

A. BARTKOWIAKOWA (Wrocław)

Ó ROZKŁADZIE OKREŚLEŃ W ZDANIACH OPISOWYCH
ŻEROMSKIEGO I SIENKIEWICZA

1. Wstęp, opis zagadnienia. W ostatnich czasach notujemy gwałtowny wzrost zapotrzebowania na matematykę ze strony różnych praktycznych dyscyplin, jak np. medycyna, biologia, agronomia. Nikt już nie wątpi w to, że matematyka w wyżej wymienionych dziedzinach jest narzędziem bardzo użytecznym, że pozwala na wyciąganie wniosków, które na innej drodze były by bardzo trudno, albo w ogóle nieosiągalne.

Stosunkowo niedawno w domenę zastosowań matematyki weszła jeszcze jedna gałąź nauki, mianowicie językoznawstwo. Stosowanie metod matematycznych w tej dziedzinie natrafiało i natrafia jeszcze nadal na duże trudności, i nie w tym dziwnego, skoro nauki humanistyczne były zawsze przeciwstawiane naukom ścisłym, matematyczno-przyrodniczym. Niemniej jednak ostatnie dwudziestolecie dowiodło, że metody statystyki matematycznej i teorii informacji mogą być cennym narzędziem badań w zakresie językoznawstwa, i potrafią ukazać rozważane tam zagadnienia w zupełnie innym oświetleniu.

Zagadnienie, którym pragnę zająć się w tej i przyszłych moich pracach, to zagadnienie stylu pisarskiego w różnych utworach prozaicznych.

Prawie wszystkie dotychczasowe klasyczne badania z tego zakresu dokonywane były na podstawie analizy pewnych charakterystycznych fragmentów utworu. W *Stylistyce polskiej* H. Kurkowskiej i S. Skorupki [6] czytamy: „Najistotniejszym momentem (w analizie językowo-stylistycznej — A. B.) jest ocena zjawisk językowych na tle kontekstu, a więc ich ocena jako elementów wypowiedzi, określająca celowość i skuteczność ich użycia, tzn. to, czy w sposób odpowiadający zamierzeniom piszącego lub mówiącego wyrażają one treść wypowiedzi i czy wywołują zamierzoną reakcję u odbiorcy”. Chodzi więc o to, aby określić, za pomocą jakich form i środków językowych wyraża pisarz treść badanego fragmentu. Analizy tej dokonuje się rozważając za każdym razem indywidualnie poszczególną sytuację. Są to więc badania odcinkowe, jednostkowe.

Oczywiście jest to jeden z możliwych sposobów podejścia do zagadnienia. W przeciwieństwie do tego klasycznego stanowiska istnieje inne, które zakłada badanie utworu jako integralnej całości biorąc pod uwagę pewne własności zbiorcze, a jako podstawowym narzędziem operuje statystyką matematyczną.

Fakt, że przy opisywaniu dzieła literackiego możemy posługiwać się metodami matematycznymi, nie powinien budzić zdziwienia. Nikt np. nie wątpi, że w termodynamice metody statystyczne są właściwe, i że dopiero one pozwalają uchwycić istotę zjawiska. Co do utworu językowego, to można przeprowadzić pewną analogię między nim a zbiorem cząsteczek gazu. Każdy z nich składa się z dużej ilości elementarnych cząstek. Cząsteczka gazu brana indywidualnie może mieć różną prędkość, rozpatrując jednak średnią prędkość całego zbioru otrzymujemy inną wielkość, charakterystyczną dla danego zbioru, mianowicie temperaturę. Podobnie, abstrahując od indywidualnych własności elementów językowych i rozważając ich zbiór jako całość, możemy otrzymać nowe charakterystyki niedostępne przy badaniu indywidualnym.

Dotychczasowe badania na tym polu szły w różnych kierunkach. I tak B. Mandelbrot [1], [7] wykazał wiele analogii między prawami mechaniki statystycznej a prawami makrolingwistycznymi (termin zaproponowany przez Mandelbrota) rządzącymi językiem. P. Guiraud [4] podał pewne wskaźniki mogące charakteryzować style różnych epok i pisarzy. W. Fuchs [2] opisuje matematycznie proces składania różnych elementów językowych z drobniejszych cegiełek. Badania nad słownikiem pisarza i parametrami różnych rozkładów posłużyły G. Udny Yulowi [11] i G. Herdanowi [5] do napisania obszernych monografii. Osobną, bogato rozwiniętą i wciąż dalej rozwijającą się dziedzinę stanowią zagadnienia przekładu maszynowego [8].

Duża część rozpatrywanych zagadnień dotyczyła raczej ogólniejszych zagadnień językowych, a jeśli nawet wprowadzono gdzieś pojęcie charakterystyki stylu, to było ono oparte na wielkościach dość abstrakcyjnych (np. Mandelbrota „temperatura utworu” lub „bogactwo słownika” Guirauda) lub też na wielkościach intuicyjnie w niewielkim stopniu związanych z pojęciem stylu (np. poprzez częstość sylab) albo wymagających dużego nakładu pracy (porównywanie słownika). Bardzo prostą charakterystyką stylu zajmował się Yule [10], który badał długość zdania wyrażoną w ilości słów, i stosował tę cechę do zagadnień autorstwa. Cecha ta sama jedna okazuje się mało dyskryminująca. Dlatego też wydawało mi się potrzebnym szukać innych charakterystyk stylu, stosunkowo prostych i łatwych do obliczenia, które by jednak były na tyle efektywne, żeby wykazywały różnice między pisarzami reprezentującymi odrębne (w intuicyjnym tego słowa znaczeniu) style. Jako cechy, które mogłyby odpowiadać powyższym wymaganiom, wzięłam:

1. Określenia do podmiotu i orzeczenia. Rozważałam długość określeń (w słowach) oraz ilość określeń przypadającą na jedno zdanie (definicja określenia — patrz następny paragraf).

