Władysław Kuraszkiewicz,Józef Łukaszewicz

Ilość różnych wyrazów w zależności od długości tekstu

Pamiętnik Literacki : czasopismo kwartalne poświęcone historii i krytyce literatury polskiej 42/1, 168-182

1951

WŁADYSŁAW KURASZKIEWICZ i JÓZEF ŁUKASZEWICZ

ILOŚĆ RÓŻNYCH WYRAZÓW W ZALEŻNOŚCI OD DŁUGOŚCI TEKSTU

Studia nad słownictwem poszczególnych autorów czy gatunków literackich są u nas dotąd ledwie rozpoczęte i fragmentaryczne. Pełnych słowników tego typu brak zupełny; można wymienić ledwie kilka pełnych indeksów wyrazowych: 1) Babiaczyka do *Biblii Zofii*, 2) Słońskiego do *Psałterza puławskiego*, 3) Kuraszkiewicza do zapisek i rot sądowych warszawskich[1]. Inne zbiory wyrazów są niezupełne, zestawione dla szczególnych celów, np. objaśnienia wyrazów obcych, archaizmów czy w ogóle wyrazów niejasnych. Potrzebę opracowania pełnych słowników podkreślali przede wszystkim językoznawcy, ale także styliści i literaci chętnie korzystaliby z prac tego typu. Przecież dobrze byłoby wiedzieć, ilu np. wyrazów używa pewien autor, jakie wyrazy lubi, jakich unika, czy używa wielu wyrazów, ale w ograniczonym, zasadniczym ich znaczeniu, czy też ma wyrazów mało, ale wydobywa całe bogactwo ich treści. Porównanie pod tym względem słownictwa różnych autorów, np. Reja i Kochanowskiego, albo różnych gatunków literackich, prozy i poezji, może przynieść bardzo interesujące wyniki, wydobywając opozycje słownikowe i semantyczne[2]. Większe prace słownikowe można wykonać wyłącznie zespołowym, zorganizowanym wysiłkiem.

---

[1] 1) Adam Babiaczyk, *Lexikon zur altpolnischen Bibel 1455. (Sophienbibel, Ausgabe von Małecki)*, Wrocław 1906. 2) Stanisław Słoński, *Psałterz puławski*, Warszawa 1916, Wyd. Tow. Nauk. Warsz., Wydz. Językoznawstwa i literatury, Komisja Językowa. 3) Władysław Kuraszkiewicz i Adam Wolf, *Zapiski i roty polskie XV—XVI wieku z ksiąg grodzkich i ziemskich ziemi warszawskiej*, Prace Komisji Językowej PAU, nr 36, Kraków 1950.

[2] Znaczenie badań słownikowych ostatnio najdobitniej określił Józef Stalin, *W sprawie marksizmu w językoznawstwie*, Pamiętnik Literacki XLI, 1950, z. 2.

Jeden człowiek utonąłby w podobnej pracy na lat wiele lub na całe życie. Wiadomo np., że słownik do Mickiewicza zaczynali kilkakrotnie różni ludzie, a obecnie trzeba było pracę rozpocząć na nowo. Właśnie Instytut Badań Literackich ma tę zasługę, że podjął się zorganizowania i uruchomienia wielkich prac zespołowych w różnych ośrodkach uniwersyteckich, i to w takiej skali, jaka w naszych warunkach przed wojną była nieosiągalna.

