}$$

– dla słownika Szymczaka

$$g_2(x) = 3.5383 - 4.5061x + 2.3029x^2 - 0.58165x^3 + 0.072125x^4 - 0.0035033x^5 \tag{11}$$

– dla słownika Dunaja

$$g_3(x) = 2.7045 - 2.9869x + 1.3234x^2 - 0.29100x^3 + 0.031635x^4 - 0.0013587x^5. \tag{12}$$

Następnie porównano wielkości zaobserwowane i oczekiwane dla każdego ze słowników, stosując test chi-kwadrat.

Tabela 6. Częstości zaobserwowane i oczekiwane dla wyrazów o x znaczeniach w słowniku Doroszewskiego

x	Częstość obserwowana	Częstość oczekiwana
1	14448	14449
2	2121	2123.9
3	570	575.29
4	197	205.31
5	87	98.668
6	35	49.802

Dla danych w Tabeli 6. $\chi^2 = 6.1682$ z $v = 5$ stopniom swobody i jest mniejszy niż $\chi^2_{0.95} = 11.1$, co oznacza, że model dobrze przybliża zaobserwowane częstości.

Tabela 7. Częstości zaobserwowane i oczekiwane dla wyrazów o x znaczeniach w słowniku Szymczaka

x	Częstość obserwowana	Częstość oczekiwana
1	14183	14183
2	2180	2180.7
3	557	578.53
4	193	195.32
5	73	75.266
6	35	35.182

Dla danych w Tabeli 7. $\chi^2 = 0.09875$ z $v = 5$ stopniom swobody i jest mniejszy niż $\chi^2_{0.95} = 11.1$, co oznacza, że model dobrze przybliża zaobserwowane częstości.

Tabela 8. Częstości zaobserwowane i oczekiwane dla wyrazów o x znaczeniach w słowniku Dunaja

x	Częstość obserwowana	Częstość oczekiwana
1	14606	14606
2	2975	2976
3	553	555.41
4	268	272.38
5	104	111.14
6	49	60.066

Dla danych w Tabeli 8. $\chi^2 = 2.5786$ z $v = 5$ stopniom swobody i jest mniejszy niż $\chi^2_{0.95} = 11.1$, co oznacza, że model dobrze przybliża zaobserwowane częstości.

5. Porównanie współczynników

Współczynniki przy potęgach x we wzorach funkcji (10) i (11) mają zbliżone wartości, ale ich podobieństwo ze współczynnikami trzeciej funkcji leży jedynie w rozmiarach liczb i ich znakach. Jednakże, jeśli porównamy współczynniki w funkcjach g, okażą się one podobne. Należy zauważyć, że dla pierwszych dwóch funkcji 99.5% danych oznacza, że analizujemy tylko leksemy o co najwyżej 5 znaczeniach:

– dla słownika Doroszewskiego

$$g_4(x) = 3.1829 - 3.6587x + 1.57625x^2 - 0.29745x^3 + 0.020625x^4 \qquad (13)$$

– dla słownika Szymczaka

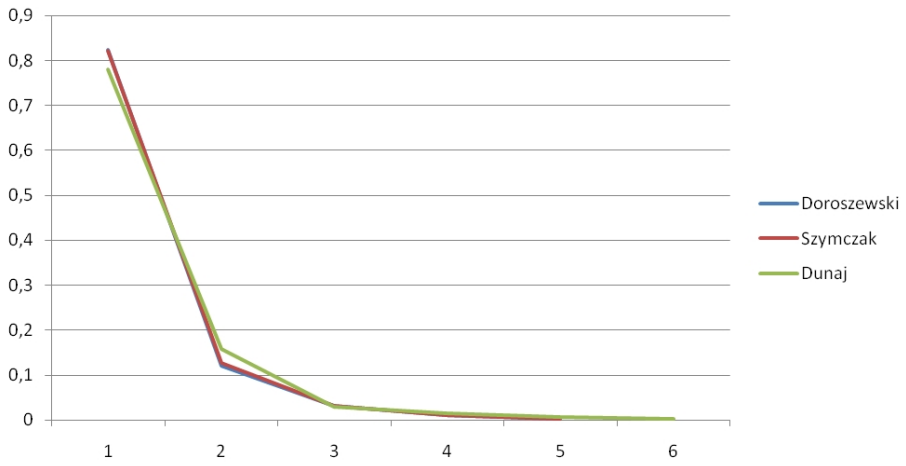
$$g_5(x) = 3.1179 - 3.5462x + 1.5146x^2 - 0.28387x^3 + 0.019576x^4 \qquad (14)$$

– dla słownika Dunaja

$$g_6(x) = 2.7045 - 2.9869x + 1.3234x^2 - 0.29100x^3 + 0.031635x^4 - 0.0013587x^5. \tag{15}$$

Oznacza to, że rozkłady statystyczne leksemów o x znaczeniach są w badanych słownikach zbliżone. Łamane na wykresie (Rysunek 2.) otrzymano przez połączenie punktów, których pierwsza współrzędna to liczba znaczeń, a druga to częstość względna leksemów o danej liczbie znaczeń, otrzymana ze wzorów (13)–(15). Jak można zauważyć, wykresy nakładają się na siebie i ich kształty są podobne.

Rysunek 2. Łamane dla funkcji danych równaniami (11)–(13)
(x – liczba znaczeń, y – częstość względna)



6. Dalsze badania

Dalsze badania mogą podążać dwiema drogami: jedną z nich jest poprawienie modelu i znalezienie fukncji niebędącej wielomianem, służącej do przybliżania danych; inną drogą będzie porównanie wielkości analizowanego typu uzyskanych na podstawie danych pochodzących ze słowników różnych języków w celu ustalenia cech języka i ich związków ze współczynnikami znalezionymi dla danego słownika.

Mathematical Model for the Krylov Law

SUMMARY

This article presents a mathematical model for Krylov's law, which fits well the empirical data. The data from a set of Polish dictionaries has been collected and the coefficients in the formula based on the mathematical model have been compared.

O Autorce

Agnieszka Kułacka - King’s College London, Birkbeck University.
Absolwentka matematyki i filologii angielskiej UWr.
Doktorat: "Statystyczne prawa językowe. Na przykładzie prawa Menzeratha-Altmanna". Obecnie pisze drugi doktorat na Wydziale Filozofii King’s College London w zakresie językoznawstwa teoretyczno-matematycznego.
Uczy matematyki, statystyki i mechaniki w jednym z londyńskich liceów, matematyki na Wydziale Ekonomii Birkbeck University oraz filozofii języka, rachunku lambda i logiki na Wydziale Filozofii King’s College London.
Zainteresowania: językoznawstwo statystyczne, filozofia języka, semantyka formalna, logika. W tych dziedzinach też publikuje.
E-mail: agnieszka.kulacka@kcl.ac.uk; a.kulacka@gmail.com