A. BARTKOWIAKOWA i B. GLEICHGEWICHT (Wrocław)

*O DŁUGOŚCI SYLABICZNEJ WYRAZÓW
W TEKSTACH AUTORÓW POLSKICH*

1. Ostatnimi czasy szerokiego rozmachu nabrały badania z zakresu lingwistyki matematycznej. Okazało się, że metody statystyki matematycznej i teorii informacji stanowią cenne narzędzie dla współczesnego językoznawcy, który za ich pomocą może uzyskiwać wyniki nieosiągalne dotychczasowymi metodami.

Ciekawą pozycję w tej dziedzinie stanowią badania W. Fucks'a nad strukturą sylabową języków ([1], [2]). W pracach tych autor zajmuje się długością sylabiczną wyrazów oraz podaje matematyczny opis budowy sylabowej słów. Badania te, mające szeroki zakres, były prowadzone na materiale dziewięciu języków: angielskiego, rosyjskiego, niemieckiego, esperanto, arabskiego, greckiego, japońskiego, łaciny i tureckiego. W pracach tych W. Fucks dochodzi do bardzo ciekawych wniosków, wykazujących pewne prawidłowości struktury rozważanych języków.

Pewne badania dotyczące długości sylabicznej wyrazów w tekstach autorów polskich były prowadzone przez M. R. Mayenową [4]. Publikacja jej zawiera znaczny materiał empiryczny, pochodzący z różnych okresów rozwoju literatury polskiej, bez dokładniejszego opracowania matematycznego uzyskanych wyników.

W niniejszej pracy pragniemy przeprowadzić badania długości sylabicznej wyrazów języka polskiego, analogiczne do cytowanych prac W. Fucks'a, co pozwoli nam na porównanie struktury sylabowej języka polskiego z odpowiednią strukturą innych języków. W pracy uwzględniono również zmiany długości sylabicznej wyrazów na przestrzeni czasu od Reja do utworów współczesnych.

W dalszym ciągu niniejszej publikacji podaje się wykaz utworów, na których prowadzono badania wraz ze sposobem pobierania z nich próbek oraz jako wyniki tych badań szeregi rozdzielenie liczby sylab w wyrazach. Jako charakterystyki szeregów przyjęto średnie i entropie. Ustala się związki między średnią i entropią. Stwierdza się następnie zgodność

jednego z rozkładów empirycznych z uogólnionym rozkładem Poissona podanym przez W. Fucksą.

2. Badania przeprowadzono na próbach z następujących utworów:

1. M. Rej, <i>Żywot człowieka pociwego</i> , [5],	} jedna próba
2. — <i>Wspólne narzekanie wszej Korony na porządną niedbałość naszą</i> , [5],	
3. Ł. Górnicki, <i>Dworzanin</i> , [6],	
4. B. Prus, <i>Placówka</i> , [7],	
5. — <i>Anielka</i> , [8],	
6. S. Żeromski, <i>Ludzie bezdomni</i> , [9],	
7. S. Lem, <i>Astronaucci</i> , [10],	
8. — <i>Obłok Magellana</i> , [11].	

W utworach 1-3 i 6-8 pobierano próbę gronową, tzn. losowano stronicę i brano pod uwagę wszystkie słowa znajdujące się na danej stronicy z wyjątkiem słów w językach obcych i nazwisk oraz nazw miejscowości. W utworach 4 i 5 brano słowa z kolejnych stron poczynając od pierwszej stronicy tekstu aż do uzyskania próby odpowiedniej wielkości. W tych ostatnich utworach uwzględniano wszystkie słowa znajdujące się w tekście oprócz słów w językach obcych. Utwory 4 i 5 były przy tym badane dwukrotnie na tym samym tekście, pierwszy raz z uwzględnieniem dialogów, drugi raz — bez uwzględnienia. Przy roboczej definicji słowa kierowano się pisownią wymienionych w odsyłaczu tekstów i każdy osobno napisany zespół liter traktowano jako zapis osobnego wyrazu. Każde słowo klasyfikowano według liczby sylab, które ono zawiera, przy czym zerosylabowe przyimki w rodzaju „w” „z” zaliczono do jednosylabowych. W ten sposób rozróżniono słowa jednosylabowe, dwusylabowe itd., aż do słów ośmiosylabowych (słów o większej ilości sylab w badanych tekstach nie znaleziono). Otrzymane dane podajemy w tabelicy 1. W tabelicy tej podane są wielkości próbek wraz ze szczegółowym wyliczeniem, ile słów było jedno-, dwu- itd., aż do ośmiosylabowych.