Jednym z takich wielkich zamierzeń IBLu jest gromadzenie materiałów słownikowych, gramatycznych i frazeologicznych z piśmiennictwa polskiego XVI wieku. Na miesięcznych konferencjach sprawozdawczych kierowników zespołów pracujących oraz przedstawicieli IBLu, w szczególności kiedy się układało plan pracy i decydowało, co zbierać, z jakich źródeł, w jaki sposób itd., jednym z trudnych zagadnień była decyzja: ile zbierać. Bo już po miesięcznej pracy zespołów (razem około 30 osób odpowiednio przygotowanych) można się było łatwo zorientować, że do skatalogowania w całości bogatego piśmiennictwa polskiego XVI wieku pod względem słownikowym, gramatycznym i frazeologicznym przy pomocy powielanych cytatów tekstowych potrzeba by było około 20 lat pracy w obecnym układzie sił. Mechanicznie zwiększyć ilości pracowników nie jest łatwo, bo pomijając odpowiednio wyższe koszty, brak obecnie takiej ilości sił kwalifikowanych do tej pracy. Pracownikami godnymi zaufania mogą być tylko uczestnicy wyższych seminariów języka polskiego; przy gorszych pracownikach kontrola pracy zajęłaby zbyt wiele energii. Trzeba zatem było ograniczyć materiał badany, ale w ten sposób, by cel naukowy pracy nie został zagrożony. Po wstępnych rozważaniach zadecydowano (z inicjatywy IBLu) wciągnąć do tej pracy tylko 20% całego materiału, biorąc do przepracowania z każdego utworu nie wszystko, tylko co piątą stronę tekstu. W ten sposób przez cały materiał XVI wieku będzie można przejść w ciągu czterech lub pięciu lat. Uzasadnieniem takiego ograniczenia materiału były dwa referaty próbne. M. R. Mayenowa wykazała na przykładzie *Biblii Zofii*, poprzez żmudne obliczenia z każdej piątej strony tekstu wyrazów sprawdzanych w ogólnym indeksie wyrazowym Babiaczyka, że z 20% tekstu *BZ* uzyskuje się około 60% wszystkich haseł słownikowych, tj. osobnych wyrazów. Podobnego przeliczenia dokonał też Wł. Kuraszkiewicz na materiale wydawanych przez niego tekstów warszawskich zapisek i rot sądowych XV—XVI w. (razem 350 stron druku), do którego przygotował również pełne

indeksy wyrazów. Obliczenia jego z 20% tekstu przyniosły około 50% wszystkich wyrazów. Wyniki obu referatów uznano za wystarczające uzasadnienie metody wciągania w dalszym ciągu pracy leksykograficznej tylko 20% badanych tekstów. Poprawkę zastosowano tylko do tekstów małych a cennych, które trzeba będzie uwzględnić w całości (100%) lub co najmniej w połowie (50%). Szczegółowsze wskazówki przyrzekł przedstawić Wł. Kuraszkiewicz po skatalogowaniu słownictwa z utworu Mikołaja Reja pt. *Wizerunek własny żywota człowieka poczciwego.*

1

W tym celu zbadano *Wizerunk* dwojako: a) w 20%, b) w całości. W pracowni IBLu zestawiono normalnie materiał słownikowy z 20% tekstu, wyzyskując co piątą stronę, a niezależnie od tego uczestnicy seminarium języka polskiego we Wrocławiu opracowali pełny tekst *Wizerunku*. Ponieważ studenci pracowali grupami i w różnym tempie (od lutego do czerwca 1950 r.), uzyskano w wyniku trzy indeksy. Indeks *A* z tekstu od strony 1 do 155 (według porządnego wydania St. Ptaszyckiego z roku 1882), indeks *B* z tekstu od strony 156 do 255 i wreszcie indeks *C* z końca tekstu od strony 256 do 278. Zatem długości grup tekstu, z których sporządzono indeksy *A*, *B* i *C*, układają się według ilości stron w stosunku jak 155 : 100 : 23, czyli w przybliżeniu jak 6 : 4 : 1. Grupy tekstu będziemy dalej oznaczali tymi samymi literami, co odpowiadające im indeksy. Nad całością pracy czuwał Wł. Kuraszkiewicz pilnując, by wszędzie stosowano te same metody wyrzucania haseł, tj. wyboru różnych wyrazów z tekstu. W lipcu i sierpniu 1950 r. z uzyskanych trzech indeksów (*A*, *B* i *C*) Wł. Kuraszkiewicz zestawił indeks całości przeprowadzając jednocześnie kontrolę i potrzebne obliczenia.

Rozbicie całego tekstu *Wizerunku* na trzy nierówne części (6 : 4 : 1) okazało się bardzo pożyteczne dla poruszanego tu zagadnienia zależności ilości różnych wyrazów w tekście od jego długości. Oprócz podstawowych indeksów: *A*, *B* i *C*, przez ich łączenie uzyskano indeksy dla połączonych grup tekstu: $A+B$, $A+C$, $B+C$, oraz indeks całości: $A+B+C$. Z przeliczenia wszystkich wyrazów otrzymano: 1) w indeksie *A* — 4672 różnych wyrazów, 2) w indeksie *B* — 3343, 3) w indeksie *C* — 1550 wyrazów. Następnie porównywano wyrazy występujące w poszczególnych indeksach, co dało następujące wyniki: 4) indeks *B* obejmuje 1150 wyrazów

nie występujących w indeksie *A*, a zatem tekst *A*+*B* musi zawierać 4672+1150=5822 różnych wyrazów, 5) indeks *C* zawiera 304 wyrazy nie występujące w *A*, więc tekst *A*+*C* zawiera 4672+304= =4976 wyrazów, 6) indeks *C* zawiera 367 wyrazów nie występujących w *B*, więc w tekście *B*+*C* występuje 3343+367=3710 różnych wyrazów, 7) wreszcie zestawienie indeksu *A*+*B* z indeksem *C* wykazało w *C* 191 wyrazów nie występujących w *A*+*B*, czyli cały tekst *Wizerunku*, *A*+*B*+*C*, zawiera 5822+191=6013 różnych wyrazów.

