ROCZNIKI POLSKIEGO TOWARZYSTWA MATEMATYCZNEGO
Seria II: WIADOMOŚCI MATEMATYCZNE VII (1963)

Ministerstwo Nauki i Szkolnictwa Wyższego

A. BARTKOWIAKOWA i L. ZUBRZYCKA (Wrocław)

O polskich badaniach języka przy użyciu metod matematycznych

Język można traktować jako populację(1) różnego rodzaju elementów. Jako elementy w języku można rozważać fonemy(2), głoski, litery, sylaby, wyrazy, zdania itp. Możemy więc raz rozpatrywać język jako populację fonemów, innym razem jako populację wyrazów lub zdań. Możemy pytać, jakie są te elementy i jaką funkcję logiczną spełniają, np. słowa i końcówki w zdaniach lub zdania w tekstach. Te pytania nabrały nowej aktualności w związku z pracami nad przekładem maszynowym z jednego języka na drugi.

Dość interesujące jest pytanie, z jakimi częstościami występują w danym języku jego elementy. Ciekawość ta może mieć rozmaite praktyczne motywy. Można pytać o najczęściej używane wyrazy, co ma znaczenie przy opracowywaniu słowników do nauki języka, lub o najczęściej występujące litery, co jest ważne dla reformy systemu stenograficznego czy też układu klawiatury maszyny do pisania.

Od dwudziestu lat badania takie prowadzone są na całym świecie z wielkim nakładem pracy. O niektórych z nich informuje artykuł R. L. Dobruszyna *Metody matematyczne w lingwistyce*, zamieszczony w Wiadomościach Matematycznych. (VI 1963)

Celem tego artykułu jest podanie informacji o polskich pracach poświęconych badaniu języka przy użyciu metod matematycznych, a w szczególności statystyki matematycznej. Badano różne elementy języka począwszy od najdrobniejszych — jak litery i fonemy — poprzez sylaby i słowa, aż do całych zdań. W tej też kolejności zamierzamy omawiać poszczególne prace.

Przy wyznaczaniu częstości elementów języka pojawiają się różne metodyczne kłopoty, zależne przede wszystkim od tego, co uważamy za

(1) Przez populację rozumiemy tu zbiorowość wszystkich elementów podlegających obserwacji w danym badaniu.

(2) Fonem jest to głoska lub grupa wzajemnie wyłączających się głosek, która w danym dialekcie (względnie języku) i w danym czasie kontrastuje z jakąkolwiek inną głoską. (Patrz W. Jassem, *Wymowa angielska*, PZWS, Warszawa 1951, str. 184.)

język, a także od tego, jakie upraszczające założenia przyjmujemy. Jeśli powiemy, że język to jest tekst jednej książki, wówczas istnieje możliwość uzyskania, za pomocą stuprocentowych badań, kompletnej odpowiedzi na pytanie, z jakimi częstościami występują poszczególne elementy w tej książce. W tym kontekście nie ma żadnej problematyki statystycznej w sensie rozważania związku między próbką i populacją. Problem taki pojawia się wówczas, jeśli z jakichś powodów nie możemy robić stuprocentowych badań (porównaj [10]). Jeśli nasze stwierdzenia o częstości elementów chcemy odnosić nie do konkretnego tekstu, ale np. do języka powieściowego jakiejś epoki, to przebadane teksty, jakkolwiek duże by one nie były, z konieczności musimy traktować jako próbki. A jeśli tak, to powstaje pytanie, czy próbka dobrze reprezentuje swoją populację macierzystą? Istotne są również pytania dotyczące różnic między tekstami pod względem obrazu częstościowego elementów. Takimi zagadnieniami zajmowały się M. Steffen i L. Zubrzycka.