3. Z otrzymanych szeregów rozdzielczych obliczono rozkłady, które przedstawiono graficznie na rysunku 1 oraz dystrybuanty, które uwidocznione są na rysunku 2. Dla większej pogładowości punkty skoku poszczególnych dystrybuant zostały na rysunku nieco przesunięte.

Zarówno rozkłady jak i dystrybuanty wykazują bardzo ciekawe prawidłowości. Jeśli rozpatrywać rozkłady w porządku chronologicznym, to obserwujemy (rys. 1) zmniejszanie się frakcji słów jedno- i dwusylabowych przy równoczesnym zwiększaniu się frakcji słów o większej liczbie sylab. To samo zjawisko zaobserwować można na wykresie dystrybuant (rys. 2): dystrybuanty odpowiadające utworom późniejszym w czasie znajdują się na wykresie całkowicie poniżej dystrybuant odpowiadających

TABLICA 1

Empiryczne rozkłady ilości sylab w słowie

M. Rej, *Żywot człowieka
pocziwego; Wspólne na-
rzekanie wszej Korony na
porządną niedbalość naszą*

<i>i</i>	<i>n_i</i>	<i>p_i</i>
1	1238	0,4127
2	1079	0,3597
3	489	0,1630
4	178	0,0593
5	16	0,0053
6	—	—
7	—	—
8	—	—

3000 1,000

B. Prus, *Placówka*
(bez dialogów)

<i>i</i>	<i>n_i</i>	<i>p_i</i>
1	1322	0,3635
2	1431	0,3942
3	622	0,1715
4	217	0,0597
5	38	0,0105
6	2	0,0006
7	—	—
8	—	—

3632 1,0000

S. Żeromski, *Ludzie
bezdomni*

<i>i</i>	<i>n_i</i>	<i>p_i</i>
1	858	0,3432
2	931	0,3724
3	494	0,1976
4	178	0,0712
5	33	0,0132
6	6	0,0024
7	—	—
8	—	—

2500 1,0000

Ł. Górnicki, *Dworzanin*

<i>i</i>	<i>n_i</i>	<i>p_i</i>
1	2504	0,4173
2	2336	0,3894
3	912	0,1520
4	220	0,0367
5	26	0,0043
6	2	0,0003
7	—	—
8	—	—

6000 1,0000

B. Prus, *Anielka*
(z dialogami)

<i>i</i>	<i>n_i</i>	<i>p_i</i>
1	2500	0,3544
2	2384	0,3380
3	1481	0,2100
4	549	0,0778
5	123	0,0174
6	17	0,0024
7	—	—
8	—	—

7054 1,0000

S. Lem, *Astronauta*

<i>i</i>	<i>n_i</i>	<i>p_i</i>
1	1999	0,3333
2	1896	0,3160
3	1324	0,2206
4	623	0,1038
5	128	0,0213
6	27	0,0045
7	3	0,0005
8	—	—

6000 1,0000

B. Prus, *Placówka*
(z dialogami)

<i>i</i>	<i>n_i</i>	<i>p_i</i>
1	2800	0,3926
2	2762	0,3873
3	1056	0,1481
4	456	0,0639
5	54	0,0075
6	4	0,0006
7	—	—
8	—	—