2. Dendrytową charakterystykę zdań czyli ilość zdań hipo- i parataktycznych w zdaniu złożonym.

W pracy niniejszej pragnę przedstawić wyniki badań nad określeniami. Znalazłam mianowicie rozkłady ilości słów w określeniu oraz ilości określeń przypadających na jedno zdanie główne. Rozkłady empiryczne tych cech można wyrazić rozkładami empirycznymi oznaczonymi w tekście wzorami (1) oraz (2). Podane są próby interpretacji tego, że otrzymano właśnie rozkłady wyrażające się poprzez wzory (1) i (2). Badania prowadzono na tekstach dwóch autorów: Żeromskiego i Sienkiewicza, przy czym okazało się, że cecha pierwsza (ilość słów w określeniu) jest dla obu pisarzy dyskryminująca.

Wyniki badań nad dendrytową charakterystyką zdań zostaną podane później.

2. Definicja określeń, materiał wyjściowy do badań. Pod nazwą *określenie* rozumiem grupę wyrazów określających podmiot i orzeczenie i spełniających w zdaniu funkcję dopełnienia, okolicznika lub przydawki. Najlepiej wyjaśnić to na przykładach. Weźmy pod uwagę dwa zdania wyjęte z *Ogniem i mieczem*: 1. „Wtem na skrócie drogi namiestnik wstrzymał konia gdyż nowy a dziwny widok uderzył jego oczy”. 2. „W pośrodku gościńca leżała na boku kolaska ze złamaną osią”. W pierwszym zdaniu podmiotem jest „namiestnik”, żadnych bliższych określeń do niego nie ma. Orzeczeniem jest „wstrzymał” i mamy tu następujące określenia: „konia” (dopełnienie bliższe), „gdź nowy a dziwny widok uderzył jego oczy” (okolicznik przyczyny), „wtem” (okolicznik czasu), „na skrócie drogi” (okolicznik miejsca). W drugim zdaniu do podmiotu „kolaska” mamy określenie „ze złamaną osią” (przydawka), orzeczenie „leżała” rozwijają szerzej grupy wyrazów „na boku” (okolicznik sposobu), „w pośrodku gościńca” (okolicznik miejsca).

Widzimy więc, że określenie jest częścią zdania głównego, w szczególności część ta może stanowić całe zdanie podrzędne.

Przy obliczaniu ilości słów w określeniu przyjąłam zasadę nie liczenia przyimków i spójników znajdujących się na początku określenia. Przyimki i spójniki znajdujące się wewnątrz były liczone podobnie jak inne słowa. Dla wyżej podanych przykładów liczność określeń jest następująca: „konia” — określenie jednowyrazowe, „gdź nowy a dziwny widok uderzył jego oczy” — siedmiowyrazowe, „wtem” — jednowyrazowe, „na skrócie drogi” — dwuwyrazowe, „ze złamaną osią” — dwuwyrazowe, „na boku” — jednowyrazowe określenie. Można się a priori spodziewać, że teksty literackie różnych autorów będą się różnić zarówno pod wzglę-

dem ilości słów składających się na jedno określenie, jak również liczbą określeń przypadających na poszczególne zdania. W dalszym ciągu zajmować się będę tymi dwoma zagadnieniami.

Badania moje ograniczyłam do utworów powieściowych Żeromskiego i Sienkiewicza, przy czym brałam pod uwagę tylko zdania opisowe, pomijając dialogi. Materiał wyjściowy do badań otrzymałam, losując z utworu próbki gronowe, liczące około 50 zdań opisowych każda. Dla wyciągnięcia pewniejszych wniosków dołączyłam do badanych utworów powieściowych jedną pracę ekonomiczną, mianowicie A. Karpińskiego *Zagadnienia socjalistycznej industrializacji Polski*.

3. Rozkład ilości słów w jednym określeniu. Rachując dla wybranej próbki ilość określeń składających się z jednego słowa (n_1), z dwóch słów (n_2) itd. otrzymałam empiryczny rozkład ilości słów w określeniu:

$$p'_{k\text{emp}} = \frac{n_k}{N}, \quad k = 1, 2, \dots, \quad N = \sum n_k.$$

Rozkład ten daje się bardzo dobrze przybliżyć przez graniczny rozkład Polya wyrażający się następująco:

$$(1) \quad \begin{cases} P_k = \left(\frac{t}{1+bt} \right)^k \cdot \frac{1(1+b) \dots [1+(k-1)b]}{k!} P_0 & \text{dla } k \geq 1, \\ P_0 = (1+bt)^{-1/b} & \text{dla } k = 0. \end{cases}$$

Parametry b i t tego rozkładu możemy estymować przez średnią \bar{x} oraz wariancję s^2 za pomocą wzorów:

$$t = \bar{x} = \sum_{k=0}^{\infty} k p_k, \quad b = \frac{s^2 - t}{t^2},$$

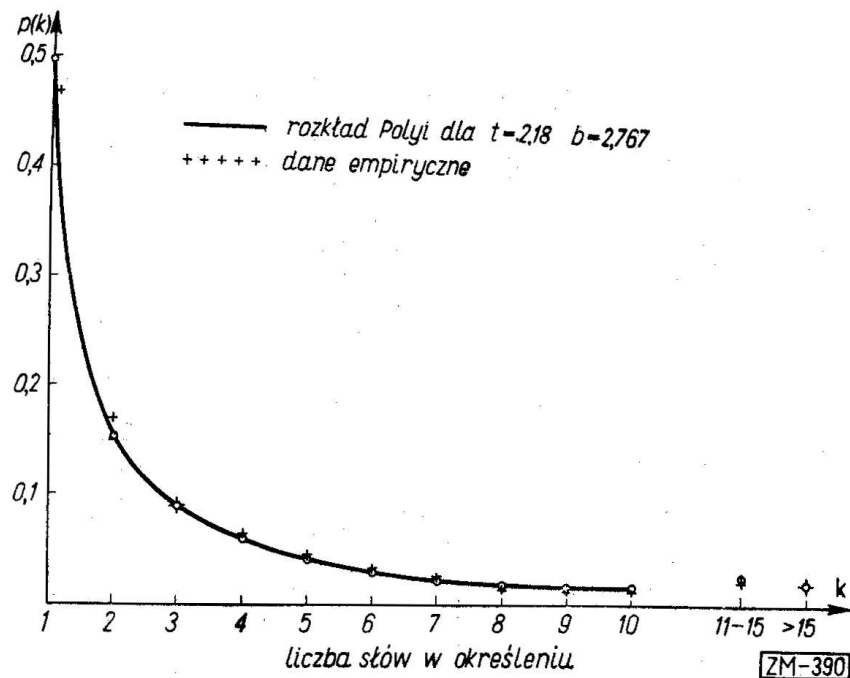
$$s^2 = \sum_{k=0}^{\infty} k^2 p_k - \left(\sum_{k=0}^{\infty} k p_k \right)^2.$$

Aby powyższy wzór opisywał otrzymany empirycznie rozkład, należy jedynie uczynić przyporządkowanie:

$$p'_{k\text{emp}} = P_{k-1} \text{ Polya}, \quad k = 1, 2, \dots,$$

ponieważ każde określenie musi zawierać co najmniej jeden element.