M. Steffen [13] postawiła sobie zadanie wyznaczenia częstości występowania fonemów i głosek w języku polskim. Materiał statystyczny zebrany przez M. Steffen obejmował 50000 głosek. Ponieważ były przypuszczenia, że teksty mogą się różnić między sobą co do częstości poszczególnych elementów, do próby wzięto teksty z 10 różnych źródeł. Była to proza literacka, publicystyczna i naukowa. Z każdego źródła wzięto tekst zwarty, obejmujący 5000 głosek. Obliczono częstości występowania poszczególnych głosek i całych fonemów dla każdego tekstu z osobna i w grupach otrzymanych przez kolejne łączenie (I+II, I+II+III, itd.). Jako ostateczne przyjęto częstości fonemów i głosek obliczone na całym materiale. Błąd oszacowania częstości obliczono według wzoru Bernouilliego $\sqrt{pq/n}$, gdzie p oznacza prawdopodobieństwo pojawienia się danej głoski, $q = 1-p$, a n jest licznością próbki. Zgodnie z tym wzorem dla próbki liczącej 50000 głosek otrzymano, że częstości głosek są wyznaczone z błędem nie przekraczającym 0,1%.

M. Steffen-Batogowa [14] zajmowała się również badaniem częstości występowania wyrazów w zależności od ilości sylab. Badania te były podjęte w celu uzyskania podstaw do opracowania polskich testów do badań nad wyrazistością i zrozumiałością mowy. Chodziło o to, żeby ułożyć listę słów, w której poszczególne fonemy występowałyby z częstościami takimi jak w języku polskim, a ponadto, aby była zachowana proporcja występowania wyrazów o różnych ilościach sylab. Zestawienia podane w tej pracy oparte są o badania statystyczne obejmujące 31788 wyrazów ze współczesnej literatury pięknej, dzienników i prac naukowych; wykorzystane zostały również materiały z pracy poprzedniej [13].

L. Zubrzycka [21] liczyła częstości występowania w języku polskim poszczególnych liter i grup spółgłoskowych. Badania te były podjęte dla potrzeb stenografii.

W systemach stenograficznych jednym ze sposobów skróconego zapisu jest zastępowanie często używanych frazesów i wyrazów krótkimi symbolami. Dlatego potrzebne jest wyłowienie najczęściej pojawiających się zwrotów, frazesów lub wyrazów. Innym sposobem ulepszania systemu stenograficznego jest przypisywanie najczęściej występującym literom najszybciej piszących się znaków alfabetu stenograficznego. W tym celu potrzebne jest uporządkowanie liter według częstości ich występowania w tekstach polskich. W związku z tym, że szybkiego stenografowania wymaga się w specjalnych okolicznościach, nasuwa się pytanie, czy stenografowane teksty nie wyróżniają się pod względem częstości liter itp. od tekstów literackich lub gazetowych, jakie zazwyczaj bierze się do statystyk językowych. Praca L. Zubrzyckiej miała na celu zbadanie, jak wielkie są te rozbieżności, jeśli w ogóle występują. W tym celu wzięto cztery teksty o różnej tematyce, po 1000 wyrazów każdy, policzono częstości liter i grup spółgłoskowych w każdym tekście z osobna i porównano te teksty co do częstości występowania badanych elementów. Dla porównania częstości liter w różnych tekstach użyto testu χ^2. Okazało się, że teksty te różnią się statystycznie istotnie co do częstości poszczególnych liter i grup spółgłoskowych. Potwierdza to przypuszczenie, że dla celów stenografii dobór odpowiednich tekstów jest potrzebny.

W pracy L. Zubrzyckiej podano ponadto sposób, w jaki można z opracowanej próbki uzyskać oszacowanie masy językowej niezbędnej do opracowania, w celu uporządkowania liter i grup spółgłoskowych według częstości. Mianowicie stawia się pytanie: o ile trzeba powiększyć próbkę, by takie różnice częstości, jakie zaobserwowano w tych czterech tekstach, można uważać za istotne? Oto odpowiedź: jeśli uporządkowanie ma być dokładne, to do uporządkowania liter potrzebne jest próbka obejmująca 100000 wyrazów, a do uporządkowania grup spółgłoskowych — 3200000 wyrazów.

L. Zubrzycka [22] rozważała również częstości znaków pisarskich i par znaków pisarskich w związku z dostosowaniem klawiatury maszyny do pisania do struktury języka.

