ZASTOSOWANIA MATEMATYKI
VII (1964)

A. BARTKOWIAKOWA i B. GLEICHGEWICHT (Wrocław)

ZASTOSOWANIE DWUPARAMETROWYCH ROZKŁADÓW FUCKSA DO OPISU DŁUGOŚCI SYLABICZNEJ WYRAZÓW W RÓŻNYCH UTWORACH PROZAICZNYCH AUTORÓW POLSKICH

1. W swych pracach [1] i [2] autorzy przeprowadzili badania dotyczące długości sylabicznej wyrazów w tekstach autorów polskich z XVI-XX w., analogiczne do badań W. Fucksa ([3], [4]). W wyniku, otrzymano między innymi dane pozwalające porównać strukturę sylabiczną języka polskiego z odpowiednią strukturą dziewięciu innych języków oraz dane dotyczące rozwoju tej struktury na przestrzeni wieków. W cytowanych wyżej pracach, przy badaniu rozkładu długości sylabicznej wyrazów, autorzy posługiwali się wzorami zaproponowanymi przez W. Fucksa (p. [4]):

$$(1)\qquad \begin{aligned} v(i) &= e^{-[\bar{i}-(1+\beta)]}\left\{(1-\beta)\frac{[\bar{i}-(1+\beta)]^{i-1}}{(i-1)!}+\beta\frac{[\bar{i}-(1+\beta)]^{i-2}}{(i-2)!}\right\} \\ v(1) &= e^{-[\bar{i}-(1+\beta)]}(1-\beta), \end{aligned} \qquad \text{dla } i \geqslant 2,$$

gdzie i wyraża liczbę sylab w słowie, $\bar{i}$ — oczekiwaną liczbę sylab w słowie, $v(i)$ jest prawdopodobieństwem tego, że dane słowo liczyć będzie i sylab, a β jest parametrem rozkładu Fucksa.

Zwróciliśmy wówczas uwagę na to, że wzór (1) nie opisuje zadowalająco omawianych rozkładów, co można było tłumaczyć tym, że wzór ten był tylko dość grubym przybliżeniem wzoru

$$(2)\qquad v(i) = e^{-(\bar{i}-\sum_{a=1}^{\infty}\beta_a)}\sum_{\nu=0}^{i}(\beta_\nu-\beta_{\nu+1})\frac{\left(\bar{i}-\sum_{a=1}^{\infty}\beta_a\right)^{i-\nu}}{(i-\nu)!},$$

gdzie β_a były dalszymi parametrami rozkładu. W rozpatrzonym wówczas przypadku *Placówki* Prusa rozkład otrzymany na podstawie wzoru (1) był na oko zgodny z empirycznym, jednak porównanie obu za pomocą testu χ^2 wykazało istotną różnicę między nimi:

$$\Pr\{\chi^2 \geqslant 15{,}00\} = 0{,}0006$$

przy dwóch stopniach swobody. Zjawisko to można było tłumaczyć również *a priori* tym, że elementy naszej próbki były wyrazami tekstu kolejno następującymi po sobie, nie było więc spełnione założenie o niezależności elementów w pobieranych próbkach. Zapowiedziano wówczas głębsze badania związane z tym zagadnieniem, polegające na porównaniu rozkładu empirycznego z ogólniejszym rozkładem Fucksa oraz zbadanie zagadnienia, czy parametry rozkładu zmieniają się w zależności od epoki czy też są konstantami językowymi.

W toku dalszych badań okazało się, że rozkłady empiryczne otrzymane dla innych utworów omawianych w cytowanych pracach autorów również różnią się istotnie od rozkładów wynikających ze wzorów (1). Dlatego też rozkładów tych nie zamieszczamy.

W niniejszej pracy rozpatrujemy przybliżenia uzyskanych wówczas empirycznie rozkładów za pomocą rozkładu typu (2), w którym przyjęto $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 \neq 0$, $\beta_3 \neq 0$, $\beta_\nu = 0$ dla $\nu > 3$. Okazuje się, że tego rodzaju rozkład opisuje zupełnie zadowalająco rozkłady empiryczne z wyjątkiem jednego utworu, mianowicie *Placówki* Prusa (z dialogami).