Jako przykład podaję rozkład uzyskany z *Przedwiośnia* (tablica 1 i wykres 1).

Rys. 1. Rozkład liczby słów w określeniu dla *Przedwiośnia*

TABLICA 1

Rozkład ilości słów w określeniu dla *Przedwiośnia* $n_{k \text{ emp}}$ — częstości otrzymane empirycznie, $n_{k \text{ Polya}}$ — częstości otrzymane z wzoru (1) dla $t = 2,18$, $b = 2,767$

k	1	2	3	4	5	6	7	8	9	10	11-15	> 15
$n_{k \text{ emp}}$	980	350	192	126	92	68	57	38	22	26	64	45
$n_{k \text{ Polya}}$	1018	316	184	124	90	67	51	40	32	25	69	44

$$\chi^2 = 10,13, \quad \Pr\{\chi^2 \geq 10,13\} \cong 0,35.$$

Widzimy, że zgodność obu rozkładów jest bardzo dobra. Przy założonym rozkładzie, odchylenia takie jak w tabeli 1 lub jeszcze większe otrzymujemy w 35 przypadkach na 100 badanych, a więc możemy śmiało przyjąć, że widoczne w tabeli 1 rozbieżności są spowodowane przyczynami losowymi i są nieistotne.

Podobne zgodności otrzymałam dla prób innych utworów. Wynika stąd, że rozkład ilości słów w określeniu opisuje się rzeczywiście wzorem (1) podanym powyżej.

Fakt, że otrzymany rozkład empiryczny daje się bardzo dobrze opisać matematycznym wzorem, jest oczywiście powodem do zadowolenia matematyka, że rzeczywistość tak pięknie daje się ująć w ramy matematycznych formuł. Ale nie tylko o to chodzi. O wiele ważniejszym jest, czy z przedstawionego faktu można wyciągnąć pewne wnioski co do istoty

mechanizmu powstawania danego zjawiska. Przypomnijmy więc jaki to proces statystyczny jest opisywany przez rozkład Polyi.

Wyjaśnimy to na modelu urnowym.

Przypuśćmy, że w urnie znajduje się Np kul czarnych i $N(1-p)$ kul białych. Wyciągamy na chybił trafił jedną kulę, po czym wkładamy na powrót do urny $(1 + N\beta)$ kul tego koloru co wyciągnięta (β jest tu pewną wielokrotnością $1/N$). Rezultat następnego ciągnięcia zależy od tego, czy poprzednim razem wyciągnęliśmy kulę białą czy też czarną.

Niech X = ilości kul czarnych otrzymanych w n ciągnięciach. Prawdopodobieństwo, że wyciągniemy k kul czarnych, wynosi:

$$p_k^n = \binom{n}{k} \frac{p(p+\beta) \dots [p+(k-1)\beta] q(q+\beta) \dots [q+(n-k-1)\beta]}{1(1+\beta)(1+2\beta) \dots [1+(n-1)\beta]}.$$

Rozkład ten jest znany jako rozkład Polyi. Przy β dążącym do zera rozkład Polyi dąży do rozkładu Bernoulliego.

Gdy p jest małe a n duże, korzystamy z wzorów asymptotycznych. Mianowicie, gdy $\lim_{n \rightarrow \infty} np = t$, $\lim_{n \rightarrow \infty} n\beta = bt$, otrzymujemy w granicy wyrażenie, podane poprzednio jako graniczny rozkład Polyi i oznaczone (1), które tak świetnie opisuje rozkład ilości słów w określeniu.

Oczekiwana ilość wyciągniętych kul czarnych i wariancja wynoszą odpowiednio:

$$E(X) = t, \quad D^2(X) = t(1 + bt),$$

skąd możemy obliczyć parametry t i b rozkładu.

W interpretacji językowej pojawieniu się kuli czarnej odpowiada wystąpienie słowa w określeniu. Proces pojawienia się dalszych słów jest „zaraźliwy”, tzn. jeśli pisarz użył jakiegoś pojęcia określającego, to kojarzy się mu ono natychmiast z jakimś innym pojęciem dodatkowym, uzupełniającym to pierwsze. Po umieszczeniu tego drugiego w określeniu następuje z kolei nowe skojarzenie, wskutek czego pojawiają się dalsze wyrazy. Mamy tu jak gdyby lawinowe pojawianie się słów; albo dalsze określenia nie pojawiają się w ogóle, albo pojawiają się ich bardzo dużo.

Wzór (1) dopuszcza jednak również inną interpretację, podaną przez Greenwooda i Yule'a [3].

Niech X będzie zmienną losową, oznaczającą ilość słów w określeniu. Załóżmy, że X ma rozkład Poissona, tzn.

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

gdzie λ jest wartością oczekiwaną zmiennej X . Na ogół zakłada się, że proces przedstawiany zmienną losową X jest wyregulowany tak, że wartość średnia λ pozostaje przez cały czas stała. Można jednak wyobrazić

sobie, że parametry procesu zmieniają się między jedną a drugą jego realizacją, w szczególności, że oczekiwana wartość λ zmienia się między jednym a drugim zdaniem. Znaczy to, że λ nie jest wielkością stałą, lecz również zmienną losową o jakiejś dystrybuancie $F(\lambda)$. W takim razie prawdopodobieństwo, że zmienna X przyjmie wartość k wyrazi się wzorem:

$$(A) \quad \Pr\{X = k\} = \int_0^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} dF(\lambda).$$

Jest to dość ogólny wzór na prawdopodobieństwo, zależny od rozkładu $F(\lambda)$.