Pisząc na maszynie, uderzamy w klawisze zgodnie z kolejnością liter w tekście. Czas między dwoma kolejnymi uderzeniami w klawisze zależny jest w głównej mierze od wzajemnego położenia tych klawiszy i na przykład jest mniejszy, gdy te klawisze są w polu pracy różnych rąk, niż gdy są w polu pracy różnych palców jednej ręki. Pożądane jest takie ułożenie liter na klawiaturze, żeby najdogodniejszym i najszybszym typom uderzeń przypisać najczęstsze pary znaków pisarskich. W podejmowanych dotychczas próbach [15] dostosowania klawiatury maszyny do pisania do struktury języka uwzględniano tylko częstości znaków pisarskich, a nie badano systematycznie ich następstwa w tekście. Pierwszy w Polsce F. Kotas [9] wyraźnie formułuje postulaty wiążące układ

klawiatury z częstością par znaków pisarskich sąsiadujących w tekście i popiera je wynikami eksperymentów przeprowadzonych w gronie uczennic po 50-godzinnym kursie pisania na maszynie. W pracy L. Zubrzyckiej przyjmuje się z pewnymi modyfikacjami zaproponowany przez F. Kotasa układ postulatów, który w sposób jednoznaczny przydziela litery poszczególnym klawiszom. W pracy przedstawiona jest metoda pozwalająca w oparciu o tablicę częstości par sąsiednich znaków piarskich skonstruować klawiaturę realizującą te postulaty. Głównym postulatem jest zapewnienie jak największej naprzemienności uderzeń, to znaczy, aby najczęściej następny klawisz uderzany był inną ręką niż poprzedni. Matematycznie zadanie sprowadza się do takiego jednoczesnego przenumerowania kolumn i wierszy w macierzy podającej prawdopodobieństwa pojawienia się w tekście poszczególnych par sąsiednich znaków pisarskich, aby uzyskać maksymalną sumę wyrazów w prawej górnej i lewej dolnej ćwiartce tablicy. W pracy podano opracowany tą metodą projekt klawiatury, przy konstrukcji którego wykorzystano statystykę 10000 par znaków pisarskich z tekstu naukowego i 10000 par znaków pisarskich z tekstu gazetowego. Podane jest również porównanie nowego projektu z wcześniejszymi projektami klawiatury, a mianowicie z projektem F. Kotasa [9], z projektem I. Gogut [9], z polskim układem uniwersalnym [15] i z będącą w powszechnym użyciu klawiaturą uniwersalną [12]. Wszyscy autorzy piszący o klawiaturze zgadzają się co do tego, że pożądane jest zapewnienie jak największej naprzemienności uderzeń. Na uwagę zasługuje fakt, że stare projekty nie różnią się pod tym względem od używanej powszechnie klawiatury uniwersalnej, a klawiatura L. Zubrzyckiej zapewnia o $10^0/_0$ większą częstość uderzeń naprzemiennych.

W dotychczas omawianych pracach, tam gdzie próbowano ocenić dokładność oszacowania pewnych prawdopodobieństw przez zaobserwowane częstości lub porównać istotność zaobserwowanych różnic między częstościami, przyjmowano upraszczające założenia, że badane teksty można traktować jako próbki losowe, w tym sensie, że kolejne litery tekstu jak gdyby były losowane niezależnie z urny o stałym składzie. Gdyby tak było, to zgubionej lub zamienionej litery nie można by w żaden sposób na podstawie kontekstu odtworzyć. Natomiast wiadomo, że w prawdziwym tekście jest to przeważnie nie bardzo trudne zadanie. Jeśli ktoś napisze „czfartek", to od razu widać, że powinno być „czwartek". Dlatego, chcąc mieć bardziej realistyczny opis języka, należałoby o nim mówić raczej jako o populacji tekstów złożonych z elementów, aniżeli jako o populacji tych elementów. Matematycznie mówiąc, chodzi tu o traktowanie tekstu jako realizacji pewnego procesu stochastycznego, w którym kolejne elementy nie są niezależne. Ponadto, na tle owej możliwości odtwarzania zgubionych lub zamienionych liter w tekście chciałoby się mieć jakąś numeryczną miarę „wielkości informacji" przy-

padającej na jedną literę tekstu. Taką miarą jest entropia. Ma ona wiele własności, jakich w zupełnie naturalny sposób można by wymagać od miary ilości informacji (patrz [25]). W przypadku, gdy możliwe są stany $A_1, A_2, \ldots, A_n$ z prawdopodobieństwami $p_1, p_2, \ldots, p_n$, entropię obliczamy według wzoru