2. Zajmijmy się obecnie wyjaśnieniem struktury rozkładu użytego przez nas w tej pracy. Według proponowanego modelu słowo składa się z dwóch rodzajów sylab: jedne z nich tworzą tzw. „zaczątek" słowny, drugie „końcówkę" (w tym miejscu nie wypowiadamy się na temat tego, jaką interpretację lingwistyczną należy podkładać pod terminy „zaczątek" i „końcówka"). Liczba sylab w „końcówce" jest zmienną losową, która ma rozkład Poissona, przyjmuje więc wartości całkowite nieujemne, $k = 0, 1, 2\ldots$, z prawdopodobieństwami

$$P_k = \frac{\lambda^k}{k!} e^{-\lambda},$$

gdzie λ jest wartością średnią rozkładu, taką samą dla wszystkich „końcówek" słownych danego języka. „Zaczątki" mogą być jednosylabowe, dwusylabowe i trójsylabowe. Mając na uwadze pewną populację słów (język), przyjmujemy, że frakcja słów o „zaczątkach" jednosylabowych wynosi $1-\beta_2$; frakcja słów o „zaczątkach" dwusylabowych $\beta_2-\beta_3$; o zaczątkach trójsylabowych β_3.

Jako wniosek z tych założeń otrzymamy szczególny przypadek wzoru (2) o następującej postaci:

$$v(i) = (1-\beta_2)P_{i-1} + (\beta_2-\beta_3)P_{i-2} + \beta_3 P_{i-3}, \tag{3}$$

gdzie P_{i-1}, P_{i-2} i P_{i-3} oznaczają prawdopodobieństwa rozkładu Poissona o średniej $\lambda = \bar{i} - (1+\beta_2+\beta_3)$, przy czym należy rozumieć, że

$$v(1) = (1-\beta_2)P_0, \qquad v(2) = (1-\beta_2)P_1 + (\beta_2-\beta_3)P_0.$$

W. Fucks podał wzory na momenty tego rodzaju rozkładów ([4]). W naszym wypadku wyrażają się one, jak łatwo sprawdzić, następująco:

$$M_2 = E[(x-\bar{x})^2] = \bar{x}-(1+\beta_2+\beta_3)^2+2(\beta_2+2\beta_3),$$

$$M_3 = E[(x-\bar{x})^3] = \bar{x}-3(1+\beta_2+\beta_3)^2+2(1+\beta_2+\beta_3)^3-$$
$$-6(\beta_2+\beta_3)(\beta_2+2\beta_3)+6\beta_3.$$

Obliczanie na podstawie tych wzorów wartości β_2 i β_3 jest bardzo żmudne; jest ono tym bardziej niecelowe, że metoda momentów nie zawsze daje dobre wyniki. Wobec tego przyjęto tu inną metodę estymacji β_2 i β_3 polegającą na dopasowywaniu rozkładu do frakcji słów jednosylabowych, która na ogół była w badanych utworach najliczniejsza. Ze wzoru na $v(1)$ otrzymujemy