Jeśli przyjąć za gęstość prawdopodobieństwa f prawo rozkładu Pearsona III typu o postaci

$$f(\lambda) = c^r \frac{\lambda^{r-1}}{\Gamma(r)} e^{-c\lambda},$$

to otrzymamy wówczas z wzoru (A) tzw. rozkład ujemny dwumianowy:

$$\Pr\{X = k\} = (-1)^k \binom{-r}{k} p^k q^r, \quad \text{gdzie} \quad p = \frac{1}{1+c}, \quad q = 1-p,$$

który jest równoważny granicznemu rozkładowi Polya (1).

W naszym przykładzie powyższa interpretacja znaczy tyle, że rozkład ilości słów w jednym określeniu jest czysto losowy, czyli poissonowski, natomiast parametr tego rozkładu, tj. wartość średnia, zmienia się w różnych określeniach. Trudność stanowi jedynie pytanie, dlaczego wartość średnia ma się zmieniać akurat według rozkładu gamma. To natomiast, że wartość średnia zmienia się w ogóle, wyczuwa się i widzi bardzo łatwo. Pisarz, stosownie do opisywanego obrazu dobiera odpowiednich środków. Aby oddać nastrój chwili, raz używa określeń długich, innym razem krótszych. Można by również przypuszczać, że każdy wątek tematyczny w utworze powieściowym ma swój oddzielny parametr.

Aby wyjaśnić bliżej tę sprawę, zbadalem jeszcze jeden utwór prozaiczny, tym razem nie powieściowy, lecz o treści ekonomicznej, mianowicie A. Karpińskiego *Zagadnienia socjalistycznej industrializacji Polski*. W przypadku takim nie ma już powodu, aby autor opisywał wykonywanie planu 6-letniego przez przemysł chemiczny innym językiem niż to samo przez przemysł elektroenergetyczny. Inaczej: Wydawało mi się, że w pracy naukowej autor nie jest zobowiązany dbać tak bardzo o nastrój chwili i dobór do tego odpowiednich środków jak powieściopisarz-artysta.

Ilość słów w określeniu dla badanego utworu ekonomicznego wyraża się również wzorem (1). Otrzymany rozkład empiryczny wraz z obliczo-

nym dla niego według wzoru (1) rozkładem teoretycznym podaje w tabelicy 2.

Zgodność jest tam wprawdzie trochę gorsza niż poprzednio, niemniej jednak istniejące odchylenia można tłumaczyć wpływami przypadkowymi.

TABLICA 2

Rozkład ilości słów w określeniu dla pracy A. Karpińskiego *Zagadnienia socjalistycznej industrializacji Polski*

$n_{k \text{ emp}}$ — częstości otrzymane empirycznie,

$n_{k \text{ Polyi}}$ — częstości obliczone według wzoru (1) dla $t = 5,11, b = 2,257$.

k	$n_{k \text{ emp}}$	$n_{k \text{ Polyi}}$	k	$n_{k \text{ emp}}$	$n_{k \text{ Polyi}}$	k	$n_{k \text{ emp}}$	$n_{k \text{ Polyi}}$
1	432	396	9	28	32	17	15	11
2	156	162	10	27	27	18	14	10
3	107	107	11	10	24	19	15	9
4	62	80	12	27	21	20	3	8
5	62	64	13	14	18	21	9	7
6	45	52	14	13	16	22-31	45	41
7	40	43	15	15	14	> 31	25	23
8	39	37	16	12	13			

$$\chi^2 = 30,00, \quad 20 \text{ stopni swobody}; \quad \Pr\{\chi^2 = 30,0\} = 0,07.$$

Tak więc otrzymany graniczny rozkład Polyi można tłumaczyć zarówno „zaraźliwym”, lawinowym pojawianiem się dodatkowych słów w określeniu, jak również zmianą parametru między określeniami.

Aby uzyskać pewne dodatkowe informacje, zbadalam sąsiednie pary określeń.

4. Badanie sąsiednich określeń. Określenia są na ogół uporządkowane liniowo w tekście. Rzadko zdarzało mi się zaobserwować, żeby jedno było przedzielone drugim. Tak więc do badań sąsiedztwa brałam pary określeń pojawiające się kolejno w utworze. Respektowałam sąsiedztwo nie tylko w obrębie jednego zdania, ale również między zdaniami. Dla dwóch zdań przykładowych, podanych w paragrafie 2 niniejszej pracy, kolejność określeń jest następująca (liczby oznaczają ilość słów liczonych w określeniu):

1, 2, 1, 7, 2, 1, 2,

skąd uzyskujemy sąsiednie pary:

(1,2) (2,1) (1,7) (7,2) (2,1) (1,2).

Pytamy się, czy liczności określeń sąsiadujących ze sobą są zależne, tzn. czy $p(k|r)$ zależy od r , przy czym $p(k|r)$ oznacza prawdopodobieństwo, że określenie będzie liczyć k słów pod warunkiem, że poprzedzające je określenie liczyło r słów.

Aby to zbadać, obliczyłam empiryczne stosunki

$$\mu_k^{(r)} = \frac{p(k|r)}{p(k)}, \quad r = 1, 2, 3-5, 6-10, > 10, \quad k = 1, 2, 3-5, 6-10, > 10.$$