7132 1,0000

B. Prus, *Anielka*
(bez dialogów)

<i>i</i>	<i>n_i</i>	<i>p_i</i>
1	1790	0,3398
2	1739	0,3301
3	1182	0,2244
4	446	0,0847
5	97	0,0184
6	14	0,0026
7	—	—
8	—	—

5268 1,0000

S. Lem, *Obłok
Magellana*

<i>i</i>	<i>n_i</i>	<i>p_i</i>
1	1075	0,3349
2	1057	0,3293
3	668	0,2081
4	322	0,1003
5	68	0,0213
6	14	0,0043
7	4	0,0012
8	2	0,0006

3210 1,0000

Rys. 1. Rozkład liczby sylab w słowie dla autorów polskich

Rys. 2. Dystrybuanty rozkładów liczby sylab w słowie dla autorów polskich

utworom wcześniejszym. Charakterystyczne jest prawie całkowite pokrywanie się rozkładów obu utworów Lema oraz stosunkowo znaczne zbliżenie do siebie rozkładów dla utworów Reja i Górnickiego. Rzuca się natomiast w oczy dość duża rozbieżność między rozkładami odnoszącymi się do dwóch utworów tego samego autora, mianowicie Prusa — *Placówką* i *Anielką*. Można to chyba wyjaśnić przystosowaniem języka autora do tematu i mowy środowiska: język *Placówki*, opisującej środowisko chłopskie, bardziej prymitywny, zbliża się do języka staropolskiego (w naszych badaniach charakterystyki te są bliskie charakterystykom języka Reja i Górnickiego), język *Anielki*, opisującej środowisko warstw szlacheckich, jako bardziej rozwiniętych intelektualnie, jest bliższy językowi współczesnemu.

Ciekawe jest porównanie rozkładów dla *Anielki* z dialogami i bez dialogów (rys. 3), to samo dla *Placówki* (rys. 4). W tekstach bez dialogów obserwujemy zwiększenie się frakcji ilości słów dłuższych, w tekstach przekazujących mowę prostą zwiększa się ilość słów krótkich. Odbija się to w sposób istotny na średniej i entropii, o czym będzie mowa niżej.

Rys. 3. Rozkład liczby sylab w słowie

Rys. 4. Rozkład liczby sylab w słowie

Na rysunku 5 przedstawiona jest zależność entropii od wartości średniej liczby sylab. Na osi odciętych są odłożone wartości średniej $\bar{i}$ liczby sylab w słowie utworu, na osi rzędnych — wartości entropii $S = -\sum p_i \log p_i$ podane w decitach, tzn. w dziesiętnych jednostkach ilości informacji. Wartość średnią $\bar{i}$ obliczono według wzoru $\bar{i} = \sum i p_i$, gdzie p_i oznacza to samo co we wzorze na entropię i jest frakcją ilości słów badanego tekstu składających się z i sylab.

Obserwując wykres na rysunku 5 spostrzegamy, że zarówno $\bar{i}$ jak S wzrastają chronologicznie z wyjątkiem jednego przypadku: Rej-Górnicki. Mimo, że Górnicki jest autorem późniejszym niż Rej, na wykresie na ry-

sunku 5 punkt odpowiadający *Dworzaninowi* wyprzedza punkt odpowiadający utworom Reja. Można to objaśnić w ten sposób, że *Dworzanin* jest utworem napisanym w dialogu. Jak już mówiliśmy wyżej, dialog wpływa na zmniejszenie się średniej liczby sylab w słowie, widzimy, że

Rys. 5. Zależność entropii od wartości średniej dla autorów polskich

zmniejsza on również entropię. Ilustruje to doskonale na rysunku 5 wzajemne położenie *Anielki* z dialogami i bez dialogów, jak również *Placówki* z dialogami i bez. W obu tych przypadkach punkty odpowiadające utworom z dialogami leżą poniżej odpowiednich punktów dla utworów bez dialogów.