$$(1) \qquad -\sum p_i \log_2 p_i.$$

Jeśli mamy 32 litery, to możemy z nich utworzyć 32^2 par liter, 32^3 trójek liter, ..., 32^n układów po n liter. Można zapytać, z jakimi prawdopodobieństwami pojawiają się dla danego n w danym języku poszczególne układy po n liter i obliczyć entropię na podstawie tych prawdopodobieństw. Gdy tę entropię podzielimy przez n, otrzymamy entropię przypadającą na jedną literę obliczoną z prawdopodobieństw układów po n liter. Granica, do której zmierza ta liczba przy rosnącym n, nazywa się entropią przypadającą na jedną literę w danym języku. Trudno zrobić statystykę grup liczących więcej niż dwie litery i dlatego jest praktycznie niemożliwe wyznaczenie entropii przypadającej na jedną literę w danym języku. Dla różnych języków wyznaczano entropię przypadającą na jedną literę obliczoną przeważnie tylko z prawdopodobieństw występowania poszczególnych liter i par liter.

Entropia obliczona na podstawie prawdopodobieństw występowania poszczególnych liter byłaby równa entropii przypadającej na jedną literę w danym języku jedynie wtedy, gdyby nie było związków między literami. Wobec istnienia tych związków, entropia przypadająca na jedną literę w danym języku jest od niej mniejsza. O jakości związków między literami w tekście możemy sądzić obserwując, o ile entropia przypadająca na jedną literę obliczona na podstawie prawdopodobieństw występowania par liter jest mniejsza od entropii obliczonej na podstawie prawdopodobieństw występowania poszczególnych liter.

Dla języka polskiego entropię przypadającą na jedną literę obliczali T. Ekiert i J. Seidler [5]. Częstości liter i par liter wyznaczyli z materiału obejmującego 40000 par liter.

Entropia obliczona zgodnie z wzorem (1) na podstawie zaobserwowanych częstości dla różnych tekstów wynosi odpowiednio:

dla tekstu gazetowego I 4,51,
dla tekstu gazetowego II 4,56,
dla tekstu gazetowego III 4,77,
dla tekstu literackiego 4,56, czyli średnio 4,60.

Entropia obliczona na podstawie zaobserwowanych częstości par liter z tych czterech tekstów łącznie wynosi 4,42.

L. Zubrzycka podaje w [22] tablicę częstości par 41 znaków pisarskich, w której potraktowano jako znaki pisarskie 32 litery, znaki prze-

stankowe: kropkę, przecinek, dwukropek, średnik, nawias, myślnik oraz odstęp między wyrazami i przeniesienie do nowego wiersza. Entropia przypadająca na jeden znak pisarski obliczona na podstawie tablicy częstości znaków pisarskich wynosi 4,67, a na podstawie częstości par znaków pisarskich wynosi 4,11. Zauważmy, że różnica entropii przy przejściu od częstości liter do częstości par u T. Ekiert i J. Seidlera jest wyraźnie mniejsza niż u L. Zubrzyckiej, co jest dość nieoczekiwane.

Oprócz badania częstości występowania poszczególnych elementów w języku można pytać o rozkłady częstości pewnych charakterystyk liczbowych tych elementów, np. ilości sylab w wyrazie, jak to robiła M. Steffen-Batogowa [14] we wspomnianej poprzednio pracy, albo zrobić jeszcze krok dalej i rozważać numeryczne charakterystyki rozkładów badanych charakterystyk, np. entropię obliczać dla rozkładu długości sylabicznej wyrazów (ilości sylab w wyrazie), a nie dla częstości samych wyrazów. Zajmuje się tym praca A. Bartkowiakowej i B. Gleichgewichta [4].

Autorzy badali ilość sylab w słowie w polskich utworach prozaicznych z różnych okresów, począwszy od Reja i Górnickiego, poprzez Prusa i Żeromskiego, aż do współczesnych — Dygata i Lema. Uzyskany rozkład ilości sylab w słowie charakteryzowano za pomocą dwóch liczb: wartości średniej oraz entropii rozkładu obliczonej według wzoru

$$E = \sum p_i \log p_i, \tag{2}$$

gdzie jako p_i brano zaobserwowaną częstość występowania słowa o długości sylabicznej i. Obydwa te wskaźniki wzrastają chronologicznie, a na wykresie średnia-entropia, tzn. na wykresie, w którym na jednej osi odkładamy średnią długość sylabiczną wyrazu, a na drugiej entropię rozkładu długości sylabicznej, punkty reprezentujące poszczególne utwory układają się w przybliżeniu wzdłuż prostej. Średnie, obliczone z wypowiedzi dialogowych, są odpowiednio mniejsze niż te, które otrzymano z fragmentów opisowych tekstu.