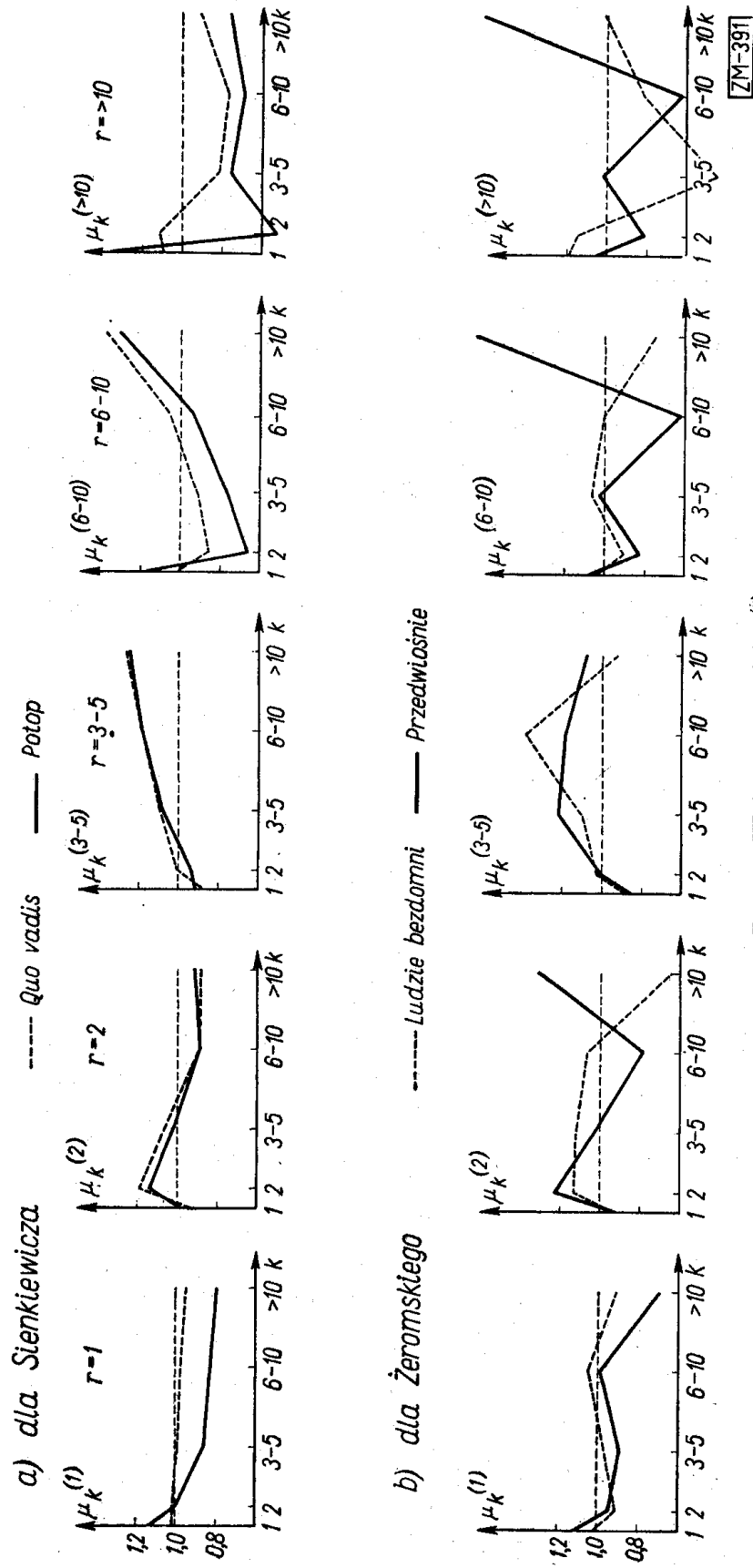
Znaczenie $p(k|r)$ podano powyżej, $p(k)$ zaś jest to prawdopodobieństwo a priori, że dane określenie liczyć będzie k słów niezależnie od tego, ile słów liczyło poprzedzające je określenie.

Nierówność $\mu_k^{(r)} > 1$ oznacza: Prawdopodobieństwo wystąpienia określenia k -elementowego po określeniu r -elementowym jest większe, niż by to wynikało z rozkładu a priori. Jeśli np. $\mu_1^{(1)} > 1$, to wyciągamy stąd wniosek, że po określeniu jednoelementowym występują znowu określenia jednoelementowe i to częściej niż określenia o innej liczności elementów, uwzględniając oczywiście ich prawdopodobieństwa a priori. Inaczej jeszcze można powiedzieć, że występuje tu „przyciąganie” określeń jednoelementowych.

Podobnie nierówność $\mu_{>10}^{(>10)} < 1$ interpretujemy jako „odpychanie się”, określeń o dużej ilości słów; prawdopodobieństwo wystąpienia określenia długiego, liczącego więcej niż dziesięć słów, po określeniu liczącym również co najmniej dziesięć słów, jest mniejsze, niż by to wynikało z rozkładu a priori.

Ustalając r można przedstawić $\mu_k^{(r)}$ jako funkcję k w postaci wykresu, jak to uwidoczniono na rysunku 2. Widzimy tam oddzielne rysunki dla $r = 1, 2, 3-5, 6-10, > 10$. Dla każdego z tych r mamy wykres wielkości $\mu_k^{(r)}$ dla $k = 1, 2, 3-5, 6-10, > 10$. Diagramy takie sporządzono dla prób z *Przedwiośnia* i *Ludzi bezdomnych* (rys. 2a) oraz *Potopu* i *Quo vadis* (rys. 2b). Na każdym rysunku przedstawiono dane z dwu utworów tego samego pisarza.

Uderzające jest to, że dla tego samego r diagramy wykazują bardzo podobny przebieg. I tak, dla $r = 1$ we wszystkich czterech przypadkach $\mu_1^{(1)}$ jest większe od jedności, następnie maleje systematycznie ze wzrostem k i dla wartości $k > 10$ osiąga minimum $\mu_{>10}^{(1)}$ mniejsze od jedności. Dla $r = 2$ maksimum $\mu_k^{(2)}$ wypada dla $k = 2$, następnie wartości funkcji maleją i przyjmują znowu minimum dla $k > 10$. Dla $r = 3-5$ we wszystkich czterech przypadkach mamy minimum dla $k = 1$, przy czym $\mu_1^{(3-5)} < 1$, następnie wartości funkcji wzrastają, dla $k = 3-5$ mamy już $\mu_{(3-5)}^{(3-5)} > 1$, dalej u Sienkiewicza wartości funkcji $\mu^{(3-5)}$ wzrastają, natomiast u Żeromskiego maleją. Przebieg funkcji μ dla $k = 6-10$ i $k > 10$ jest również podobny. Wszędzie spotykamy wartości większe od jedności dla $k = 1$, następnie wartości funkcji maleją, po czym dla dużych k następuje pewien nieznaczny wzrost funkcji, różny u Żeromskiego i Sienkiewicza. W *Potopie* i *Quo vadis* wzrost ten jest nieznaczny i wartość funkcji dla $k > 10$ w obu badanych utworach jest mniejsza od jedności, natomiast

Rys. 2. Wykres wartości $\mu_k^{(r)}$

w *Przedwiośniu* i *Ludziach bezdomnych* wzrost ten jest większy i mamy nawet $\mu_{\geq 10}^{(>10)} > 1$.