$$\log_{10}\frac{v(1)}{1-\beta_2} = -[\bar{i}-(1+\beta_2+\beta_3)]\log_{10}e,$$

skąd, przy ustalonym β_2, wyznaczamy tak β_3, by dokładnie zachodziła napisana równość. Dla otrzymanych w ten sposób parametrów β_2 i β_3 obliczamy rozkład (teoretyczny) i porównujemy go z rozkładem empirycznym obliczając wielkość χ^2. Czynność tę powtarzamy zmieniając β_2 aż do uzyskania przybliżonego minimum χ^2. Uzyskane wyniki przedstawiono w załączonej tablicy, w której n_i oznacza empiryczny n_i' zaś teoretyczny — obliczony według wzoru (3) — rozkład liczby sylab w słowie każdego z podanych tam utworów. Pod każdym rozkładem podano liczność próbki, wartość średnią rozkładu oraz wartości parametrów β_2 i β_3 uzyskane w sposób opisany wyżej przy minimalnym χ^2. Podano również wartości χ^2 i prawdopodobieństwa uzyskania nie mniejszych wielkości χ^2 przy założeniu, że odchylenia są przypadkowe. Liczbę stopni swobody dla χ^2 uzyskano odejmując od liczby porównywanych klas liczbę 4 (ze względu na to, że wyznaczamy tu empirycznie średnią i dwa parametry). Zbadano w ten sposób wszystkie rozkłady rozpatrywane w pracach [1] i [2] i uzyskano zadowalające wyniki z wyjątkiem *Placówki* Prusa. W poprzednich pracach rozpatrywaliśmy rozkłady dla *Placówki* dwukrotnie: raz brano pod uwagę pełny tekst, a drugi raz ten sam tekst, lecz z wyłączeniem dialogów. W obecnych badaniach do tekstu *Placówki* bez dialogów udało się nam dobrać β_2 i β_3 w ten sposób, że otrzymano przybliżenie rozkładu empirycznego za pomocą wzoru (3) na poziomie istotności 0,006. Natomiast nie udało się uzyskać tego dla pełnego tekstu *Placówki* (tzn. łącznie z dialogami): prawdopodobieństwo przypadkowego uzyskania zaobserwowanej lub jeszcze większej wartości χ^2 wypadało tu dużo mniejsze od 0,001. Można było oczekiwać, że przyczyną tego stanu rzeczy jest niejednorodność tekstu składającego się z opisów i dialogów, wobec czego zbadano rozkład dla dialogów oddzielnie; i w tym wypadku nie udało się otrzymać zadowalającego przybliżenia rozkładu empirycznego. Można by stąd wyciągnąć wniosek, że dialogi w *Placówce* są zbudowane

w inny sposób niż to zakłada proponowany przez nas model. Trudno jest nam sądzić o zjawisku lingwistycznym leżącym u podłoża tego faktu, jest nim, być może, dopasowywanie dialogu *Placówki* do gwary chłopskiej, natomiast matematycznie może to oznaczać, że liczba dołączonych sylab w „końcówce" nie wyraża się rozkładem Poissona, lub też wyraża się rozkładem Poissona, lecz z różnymi wartościami średnimi dla różnych rodzajów słów.

Jest rzeczą ciekawą, że zjawiska tego nie zaobserwowaliśmy w przypadku *Anielki* Prusa: tu zarówno pełny tekst jak i oddzielnie tekst bez dialogów i dialogi dają się bardzo dobrze opisać wzorem (3).

Pewne wątpliwości mogą wzbudzać również obydwie pozycje Lema. Uzyskano tu przybliżenia na poziomie istotności 0,004 i 0,02, co leży już na granicy poziomu ufności. Nie wydaje się, żeby przez dokładniejsze wyznaczenie β_2 i β_3 można uzyskać dużo lepsze przybliżenie. W rozkładach tych zaobserwowano największe wartości średnie i największe frakcje zaczątków trójsylabowych. Może więc mała dokładność przybliżenia wywołana jest przez częstsze pojawianie się zaczątków czterosylabowych, których w modelu nie brano pod uwagę.

3. Jak wspomnieliśmy, ilość zaczątków jednosylabowych charakteryzuje wielkość $1-\beta_2$, dwusylabowych $\beta_2-\beta_3$, trójsylabowych β_3. Wielkości te są oczywiście charakterystyczne dla języka, w którym tekst został napisany, w danym przypadku dla polskiego. Można by się spodziewać, że parametry β dla języka z różnych epok lub różnych rodzajów literackich będą się różniły między sobą. Jest rzeczą interesującą, jak wielkim zmianom mogą ulegać te parametry, tzn. w jakim stopniu są one charakterystyką autora (rodzaju literackiego lub epoki), a w jakim stopniu są konstantami językowymi. Do wyjaśnienia tego zilustrowaliśmy na rysunku 1 wielkości $1-\beta_2$, $\beta_2-\beta_3$ i β_3 dla badanych utworów. Jak widać, β_2 jest wielkością mało zmieniającą się dla badanych utworów: $0{,}300 \leqslant \beta_2 \leqslant 0{,}390$, natomiast β_3 zmienia się od 0,005 do 0,179, a więc dość znacznie. Ciekawe byłoby porównanie tych parametrów z odpowiednimi parametrami w innych językach.