Instytut Badań Literackich postanowił opracować słownik szesnastowiecznych wyrazów. Okazało się, że jest rzeczą technicznie niemożliwą zbadanie całego dostępnego materiału i trzeba było zdecydować się jedynie na pewną część (20% wszystkich dostępnych tekstów). O ile jednak taka próbka będzie reprezentacyjna dla całej epoki?

W. Kuraszkiewicz i J. Łukaszewicz [11] badali szybkość narastania liczby różnych wyrazów w miarę powiększającej się długości tekstu. Porównując dane dla materiału opracowanego na podstawie całego utworu oraz jego próbek doszli oni do następujących praktycznych wniosków: ilość różnych wyrazów w 20% próbce rośnie z początku bardzo szybko, a później już bardzo powoli; 20% próbka dla dłuższych utworów (ponad 300000 liter) jest dość reprezentacyjna, bo wyczerpuje przeszło 50%

różnych wyrazów zawartych w tekście; lepiej jest badać $20^0/_0$ materiału badając co piątą stronę, aniżeli jeden zwarty fragment tekstu; ilość różnych wyrazów z $20^0/_0$ próbki można powiększyć losując drobniejsze fragmenty niż strony, np. wiersze. Wnioski te zastosowano w pracowniach leksykograficznych Instytutu Badań Literackich przy dalszych badaniach.

Przypuśćmy teraz, że wyrazy występujące w jakimś tekście numerujemy według częstości ich występowania, od najczęściej do najrzadziej występujących. Niech P_r oznacza częstość r-tego wyrazu. Mandelbrot [11] odkrył, że związek między r i P_r daje się opisać przez dwuparametrową rodzinę rozkładów prawdopodobieństwa

$$p_r = p(r+\varrho)^{-B}. \tag{3}$$

Parametry ϱ i B tego rozkładu proponuje on używać jako wskaźników bogactwa słownikowego. Jednak szacowanie tych parametrów na podstawie próbki jest dość trudne. Mianowicie z próbki możemy wyznaczyć tylko częstości P_r będące przybliżeniem prawdopodobieństw p_r, a następnie musimy dobrać parametry ϱ i B tak, aby uzyskać dobrą zgodność prawdopodobieństw p_r z zaobserwowanymi częstościami P_r. Na ogół wyznacza się te parametry pewnymi metodami graficznymi. J. Woronczak [18] zaproponował inny sposób szacowania tych parametrów, odznaczający się prostotą i przejrzystością.

Niech $\zeta(z, q) = \sum_{n=0}^{\infty} \frac{1}{(n+q)^z}$ będzie uogólnioną funkcją ζ Riemanna. Przyjmując $q = \varrho+1$, $z = B$ otrzymujemy następujące równości:

$$\sum_{r=1}^{\infty} p_r^2 = \frac{\zeta(2z, q)}{\zeta^2(z, q)} = Z_2(z, q),$$

$$\sum_{r=1}^{\infty} p_r^3 = \frac{\zeta(3z, q)}{\zeta^3(z, q)} = Z_3(z, q).$$

Niech teraz s_i oznacza ilość występowania i-tego słowa w próbie o liczności N, $i = 1, 2, \ldots, k$, gdzie k jest liczbą różnych słów w próbie. Wówczas można udowodnić, że

$$II_2 = \frac{\sum_{i=1}^{k} s_i(s_i-1)}{N(N-1)} = \frac{\sum_{i=1}^{k} s_i^2 - N}{N(N-1)}$$

jest nieobciążonym estymatorem(3) liczby $\sum_{r=1}^{\infty} p_r^2$, a

(3) Estymatorem nazywa się w statystyce matematycznej funkcję przyporządkowującą wynikom obserwacji wartości liczbowe.

$$\Pi_3 = \frac{\sum_{i=1}^{k} s_i(s_i-1)(s_i-2)}{N(N-1)(N-2)} = \frac{\sum_{i=1}^{k} s_i^3 - 3\sum_{i=1}^{k} s_i^2 + 2N}{N(N-1)(N-2)}$$

jest nieobciążonym estymatorem liczby $\sum_{r=1}^{\infty} p_r^3$.