Wiedząc, że $\mu_k^{(r)} > 1$ oznacza przyciąganie się określeń o r i k elementach, podczas gdy $\mu_k^{(r)} < 1$ odpowiada ich odpychaniu się, możemy opisanemu powyżej przebiegowi funkcji $\mu_k^{(r)}$ nadać ściśle określony sens językowy i wyciągnąć odpowiednie wnioski co do praw rządzących sąsiedztwem określeń o danej długości.

Wyniki powyższe można również ująć w następujący schemat:

Jeśli stosunek $\mu_k^{(r)} = \frac{p(k|r)}{p(k)} \geq 1$, to oznaczam go symbolem „+”
jeśli $\mu_k^{(r)} = \frac{p(k|r)}{p(k)} < 1$, to oznaczę go „-”.

Otrzymane wyniki można ująć schematycznie za pomocą plusów i minusów, jak to przedstawia przykładowo tablica 3 dla *Quo vadis*.

TABLICA 3

Prawdopodobieństwa warunkowe wystąpienia określenia k -elementowego po określeniu r -elementowym dla *Quo vadis* Sienkiewicza.

„+” oznacza $\frac{p(k|r)}{p(k)} > 1$, „-” oznacza $\frac{p(k|r)}{p(k)} < 1$

$r \backslash k$	1	2	3-5	6-10	> 10
1	+	+	-	-	-
2	-	+	+	-	-
3-5	-	+	+	+	+
6-10	+	-	-	+	+
> 10	+	+	-	-	-

W tablicy tej uderza nas systematyczne występowanie nadwyżek (plusów) na przekątnej (poza ostatnim wierszem). Nadwyżki te są niewielkie i na ogół nieistotne, jeśli rozpatrywać każdą oddzielnie, niemniej jednak powtarzają się regularnie również w innych tabliczkach, których nie zamieszczałam. Muszę jednak dodać, że podobne tabliczki ułożyłam dla *Ludzi bezdomnych*, *Przedwiośnia* i *Potopu*. Wszędzie otrzymałam w trzech pierwszych wierszach na przekątnej plusy. Dopiero w piątej tabliczce obliczonej dla *Zagadnień socjalistycznej industrializacji Polski* Karpińskiego otrzymałam jeden minus i jeden znak równości. Tak więc ilość plusów na wymienionych miejscach wynosi 13 na 15.

Tworząc statystykę

$$u = \frac{|z - Np|}{\sqrt{Np(1-p)}}$$

i podstawiając $z = 13$, $N = 15$, $p = \frac{1}{2}$, otrzymujemy $u = 2,84$, przy czym wiadomo, że $\Pr\{u \geq 2,58\} = 0,01$.

Zjawisko występowania nadwyżek dla określeń o niewielkiej ilości słów możemy więc uznać za istotne. Znaczy to, że określenia słabo rozwinięte sąsiadują również z określeniami słabo rozwiniętymi. Wnioskujemy stąd, że są w książce miejsca, gdzie pisarz używa określeń podobnych co do długości: w jednych fragmentach po określeniach jednowyrazowych występują również jednowyrazowe, gdzie indziej określenia kilkuwyrazowe (trzy do pięć wyrazów) występują również obok kilkuwyrazowych. Inaczej wygląda sprawa, gdy weźmiemy pod uwagę następstwa po określeniach dobrze rozwiniętych. Byłoby rzeczą nużącą, gdyby po jednym długim określeniu następowało znowu drugie kilkunastowyrazowe. Dlatego też długie przystawki są przeplatane krótkimi. Jest to uwidocznione w tablicy 3 występowaniem minusa na ostatnim miejscu w ostatnim wierszu, oraz występowaniem plusa oznaczającego nadwyżkę określeń jednoelementowych po określeniach wieloelementowych.

Tak więc badanie następstw dało następujące wyniki: Istnieje pewna, chociaż niewielka, zależność między liczebnością sąsiednich określeń. Wyraża się ona dodatnią korelacją dla niezbyt licznych określeń ($k \leq 5$). Dla $k > 10$ prawdopodobieństwo wystąpienia określenia dziesięcio lub więcej elementowego po określeniu również kilkunastuelementowym jest różne dla obu badanych pisarzy: U Sienkiewicza, który potrafi określenia rozbudowywać w długie zdania, po określeniach długich następują krótkie, natomiast u Żeromskiego, który nie używa zbyt obszernych określeń, następstwa po długich określeniach są różne, albo nie zależą od poprzedniego (*Ludzie bezdomni*), albo nawet spotykamy pewne nagromadzenie się długich określeń obok siebie, prawdopodobnie dla wywołania odpowiedniego efektu. Może tu właśnie tkwi jedna z przyczyn głębokiej impresyjności stylu Żeromskiego.

5. Ilość słów w określeniu u różnych autorów. Aby stwierdzić do jakiego stopnia ilość słów w określeniu jest cechą indywidualną pisarza, zbadałam dziewięć utworów Żeromskiego i Sienkiewicza. Dla rozkładów ilości słów w określeniu, otrzymanych w sposób opisany w poprzednich punktach pracy, obliczyłam średnią, wariancję, entropię, współczynnik wariacji i współczynnik przyciągania b rozkładu Polya odpowiadającego danemu rozkładowi. Wyniki przedstawione są w tablicy 4.

Jak widać z tablicy, wszystkie średnie dla utworów Żeromskiego są niższe od średnich dla powieści Sienkiewicza. Różnice te są niewielkie, powtarzają się jednak systematycznie. Testując hipotezę nierówności średnich dla obu autorów można otrzymać przy pomocy testu X van der Waerdena [9], że średnie te są istotnie różne na poziomie istotności $\alpha \leq 0,01$. Wynika stąd, że wartości średnie ilości słów w określeniu dyskryminują utwory powieściowe Żeromskiego i Sienkiewicza. Określenia autora *Ogniem i mieczem* są nieco dłuższe niż autora *Syzyfowych prac*.