Najbardziej uderzające w wykresie na rysunku 5 jest jednak to, że mimo tak różnorodnych tekstów o różnej treści, pisanych w różnych czasach przez różnych autorów, wszystkie otrzymane przez nas punkty układają się prawie dokładnie wzdłuż linii prostej. Czy jest to jednak rzeczywiście prosta? Otrzymane przez nas punkty leżą w wąskim zakresie zmiany $\bar{i}$ i S : $1,8 < \bar{i} < 2,2$ zaś $0,50 < S < 0,62$. Zależność tę dla szerszego zakresu badał W. Fucks [2], który stwierdził, że zależność między średnią i entropią nie jest liniowa lecz przedstawia się w sposób pokazany na rysunku 6. Na wykresie tym obok krzywej teoretycznej, wyrażającej zależność entropii od średniej, widzimy również punkty empiryczne średnia-entropia, uzyskane przez Fucksę dla dziewięciu języków. Punkty te leżą dość dobrze na wspomnianej krzywej teoretycznej z wyjątkiem języka arabskiego, co do którego jednak sam Fucks zastrzega się, że odpowiednie dane zostały otrzymane z niewielkiej, mało reprezentatywnej próbki. Pogrubiona część wykresu odpowiada tu naszemu wykresowi z rysunku 5. Jak widzimy, wykres nasz dobrze pokrywa się z wykresem teoretycznym podanym przez Fucksę, odpowiadającym następującej zależności:

$$S = - \sum_{i=0}^{\infty} \frac{e^{-(\bar{i}-1)} (\bar{i}-1)^{i-1}}{(i-1)!} \log \frac{e^{-(\bar{i}-1)} (\bar{i}-1)^{i-1}}{(i-1)!}.$$

Widzimy również, że wykres otrzymany przez nas, będąc łukiem wspomnianej krzywej teoretycznej, dla niewielkiego zakresu zmiany parametrów mało różni się od prostej.

Rys. 6. Autorzy polscy na tle dziewięciu języków

Przy sposobności chcielibyśmy zrobić uwagę co do samego określenia punktu odpowiadającego na wykresie temu czy innemu językowi. Cennaść takiego punktu jest chyba dość względna, ponieważ położenie jego nie może być ściśle określone ze względu na trudności zdefiniowania języka polskiego czy innego. Z tych względów podaliśmy cały łuk odpowiadający różnym autorom polskim zamiast obliczać współrzędne jednego mało sprecyzowanego punktu odpowiadającego językowi polskiemu.

Byłoby rzeczą ciekawą, zwłaszcza dla lingwistów, rozpatrzenie otrzymanych wyżej wyników w świetle koncepcji G. Herdana ([3], str. 187) o tendencjach zmian zachodzących w językach w kierunku monosylabizmu i polisylabizmu.

4. Rozpatrzmy teraz rozkład liczby sylab w słowie. Według przyjętej przez nas definicji każde słowo musi zawierać przynajmniej jedną sylabę. Jeśliby słowa powstawały z sylab na drodze czysto losowej, to otrzymany rozkład dodatkowej ilości sylab (tj. liczby sylab w słowie minus jedna) byłby poissonowski. Tak jednak nie jest. W. Fucks zauważył, że rozkłady empiryczne wyrażają się pewnymi uogólnionymi schematami Poissona. Powstają one jako suma rozkładów zerojedynekowych i niezależnych od nich rozkładów poissonowskich. Dla badanego przez nas roz-

kładu liczby sylab w słowie w *Placówce* (z dialogami) omawiana zależność teoretyczna wyraża się wzorem

$$(1) \quad v(i) = e^{-[\bar{i}-(1+\beta)]} \left\{ (1-\beta) \frac{[\bar{i}-(1+\beta)]^{i-1}}{(i-1)!} + \beta \frac{[\bar{i}-(1+\beta)]^{i-2}}{(i-2)!} \right\} \text{ dla } i \geq 2, \\ v(1) = e^{-[\bar{i}-(1+\beta)]} (1-\beta),$$

gdzie i wyraża liczbę sylab w słowie, $\bar{i}$ — oczekiwaną liczbę sylab w słowie, $v(i)$ jest prawdopodobieństwem, że dane słowo liczyć będzie i sylab, zaś β jest parametrem rozkładu Fuckska; w naszym przypadku $\beta = 0,2$, co obliczono metodą momentów.