J. Woronczak proponuje wyznaczać z i q, a zatem ϱ i B z układu równań

$$Z_2(z, q) = \Pi_2, \quad Z_3(z, q) = \Pi_3.$$

Aby ułatwić rozwiązywanie tego układu równań, sporządza on wspólny nomogram siatkowy dla funkcji $Z_2(z, q)$ i $Z_3(z, q)$.

Proponowano również inne wskaźniki bogactwa słownikowego, oparte na częstości występowania poszczególnych słów w badanym tekście (G. Herdan [7] i G. U. Yule [24]). Probabilistyczne wyjaśnienie zjawiska, że wartości liczbowe tych charakterystyk praktycznie nie zależą od wielkości próbki, podał S. Zubrzycki [23].

Przechodząc do jeszcze większych jednostek tekstowych należy wymienić składniki syntaktyczne w zdaniu, jak podmiot, orzeczenie, dopełnienie, okolicznik i przydawka. Ilość słów w składniku syntaktycznym, jak również ilość takich składników w jednym zdaniu, jest pewną cechą stylistyczno-składniową, która może charakteryzować zarówno autora, jak i rodzaj literacki. M. Kniagininowa [8] badała takie cechy dla różnych gatunków literackich twórczości Żeromskiego: dramatu, powieści, prozy, poetyki i publicystyki. Okazało się, że we wszystkich badanych utworach najbardziej zmienna jest długość przydawek (ilość słów w przydawce). Znacznie węższy obszar zmienności mają dopełnienia i okoliczniki. M. Kniagininowa wykazuje, że pod względem długości składników syntaktycznych zachodzi istotna różnica między dramatem a wszystkimi pozostałymi gatunkami literackimi, między powieścią a prozą poetycką i publicystyką. Nieistotna jest natomiast różnica pod tym względem między prozą poetycką i publicystyką. Różnice między średnimi długościami przydawki są dla wszystkich gatunków literackich Żeromskiego istotne.

A. Bartkowiakowa [1] badała tzw. określenia do podmiotu i orzeczenia, czyli okoliczniki, przydawki i dopełnienia wzięte łącznie dla danego zdania. Rozkłady długości określeń liczonej w słowach dały się przybliżyć przez tzw. graniczny rozkład Pólyi (ujemny dwumianowy). Znaczy to, że całe szeregi rozdzielcze długości określeń, a więc i utwory, można scharakteryzować dwiema liczbami, mianowicie dwoma parametrami rozkładu Pólyi. Badania te były robione dla utworów powieściowych Żeromskiego i Sienkiewicza. Określenia sąsiadujące ze sobą wykazują pewne „przyciąganie", tzn. po krótkich określeniach występują

częściej dalsze krótkie określenia. Między długością określenia u Żeromskiego i Sienkiewicza istnieją pewne (zresztą niewielkie) różnice. Sienkiewicz używa określeń nieco dłuższych niż Żeromski. Częstość określeń natomiast, tzn. średnia ich liczba przypadająca na jedno zdanie, nie różni się dla obu badanych pisarzy.

A. Bartkowiakowa [2] badała również rozkłady ilości zdań złożonych współrzędnie i podrzędnie. Można je też wyrazić granicznym rozkładem Pólyi. Znaczy to, że mechanizm powstawania określeń ze słów i zdań złożonych ze zdań pojedynczych jest podobny. Kolejność zdań prostych i złożonych również nie jest przypadkowa, np. jest wyraźnie mniej serii o długości 1 składających się ze zdań prostych i złożonych. Świadczy to, że autor nie lubi używać raptownych przejść: zdanie złożone — proste — złożone, lub odwrotnie, lecz że tendencja do posługiwania się jednym rodzajem zdań (prostymi lub złożonymi) utrzymuje się jakiś czas w trakcie pisania.