TABLICA 4

Średnie, wariancje, współczynniki wariacji W , entropie, współczynniki przyciągania b dla różnych utworów prozaicznych

Autor i dzieło	\bar{x}	s^2	W	E	b
Żeromski					
Popioły	2,73	9,31	1,12	0,7455	2,53
Uroda życia	3,07	16,73	1,33	0,7741	3,33
Przedwiośnie	3,18	15,36	1,23	0,8121	2,78
Ludzie bezdomni	3,01	12,13	1,16	0,7847	2,51
Szyfrowe prace	3,21	17,34	1,30	0,8062	3,10
Sienkiewicz					
W pustyni i w puszczy	3,41	19,34	1,30	0,8269	3,48
Rodzina Połanieckich	3,66	23,04	1,32	0,8514	2,88
Potop	3,58	24,36	1,37	0,8459	3,26
Quo vadis	3,68	24,31	1,33	0,8617	3,03
Karpiński					
Zagadnienie socjalistycznej industrializacji Polski	6,11	64,05	1,32	1,0885	2,26

Również wariancje empiryczne s^2 rozkładów Sienkiewicza wypadły wszystkie większe od wariacji rozkładów Żeromskiego. Może to jednak być związane z różnicami między średnimi, mianowicie większa średnia powoduje większy rozrzut. Aby uniknąć tej zależności, obliczamy współczynnik wariacji $W = s/\bar{x}$. Jak widać z tablicy 4 współczynnik wariacji W waha się w granicach 1,12-1,37 niezależnie od autora. W tablicy 4 podano również parametr b rozkładu Polya przybliżającego dany rozkład empiryczny. Parametr ten nazywamy *współczynnikiem przyciągania*. Okazuje się, że waha się on w granicach 2,2-3,5 dla wszystkich autorów badanych, czyli jest praktycznie stały i niezależny od pisarza. Wobec tego możemy przypuszczać, że współczynnik przyciągania b , podobnie jak współczynnik wariacji W , jest pewną charakterystyką, związaną nie z indywidualnym stylem, ale w ogóle ze sposobem pisania, czyli jest pewną charakterystyką językową, a nie stylistyczną. Można by tu wnieść zarzut, że współczynniki b i W wypadły dlatego takie podobne, że i średnie odpowiednich utworów są bardzo zbliżone. Aby tego uniknąć, obliczyłam b i W dla utworu o znacznie większej średniej, mianowicie Karpińskiego *Zagadnienia socjalistycznej industrializacji Polski*. Wartości współczynników b i W wypadły w tych samych granicach.

Powróćmy jeszcze do faktu, że rozkład ilości słów w określeniu nie jest jednostajny, ale że występuje swoiste przyciąganie się określeń. Można stąd wyciągnąć wniosek, że pisarz nie przedstawia wszystkich faktów jednakowo dokładnie. Nie jest tak, żeby ilość skojarzeń była ściśle losowa, przypadkowa. Mamy tu do czynienia ze świadomą organizacją

określeń. Niektóre fakty są przez pisarza zaledwie naznaczone, inne rozbudowane o wiele bardziej niżby to wynikało z rozkładu czysto przypadkowego. Taką organizację możemy również wyrazić matematycznie jako pomniejszanie entropii układu. Wiadomo, że układy o dużym stopniu dezorganizacji, zbliżające się do chaosu, charakteryzują się dużą entropią. Zmniejszanie się entropii świadczy o wprowadzeniu pewnego uporządkowania. Wartości entropii obliczonej jako

$$E = \sum p_i \log_{10} p_i$$

podane są w tabelicy 4. Jest rzeczą ciekawą, że utwory Żeromskiego charakteryzują się mniejszą entropią niż utwory Sienkiewicza. Widzimy jednak, że zmiany entropii idą na ogół równoległe ze zmianami średnich, a więc są to dwie wielkości powiązane ze sobą.

6. Ilość określeń przypadająca na jedno zdanie. Przez jedno zdanie rozumiem zdanie główne z należącymi do niego zdaniami podrzędnymi. Dwa zdania złożone współrzędnie uważam za dwa różne zdania. Rozkład ilości określeń przypadających na jedno zdanie podany jest przykładowo w tabelicy 5. Wzięty jest on z utworu *W pustyni i w puszczy* Sienkiewicza i różni się w sposób znamieny od rozkładu Poissona, co można zobaczyć na rysunku 5. Charakteryzuje się on mianowicie podnormalną wariancją (dla rozkładu na wykresie $\bar{x} = 2,466$, $s^2 = 1,9886$). Zdań o bardzo małej lub bardzo dużej ilości określeń jest o wiele mniej, niżby to wynikało z rozkładu Poissona. Zamiast tego otrzymany empirycznie rozkład charakteryzuje się znaczną kurtozą.

Rozkład ten daje się opisać uogólnionym rozkładem Poissona, będącym rozkładem sumy dwu niezależnych zmiennych losowych: o rozkładzie Poissona i zerojedynkowym.

Sumaryczny rozkład ma postać następującą:

$$(2) \quad P(i) = \begin{cases} e^{-(\bar{x}-\beta_1)} \left[(1-\beta_1) \frac{(\bar{x}-\beta_1)^i}{i!} + \beta_1 \frac{(\bar{x}-\beta_1)^{i-1}}{(i-1)!} \right] & \text{dla } i \geq 1, \\ (1-\beta_1) e^{-(\bar{x}-\beta_1)} & \text{dla } i = 0. \end{cases}$$

Parametr β_1 można oszacować metodą momentów. Mamy $M_1 = \bar{x}$, $M_2 = \bar{x} - \beta_1^2$, skąd znajdujemy $\beta_1 = \sqrt{\bar{x} - M_2}$.

Widzimy, że warunkiem koniecznym należenia do tej klasy rozkładów jest podnormalna wariancja.

Można tu jeszcze dodać, że wzór (2) jest szczególnym przypadkiem innego, ogólniejszego:

$$P(i) = e^{-(\bar{x} - \sum_{a=1}^{\infty} \beta_a)} \cdot \sum_{r=0}^{\infty} (\beta_r - \beta_{r+1}) \cdot \frac{(\bar{x} - \sum_{a=1}^{\infty} \beta_a)^{i-r}}{(i-r)!}.$$

Wyprowadzenie tego wzoru wraz z przykładami znajdzie czytelnik w książce Fuchsa [2].