Podany przez nas wzór (1), wystarczający do opisu badanego przez nas rozkładu, jest przypadkiem szczególnym ogólniejszej klasy rozkładów zbadanych przez Fuckska w [2]:

$$(2) \quad v(i) = e^{-(\bar{i} - \sum_{\alpha=1}^{\infty} \beta_{\alpha})} \sum_{v=0}^{\infty} (\beta_v - \beta_{v+1}) \frac{(\bar{i} - \sum_{\alpha=1}^{\infty} \beta_{\alpha})^{i-v}}{(i-v)!}.$$

Rys. 7. Rozkład liczby sylab w słowie dla *Placówki* Prusa (z dialogami)

Wzór (1) otrzymuje się z wzoru (2) przez podstawienie $\beta_0 = 1$, $\beta_1 = 1$, $\beta_2 = \beta = 0,2$. Metody estymacji parametrów β_{α} można znaleźć w cytowanej pracy [2].

Na rysunku 7 wykreślono ⁽¹⁾ rozkłady teoretyczny i empiryczny ilości sylab w słowie dla *Placówki*. Na podstawie wykresów widzimy, że zgodność obu rozkładów jest na oko zadowalająca. Niemniej jednak porównywanie obu rozkładów testem χ^2 daje istotne różnice:

i = liczba sylab w słowie	1	2	3	4	> 5
liczności zaobserwowane	2800	2762	1056	456	58
liczności teoretyczne	2777	2693	1220	353	89

$$\chi_a^2 \approx 60,$$

$$\Pr\{\chi^2 > \chi_a^2\} \ll 0,001.$$

Jest rzeczą ciekawą, że wysoka wartość χ_a^2 jest wywołana przede wszystkim dużymi różnicami w klasach $i = 3$ oraz $i = 4$, przy czym w klasie $i = 3$ mamy niedobór a w klasie $i = 4$ znaczną nadwyżkę licznosci zaobserwowanych w stosunku do licznosci teoretycznych wynikających z wzoru (1) dla wartości $\beta = 0,2$. Po połączeniu tych klas różnice między obu rozkładami znacznie się zmniejszają:

i = liczba sylab w słowie	1	2	3-4	> 5
liczności zaobserwowane	2800	2762	1512	58
liczności teoretyczne	2777	2693	1573	89

$$\chi_a^2 = 15,13,$$

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

Jakie mogą być przyczyny wykrytej powyżej testem χ^2 niezgodności? Otóż po pierwsze, wzór (1) był wyprowadzony przy założeniu niezależności kolejno losowanych elementów. Tymczasem elementy naszej próbki były wyrazami tekstu kolejno następującymi po sobie, a więc założenie o niezależności nie było spełnione. Powiązanie słów tematyką i treścią utworu mogło spowodować wystąpienie lokalnych nadwyżek słów i sylabowych, co z kolei spowodowało pewne odstępstwa od postulowanego rozkładu. Po drugie, zaobserwowane odstępstwo można tłumaczyć również tym, że wzór (1) nie opisuje zadowalająco omawianego rozkładu. Zwróćmy uwagę na to, że mamy tu wzór w stosunku do formuły (2) bardzo uproszczony z uwzględnieniem tylko jednego parametru β . Fucks w swej pracy [2], przy opisywaniu uzyskanych przez siebie rozkładów, posługuje się dwoma

⁽¹⁾ Odcinki łamanych linii prostej służą tylko do przejrzystego powiązania grup punktów.

lub trzema parametrami. Być może i w naszym przypadku po dobraniu większej ilości parametrów zdołalibyśmy uzyskać lepszą zgodność.