Pozostaje jeszcze sprawa dość zasadnicza: czy owym hipotetycznym „zaczątkom" i „końcówkom" można przypisać jakąś interpretację lingwistyczną. Mieliśmy nadzieję, że mogą nimi być pewne części wyrazu, znane już językoznawcom, np. morfemy główne lub tematy słowne oraz końcówki wyrazów w potocznym tego słowa znaczeniu, z dołączeniem być może również przedrostków. Jednak wstępne badania zdają się nie potwierdzać tego przypuszczenia. Zagadnieniem tym nie zajmował się W. Fucks w cytowanych pracach, a autorom nie są znane inne badania dotyczące tego zagadnienia[1].

[1] Przypisek w korekcie: Ostatnio dowiedzieliśmy się od prof. A. N. Kołmogorowa, że zagadnieniem tym zajmowano się w Moskwie.

TABLICA EMPIRYCZNYCH (n_i) I TEORETYCZNYCH (n'_i) ROZKŁADÓW LICZBY SYLAB W SŁOWIE

1. M. Rej, *Żywot...*

i	n_i	n'_i
1	1737	1737,00
2	1474	1485,60
3	638	607,08
4	120	144,12
5	26	23,08
6	5	3,12

$n = 4000$, $\bar{i} = 1{,}8100$
$\beta_2 = 0{,}345$, $\beta_3 = 0{,}054$
$\chi_0^2 = 7{,}22$
$\Pr\{\chi^2 > \chi_0^2\} \approx 0{,}02$
2 stopnie swobody

2. Ł. Górnicki, *Dworzanin*

i	n_i	n'_i
1	2504	2504,00
2	2336	2350,33
3	912	899,57
4	220	208,13
5	26	38,97
6	2	

$n = 6000$, $\bar{i} = 1{,}8223$
$\beta_2 = 0{,}300$, $\beta_3 = 0{,}005$
$\chi_0^2 = 4{,}054$
$\Pr\{\chi^2 > \chi_0^2\} \approx 0{,}02$
1 stopień swobody

3. B. Prus, *Placówka* (bez dialogów)

i	n_i	n'_i
1	1322	1322,00
2	1431	1415,35
3	622	662,60
4	217	188,15
5	38	43,90
6	2	

$n = 3632$, $\bar{i} = 1{,}9603$
$\beta_2 = 0{,}360$, $\beta_3 = 0{,}036$
$\chi_0^2 = 7{,}431$
$\Pr\{\chi^2 > \chi_0^2\} \approx 0{,}006$
1 stopień swobody

4. B. Prus, *Anielka* (dialogi)

i	n_i	n'_i
1	710	710,00
2	645	632,76
3	299	319,87
4	103	98,85
5	26	24,52
6	3	

$n = 1782$, $\bar{i} = 1{,}9356$
$\beta_2 = 320$, $\beta_3 = 0{,}079$
$\chi_0^2 = 2{,}591$
$\Pr\{\chi^2 > \chi_0^2\} \approx 0{,}10$
1 stopień swobody

5. B. Prus, *Anielka* (z dialogami)

i	n_i	n'_i
1	2500	2500,00
2	2384	2390,14
3	1481	1486,05
4	549	530,21
5	123	123,61
6	17	23,98

$n = 7054$, $\bar{i} = 2{,}0732$
$\beta_2 = 0{,}380$, $\beta_3 = 0{,}134$
$\chi_0^2 = 2{,}443$
$\Pr\{\chi^2 > \chi_0^2\} \approx 0{,}30$
2 stopnie swobody