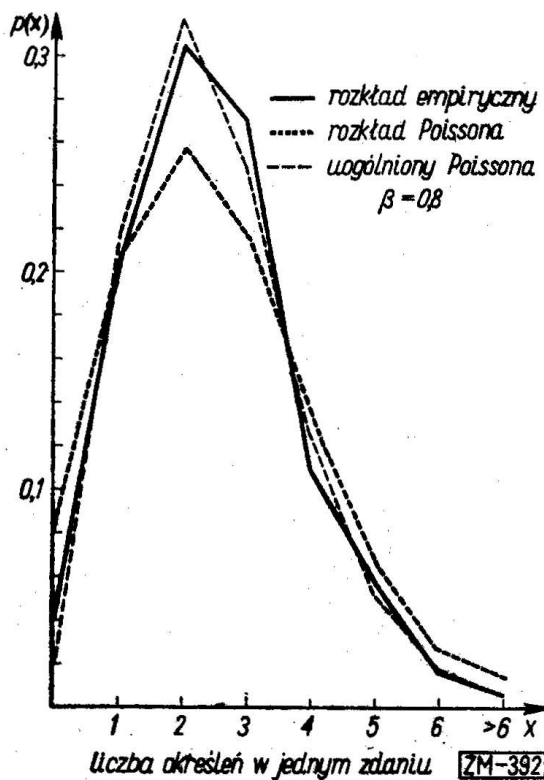
Uzyskane z wzoru (2) przybliżenia rozkładu ilości określeń w zdaniu przedstawiają tablica 5 i wykres 3.

TABLICA 5

Ilość określeń przypadająca na jedno zdanie otrzymana empirycznie — n_{iemp} oraz obliczona teoretycznie n_{iug} według wzoru (2). Utwór *W pustyni i w puszczy*

i	0	1	2	3	4	5	6	> 6	razem
n_{iemp}	26	122	182	161	63	32	11	5	602
n_{iug}	29	126	177	141	79	34	11	5	602

Dane w tablicy 5 podane są dla jednej tylko książki, lecz rozkłady dla innych utworów powieściowych, jak *Syzyfowe prace*, *Ludzie bezdomni*, *Potop*, *Quo vadis* nie odbiegają istotnie od podanego.



Rys. 3. Rozkład liczby określeń przypadającej na jedno zdanie, utwór *W pustyni i w puszczy*

Zgodność danych empirycznych i obliczonych teoretycznie badano testem χ^2 . Otrzymane odchylenia mieszczą się w granicach przypadkowego błędu.

Pozostaje sprawa interpretacji otrzymanego wyniku.

Wzór (2) można rozumieć w sposób następujący: Pewna część (β_1) ogólnej ilości zdań musi mieć co najmniej jedno określenie. Poza tym ilość określeń w zdaniu jest czysto losowa czyli poissonowska. W stylistyce nie są pożądane zdania proste, nierozwinięte, tzw. „zдания nagie”. Tym samym nakłada się na pisarza pewne ograniczenie jego swobody, dowolności w konstruowaniu zdania, mianowicie zdania w przeważającej części powinny być rozwinięte, mieć co najmniej jedno określenie. Ta przeważająca część oznaczona jest w równaniu (2) parametrem β_1 .

Dopiero po tym ograniczeniu występowanie dalszej ilości określeń jest już czysto przypadkowe.

Streszczając wyniki pracy można stwierdzić: Określenia w utworach powieściowych Żeromskiego i Sienkiewicza różnią się jakością: określenia

Żeromskiego są krótsze i bardziej kontrastowe niż określenia Sienkiewicza; nie różnią się natomiast częstością w zdaniu.

Pragnę złożyć podziękowanie prof. J. Perkalowi, prof. J. Łukasiewiczowi oraz dr. J. Woronczakowi za cenne uwagi przy pisaniu i redagowaniu niniejszej pracy.

Prace cytowane

- [1] L. Apostel, B. Mandelbrot, A. Morf, *Logique, Language et Théorie d'information*, Paris 1957.
- [2] W. Fuchs, *Mathematical theory of word-formation*, Aachen 1955.
- [3] M. Greenwood, G. U. Yule, *An inquiry into the nature of frequency distribution*, Journ. of the R. S. S. 83 (1920).
- [4] P. Guiraud, *Les caractères statistiques du vocabulaire*, Paris 1947.
- [5] G. Herdan, *Language as choice and chance*, Groningen 1956.
- [6] H. Kurkowska, S. Skorupka, *Stylistyka polska*, Warszawa 1959.
- [7] B. Mandelbrot, *Mécanique statistique et théorie d'information*, Comptes rendus des Seances de l'Academie des Sciences t. 232.
- [8] *Проблемы кибернетики*, Москва 1960.
- [9] B. L. van der Waerden-E. Nievergelt, *Tafeln zum Vergleich zweier Stichproben mittels X-Test*, Springer Verlag 1955.
- [10] G. U. Yule, *On sentence length as statistical characteristic of style in prose*, Biometrika 30, str. 363.
- [11] — *The statistical study of literary vocabulary*, Cambridge 1944.

Praca wpłynęła 15. 1. 1961

A. БАРТКОВЯК (Вроцлав)

О РАСПРЕДЕЛЕНИИ ВТОРОСТЕПЕННЫХ ЧЛЕНОВ В ПРЕДЛОЖЕНИЯХ ЖЕРОМСКОГО И СЕНКЕВИЧА

РЕЗЮМЕ

В статье исследовано второстепенные члены в качестве некоторых характеристик стиля писателя.

Найдено, что число слов во второстепенном члене выражается через распределение Пуассона (уравнение (1)). Среднее значение распределения и энтропия естественно разны для книг Жеромского и Сенкевича. Число второстепенных членов в одном предложении аппроксимируется уравнением (2) и не отличается у исследованных писателей.

A. BARTKOWIAKOWA (Wrocław)

*ON THE DISTRIBUTION OF COMPLEMENTS IN DESCRIPTIVE
SENTENCES OF ŻEROMSKI AND SIENKIEWICZ*

SUMMARY

In the paper the complements of subject and predicate as certain features of literary style are investigated. It is found that the number of words in a complement is governed by the Polya distribution, given as formula (1). The mean number of words in a complement and the entropy taken for different writers Sienkiewicz and Żeromski show a significant difference.

The number of complements in a sentence can be expressed by the distribution given by formula (2). For the authors mentioned above no significant difference has been found.