Do badania zależności ilości różnych wyrazów od długości tekstu nie wystarczy wyrażenie długości tekstu, tak jak podaliśmy poprzednio, liczbą stronic. Tekst *A* obejmuje 6595 wierszy 13-zgłoskowych właściwego tekstu *Wizerunku* (s. 14 —155) oraz teksty wstępne: na stronach tytułowych, wstęp na 180 wierszy prozą i dwa wstępne wiersze *Do czytelnika* (31 wierszy 8-zgłoskowych i 42 wiersze 13-zgłoskowe). Licząc przeciętnie jeden wiersz 13-zgłoskowy na 36 liter, wiersz prozą na 62 litery i wiersz 8-zgłoskowy na 22 litery, wypada na tekst *A* 250.920 liter, co po zaokrągleniu daje 251 tysięcy liter. Tekst *B* obejmuje 4495 wierszy, czyli około 162 tysiące liter. Tekst *C* ÷ 1035 wierszy, czyli około 37 tysięcy liter.

Zamieszczona poniżej tabela I podaje długości poszczególnych grup tekstu (wyrażone w ilości liter i w procentach całkowitej długości utworu) oraz ilości różnych wyrazów występujących w tych tekstach (w liczbie bezwzględnej i w procentach ilości wszystkich wyrazów *Wizerunku*).

TABELA I.

L. p.	Grupa tekstu	Długość tekstu		Ilość różnych wyrazów	
		Ilość liter (w tysiącach)	% całego tekstu *Wizerunku*	Liczba bezwzględna	% wszystkich wyrazów *Wizerunku*
1	*C*	37	8%	1550	25,8%
2	*B*	162	36%	3343	55,6%
3	*B*+*C*	199	44%	3710	61,7%
4	*A*	251	56%	4672	77,7%
5	*A*+*C*	288	64%	4976	82,7%
6	*A*+*B*	413	92%	5822	96,8%
7	*A*+*B*+*C*	450	100%	6013	100,0%

Niezależnie od omówionych prac, we wrocławskiej pracowni leksykograficznej IBLu rozpatrzono normalnie 20% materiału z *Wizerunku*, wyzyskując co piątą stronę tekstu, i spisano indeks

występujących tu wyrazów, których ogólna ilość wynosi 3183. W indeksie tym stosunkowo łatwo można było wydzielić materiał wzięty z grup tekstu *A*, *B* i *C*. Te znów grupy tekstów przepracowanych przez IBL będziemy nazywali odpowiednio α, β i γ. Zatem tekst α obejmuje 20% tekstu *A*, podobnie β obejmuje 20% tekstu *B*, i $\gamma \div 20\%$ tekstu *C*. Zgodnie z tym również tekst $\alpha+\beta$ stanowi 20% tekstu $A+B$, tekst $\alpha+\gamma \div 20\%$ tekstu $A+C$, tekst $\beta+\gamma \div 20\%$ tekstu $B+C$ i wreszcie tekst $\alpha+\beta+\gamma$, to znaczy cały materiał opracowany w IBLu stanowi 20% tekstu $A+B+C$, czyli 20% całości *Wizerunku*. Wł. Kuraszkiewicz przeliczył odrębne, tj. nie znane innym grupom tekstu wyrazy; jest ich w grupie $\alpha \div 1364$, w grupie $\beta \div 677$ i w grupie $\gamma \div 150$ odrębnych wyrazów. Następnie wyodrębniono wyrazy występujące tylko w dwu spośród grup α, β, γ; wyrazów występujących w grupach α i β jest 574, w grupach α i $\gamma \div 74$, a w grupach β i $\gamma \div 41$ wyrazów. Wreszcie 303 wyrazy występują we wszystkich trzech grupach α, β i γ. Sumując razem wszystkie te liczby wyrazów odrębnych i wspólnych w poszczególnych grupach tekstu, otrzymamy sumę 3183, tj. właśnie ogólną liczbę wyrazów z tekstu $\alpha+\beta+\gamma$, czyli ilość wyrazów całego indeksu z pracowni IBLu. Dalsze obliczenia dają następujące wyniki: grupa α zawiera 2315 różnych wyrazów, grupa $\beta \div 1595$ i grupa $\gamma \div 568$ wyrazów; grupa β i γ obejmuje wszystkie wyrazy grupy α, odrębne wyrazy grupy β oraz wyrazy wspólne grupom α i β, tj. $2315+677+41=3033$ różne wyrazy, grupa $\alpha+\gamma$ analogicznie obejmuje $2315+150+41=2506$ wyrazów, grupa $\beta+\gamma \div 1595+150+74=1819$ różnych wyrazów. Powyższe dane przedstawiamy na tabeli II, będącej przedłużeniem tabeli I.