Metod statystyki matematycznej używano także w wersologii. Do prac z tego zakresu należy praca J. Woronczaka [17], który prowadził dość rozległe badania nad utworami z okresu średniowiecza i odrodzenia. Badał on długość sylabiczną zdań lub wersów oraz rozkłady prawdopodobieństw i współczynniki wariacji tych długości, a także ilości kolejno po sobie następujących wersów o tej samej długości sylabicznej. Wykazuje on dobrą zgodność rozkładu długości sylabicznej wersu z rozkładem gamma. Współczynniki wariacji długości sylabicznej dla zdań pochodzących z utworów wierszowanych są mniejsze, niż dla utworów prozaicznych. Jako ciekawy szczegół podajemy, że obliczył on również współczynnik wariacji dla niektórych wierszy Przybosia, i to takich, które pierwotnie były wydane w formie wierszowanej, a w następnym wydaniu sam pisarz zmienił szyk na prozaiczny. Współczynnik wariacji dla tych utworów wypada w granicach odpowiadających prozie.

Na zakończenie jeszcze parę słów o tzw. zagadnieniu autorstwa. Zdarza się, że jakiś utwór został wydany anonimowo lub pod zmienionym celowo nazwiskiem. Obszerna literatura dotyczy zagadnienia autorstwa „De imitatione", który to traktat oprócz uznawanego ogólnie Tomasza à Kempis przypisywany jest przez niektórych historyków Kościoła innym osobom, np. Janowi Gersonowi. Podobnie w naszej literaturze w wydaniu zbiorowym dzieł Reja pt. „Zwierciadło" dołączona została biografia Reja pt. „Żywot i sprawy poćciwego ślachcica polskiego M. Reja z Nagłowic, który napisał Andrzej Trzecieski, jego dobry towarzysz, który wiedział wszystkie sprawy jego".

Mimo, iż w tytule wymieniony jest inny autor, utwór ten przypisywany jest przez większość historyków literatury samemu Rejowi, który podszył się jakoby pod nazwisko przyjaciela.

- Są wreszcie bezimienne rękopisy średniowieczne, pochodzące z różnych szkół. Wyłania się więc zagadnienie oszacowania stopnia pokrewieństwa tych rękopisów. Zagadnienie takie można badać porównując drogą szczegółowej analizy odpowiednie zwroty, wyrażenia, słownictwo.

J. Woronczak [16], porównując częstość występowania różnych form czy wyrazów w wersach o różnej długości, próbował uzyskać argumenty przemawiające za istnieniem przeróbek pochodzących od kopistów w bajkach Biernata z Lublina.

Językoznawca oblicza nieraz kilka wskaźników liczbowych dla danego utworu. Wówczas utwór można matematycznie traktować jako punkt w przestrzeni wielowymiarowej. Mając układ takich punktów reprezentujących różne utwory możemy, przyjmując odpowiednią konwencję, obliczyć wzajemne odległości tych utworów i na tle tych odległości, przy pomocy taksonomii wrocławskiej (porównaj [6]) śledzić wzajemne ich sąsiedztwa sporządzając dendryt. Można się spodziewać, że utwory tego samego autora wystąpią w takim dendrycie w zwartej grupie. Dlatego też postać dendrytu sporządzonego dla grupy utworów niewątpliwie pochodzących od pewnego autora oraz utworu li tylko przypisywanego temu autorowi może być argumentem przemawiającym za lub przeciw wątpliwemu autorstwu.

Badania tego rodzaju przeprowadziła A. Bartkowiakowa [3] porównując „Żywot i sprawy M. Reja" z kilkoma innymi utworami Reja i jednym utworem A. Trzecieskiego. W sporządzonym dendrycie tych utworów „Żywot i sprawy M. Reja" znalazł się pomiędzy grupą traktatów moralnych Reja a grupą ksiąg „Żywota człowieka poczciwego", podczas gdy utwór Trzecieskiego leżał poza grupą utworów związanych z Rejem. Można w tym widzieć ważki argument za autorstwem Reja.

J. Woronczak [19] sporządził dendryt szesnastowiecznych mszałów, opierając obliczanie odległości między mszałami na występowaniu określonych typów sekwencji. Zaobserwował zgodność grup, wyróżnionych przy pomocy dendrytu, z podziałem terytorialnym.

J. Woronczak [20] uszeregował również 60 zachowanych przekazów Bogurodzicy przy pomocy dendrytu opartego na występowaniu wariantów pewnych słów lub szyku słów.