6. B. Prus, *Anielka* (bez dialogów)

i	n_i	n'_i
1	1790	1790,00
2	1739	1766,33
3	1782	1152,60
4	446	431,81
5	97	105,59
6	14	21,65

$n = 5268$, $\bar{i} = 2{,}1198$
$\beta_2 = 0{,}390$, $\beta_3 = 0{,}145$
$\chi_0^2 = 5{,}041$
$\Pr\{\chi^2 > \chi_0^2\} \approx 0{,}08$
2 stopnie swobody

7. S. Żeromski, *Ludzie bezdomni*

i	n_i	n'_i
1	858	858,00
2	931	931,00
3	494	501,24
4	178	164,74
5	33	37,53
6	6	7,41

$n = 2500$, $\bar{i} = 2{,}0460$
$\beta_2 = 0{,}370$ $\beta_3 = 0{,}069$
$\chi_0^2 = 1{,}987$
$\Pr\{\chi^2 > \chi_0^2\} \approx 0{,}36$
2 stopnie swobody

8. S. Lem, *Astronauci*

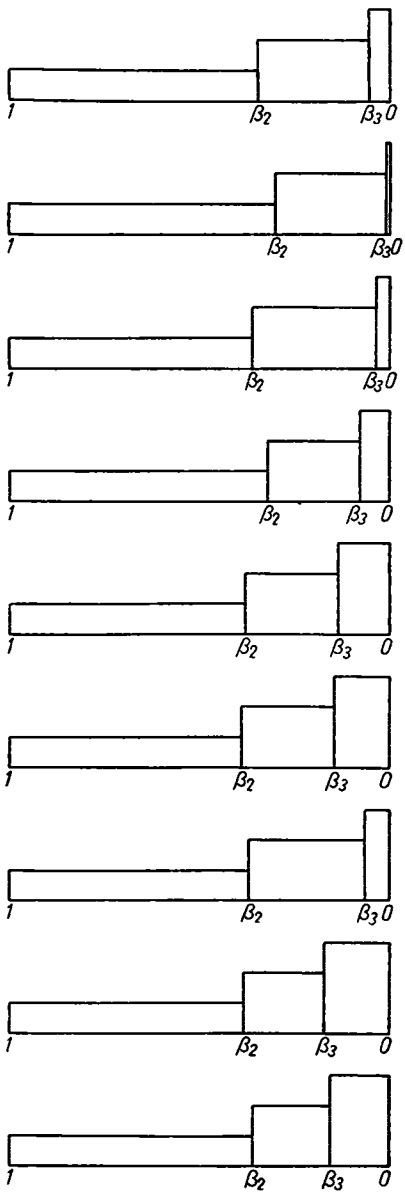
i	n_i	n'_i
1	1999	1999,00
2	1896	1889,53
3	1324	1364,92
4	623	562,85
5	128	149,57
6	27	34,33
7	3	

$n = 6000$, $\bar{i} = 2{,}1797$
$\beta_2 = 0{,}385$, $\beta_3 = 0{,}182$
$\chi_0^2 = 10{,}95$
$\Pr\{\chi^2 > \chi_0^2\} \approx 0{,}004$
2 stopnie swobody

9. S. Lem, *Obłok Magellana*

i	n_i	n'_i
1	1075	1075,00
2	1057	1037,22
3	668	710,00
4	322	290,98
5	68	78,62
6	14	18,18
7	4	
8	2	

$n = 3210$, $\bar{i} = 2{,}1648$
$\beta_2 = 360$, $\beta_3 = 0{,}157$
$\chi_0^2 = 7{,}784$
$\Pr\{\chi^2 > \chi_0^2\} \approx 0{,}02$
2 stopnie swobody



Rys. 1. Frakcje „zaczątków" słów

Prace cytowane

[1] A. Bartkowiakowa i B. Gleichgewicht, *O długości sylabicznej wyrazów w tekstach autorów polskich*, Zastosow. Mat. 6 (1962), str. 309-319.