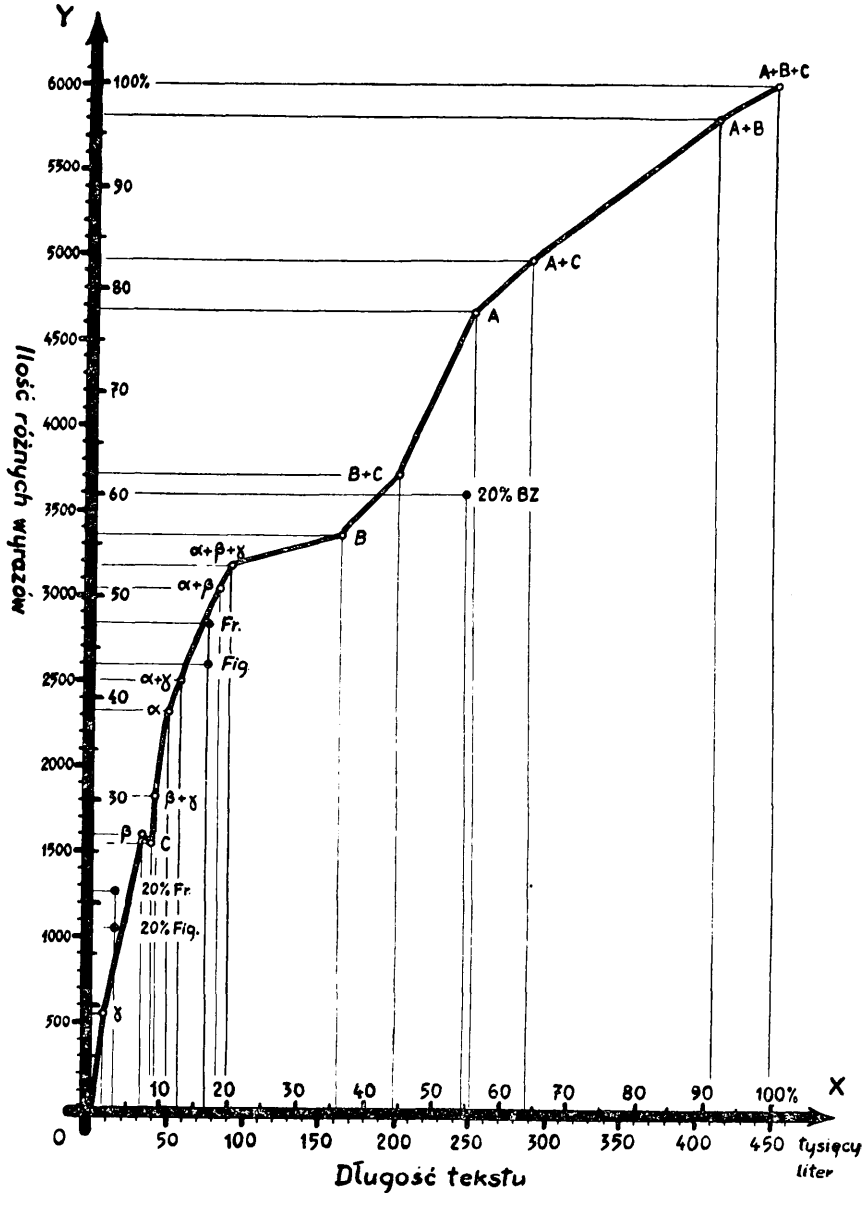
TABELA II.