Głębsze badania związane z tym ostatnim zagadnieniem będą przeprowadzone oddzielnie.

5. Autorowie prowadzą obecnie badania nad porównaniem omawianych w pracy rozkładów empirycznych z ogólniejszymi rozkładami Fucks'a (patrz wzór (2)). Będzie rzeczą ciekawą wyjaśnić, czy zachodzą zmiany parametrów dla utworów pochodzących z różnych epok, czy też parametry te są konstantami językowymi. Prowadzone są również badania nad efektywnością próbek, zmierzające do ustalenia wielkości próbki potrzebnej do uzyskania odpowiednich danych.

Powyższe zagadnienia będą tematem innej publikacji autorów.

Prace cytowane

[1] W. Fucks, *Mathematische Analyse von Sprachelementen*, Sprachstil und Sprachen, Westdeutscher Verlag Koln und Oplanden, 1955.

[2] — *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 статей, Москва 1957).

[3] G. Herdan, *Language as choice and chance*, Groningen 1956.

[4] M. R. Mayenowa, *O współzależności rozwoju języka literackiego i form literackich*, Nauka Polska 2 (26), 1959.

[5] M. Rej, *Pisma prozą i wierszem*: opracował A. Brueckner, wyd. 2, Kraków 1926.

[6] Ł. Górnicki, *Dworzanin*, Biblioteka Pisarzy Polskich i Obcych, Książka i Wiedza, 1950.

[7] B. Prus, *Placówka*, Biblioteka Szkolna, 1960.

[8] — *Anielka*, Biblioteka Szkolna, 1960.

[9] S. Żeromski, *Ludzie bezdomni*, Czytelnik, 1954.

[10] S. Lem, *Astronauta*, Czytelnik, 1952.

[11] — *Obłok Magellana*, Iskry, 1955.

INSTYTUT MATEMATYCZNY POLSKIEJ AKADEMII NAUK

Praca wpłynęła 19. 6. 1961

A. БАРТКОВЯК и Б. ГЛЕЙХГЕВИХТ (Вроцлав)

ХАРАКТЕРИСТИКА ПОЛЬСКОГО ЯЗЫКА ЧИСЛОМ СЛОГОВ В СЛОВЕ

РЕЗЮМЕ

Число слогов в слове в данном тексте является случайной величиной, принимающей дискретные значения: 1, 2, 3, ... Каждый писатель может быть схарактеризован распределением этой случайной величины. В этой работе рассматри-

вается язык нескольких польских писателей из разных столетий. Устанавливаются некоторые интересные закономерности: среднее значение и энтропия растут систематически с временем, что показано на рис. 5. Находим также место польского языка среди других языков на кривой зависимости энтропии от средней (рис. 6). Устанавливается разница между текстом с диалогом и текстом без диалога. Полученная дальше формула (1) даёт хорошую аппроксимацию эмпирической функции распределения, полученной для книги „*Placówka*” („Форпост”) Б. Пруса, рассматриваемой без диалогов.

A. BARTKOWIAKOWA and B. GLEICHGEWICHT (Wrocław)

ON THE SYLLABLE CHARACTERISTIC OF THE POLISH LANGUAGE

SUMMARY

The number of syllables in a word of a text is a random variable, taking discrete values 1, 2, 3, ... Every writer can be characterised by the distribution of this random variable. In our paper we investigate the language of several Polish writers representing different centuries. We have found some interesting regularities: the average mean number and the entropy grow systematically with time, which is shown in figure 5. We have also established the place of the Polish language among other languages on the entropy-means curve (figure 6). We point out some differences between texts with dialogues and without them and a mathematical formula (1) giving a good approximation to the empirical distribution obtained for B. Prus, *Placówka*, the words being counted in the text without dialogues.