[2] — *O długości sylabicznej wyrazów w różnych tekstach*, Pamiętnik Literacki (w druku).

[3] W. Fucks, *Mathematische Analyse von Sprachelementen, Sprachstil und Sprachen*, Köln und Opladen 1955.

[4] — *Mathematical theory of word-formation*, Physics Institute of the University of Technology, Aachen 1955 (cytowane na podstawie *Теория передачи сообщений*, Труды третьей международной конференции, Сборник статьей, Москва 1957.

INSTYTUT MATEMATYCZNY POLSKIEJ AKADEMII NAUK
INSTYTUT MATEMATYCZNY UNIWERSYTETU WROCŁAWSKIEGO

Praca wpłynęła 22. 3. 1963

А. БАРТКОВЯКОВА и Б. ГЛЕЙХГЕВИХТ (Вроцлав)

ПРИМЕНЕНИЕ ДВУПАРАМЕТРИЧЕСКИХ РАСПРЕДЕЛЕНИЙ ФУКСА ДЛЯ ОПИСАНИЯ ЧИСЛА СЛОГОВ В СЛОВАХ РАЗЛИЧНИХ ПРОИЗВЕДЕНИЙ ПРОЗЫ ПОЛЬСКИХ АВТОРОВ

РЕЗЮМЕ

Предполагая, что каждое слово содержит по крайней мере один слог, исследовано согласованность распределений числа слогов в слове с распределением следующего вида:

$$v(i) = \begin{cases} (1-\beta_2)\dfrac{\lambda^{i-1}}{(i-1)!}e^{-\lambda} + (\beta_2-\beta_3)\dfrac{\lambda^{i-2}}{(i-2)!}e^{-\lambda} + \beta_3\dfrac{\lambda^{i-3}}{(i-3)!}e^{-\lambda} & \text{при} \quad i \geqslant 3; \\ (1-\beta_2)\lambda e^{-\lambda} + (\beta_2-\beta_3)e^{-\lambda}, & \text{при} \quad i = 2; \\ (1-\beta_2)e^{-\lambda}, & \text{при} \quad i = 1. \end{cases}$$

В таблице (стр. 349) приведены эмпирические распределения, полученные в результате исследования текста выбранного из различных произведений. Рядом приведены подобраные теоретические распределения. Существенность различия исследовано тестом χ^2. В семи из девяти исследуемых случаев получено хорошую согласованность ($\Pr\{\chi^2 \geqslant \chi_0^2\} \geqslant 0{,}02$).

A. BARTKOWIAKOWA and B. GLEICHGEWICHT (Wrocław)

APPLICATION OF TWO-PARAMETER FUCKS' DISTRIBUTIONS TO THE STUDY OF SYLLABIC LENGTH OF WORDS FOR VARIOUS POLISH NOVELISTS

SUMMARY

On the assumption that each word contains at least one syllable, the authors compared the distribution of the syllabic length of words with the following distribution

$$v(i) = \begin{cases} (1-\beta_2)\dfrac{\lambda^{i-1}}{(i-1)!}e^{-\lambda} + (\beta_2-\beta_3)\dfrac{\lambda^{i-2}}{(i-2)!}e^{-\lambda} + \beta_3\dfrac{\lambda^{i-3}}{(i-3)!}e^{-\lambda} & \text{for} \quad i > 3; \\ (1-\beta_2)\lambda e^{-\lambda} + (\beta_2-\beta_3)e^{-\lambda} & \text{for} \quad i = 2; \\ (1-\beta_2)e^{-\lambda} & \text{for} \quad i = 1. \end{cases}$$

A table (p. 349) gives the empirical distributions obtained from samples of prose of various authors. The same table gives also the fitted theoretical distributions. The significance of deviations has been tested with the χ^2 test; in seven out of nine cases studied these deviations were found not significant ($\Pr\{\chi^2 > \chi_0^2\} > 0{,}02$).