Zreferowałyśmy tu sporo prac, w których próbowano rozwiązywać zadania dotyczące języka przy użyciu metod matematycznych. O wzrastającym zainteresowaniu tym kierunkiem badań świadczyć może fakt zorganizowania Międzynarodowej Konferencji Poetyki (Jabłonna, 15-25 września 1961 r.) poświęconej stosowaniu metod matematycznych. Należy oczekiwać dalszych wyników badań prowadzonych w tym kierunku.

Prace cytowane

[1] A. Bartkowiakowa, *O rozkładzie określeń w zdaniach opisowych Żeromskiego i Sienkiewicza*, Zastosowania Matematyki 6 (1962), str. 287-303.

[2] — *Statystyka zdań współrzędnych i podrzędnych na przykładzie zdań opisowych Żeromskiego i Sienkiewicza*, Zastosowania Matematyki (w druku).

[3] — *O pewnym sposobie ustalania autorstwa anonimowych tekstów*, referat wygłoszony na Międzynarodowej Konferencji Poetyki, Jabłonna 1961.

[4] — i B. Gleichgewicht, *O długości sylabicznej wyrazów w tekstach autorów polskich*, Zastosowania Matematyki 6 (1962), str. 309-319.

[5] T. Ekiert i J. Seidler, *Badanie struktury słów języka polskiego*, maszynopis nieopublikowany.

[6] K. Florek, J. Łukaszewicz, J. Perkal, H. Steinhaus, S. Zubrzycki, *Taksonomia Wrocławska*, Przegląd Antropologiczny 17 (1951), str. 193-211.

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

[8] M. Kniaginińowa, *Próba zastosowania metod statystycznych w badaniach stylistyczno-składniowych*, Język Polski 13, Nr 2 (1962), str. 92.

[9] F. Kotas, *Projekt nowej racjonalnej klawiatury*, maszynopis nieopublikowany.

[10] W. Kuraszkiewicz i J. Łukaszewicz, *Ilość różnych wyrazów w zależności od długości tekstu*, Pamiętnik Literacki 1 (1951), str. 168-182.

[11] B. Mandelbrot, *Structure formelle des textes et communication*, Word 10 (1954), str. 1-27.

[12] *Polska Norma, PN-58/F-02000, maszyny do pisania. Układy znaków na klawiaturach.*

[13] M. Steffen, *Częstość występowania głosek polskich*, Biuletyn Polskiego Towarzystwa Językoznawczego 16 (1957), str. 145-164.

[14] M. Steffen-Batogowa, *Polskie testy do badań nad wyrazistością i zrozumiałością mowy*, Przegląd Telekomunikacyjny 3 (1962), str. 82-89.

[15] T. Szeffel, *Zagadnienia normalizacji klawiatury w maszynach do pisania*, Wiadomości PKN, XIX. 5 (1951), str. 321-329.

[16] J. Woronczak, *Z badań nad wierszami Biernata z Lublina*, Pamiętnik Literacki 3 (1958), str. 97-118.

[17] — *Statistische Methoden in der Verslehre*, referat wygłoszony na Konferencji Poetyki, Warszawa 1960.

[18] — *Metody obliczania wskaźników bogactwa słownikowego tekstów*, referat wygłoszony na Międzynarodowej Konferencji Poetyki, Jabłonna 1961.

[19] — *Repertuar sekwencyjny w Polsce na przełomie XV i XVI wieku.*

[20] — E. Ostrowska, H. Feicht, *Bogurodzica*, Wrocław-Warszawa-Kraków 1962.

[21] L. Zubrzycka, *Uwagi o statystycznym badaniu struktury języka polskiego dla potrzeb stenografii*, maszynopis.

[22] — *O dostosowywaniu klawiatury maszyny do pisania do struktury języka*, Zastosowania Matematyki 6 (1963), str. 419-439.

[23] S. Zubrzycki, *Concerning Yule's characteristic of style*, Zastosowania Matematyki 4 (1959), str. 328-331.

[24] G. U. Yule, *Statistical study of vocabulary*, Cambridge 1944.

[25] Д. К. Фаддеев, *К понятию энтропии конечной вероятностной схемы*, Успехи Математических Наук 11(1956), str. 227-231.