Lp.	Grupa tekstu	Długość tekstu		Ilość różnych wyrazów	
		Ilość liter (w tysiącach)	% całego tekstu *Wizerunku*	Liczba bezwzględna	% wszystkich wyrazów *Wizerunku*
8	γ	7,4	1,6%	568	9,4%
9	β	32,4	7,2%	1595	26,5%
10	$\beta+\gamma$	40	9%	1819	30,2%
11	α	50	11%	2315	38,5%
12	$\alpha+\gamma$	57,6	13%	2506	41,7%
13	$\alpha+\beta$	82,6	18%	3033	50,4%
14	$\alpha+\beta+\gamma$	90	20%	3183	52,9%

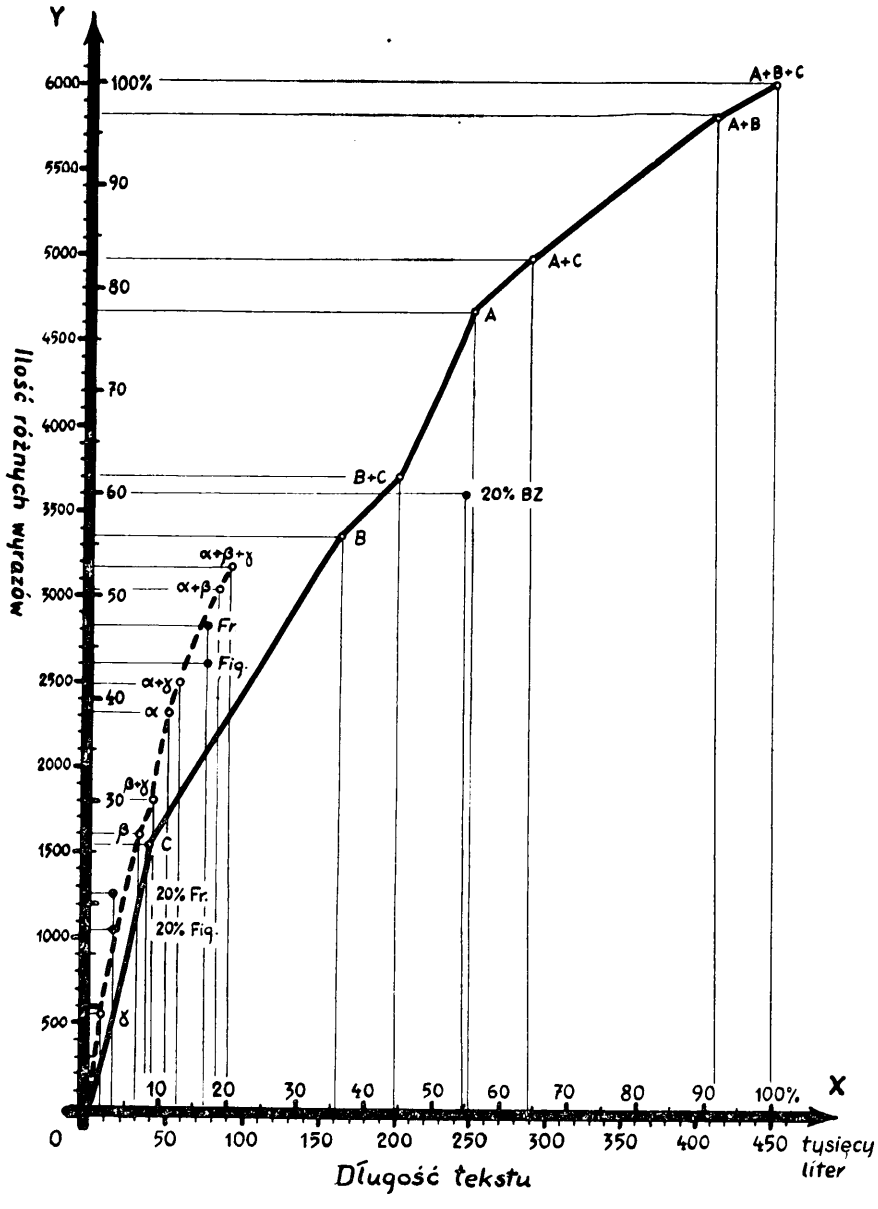
2

Ujętą liczbowo w tabelach I i II zależność między ilością różnych wyrazów w tekście a jego długością można przedstawić graficznie (wykres I i II). Na osi poziomej (X) odkładamy długości poszczególnych grup tekstu, przy czym podziałka pod osią odpowiada długości tekstu, wyrażonej w ilości liter, podczas gdy podziałka nad osią oznacza względną długość grupy tekstu w procentach długości pełnego tekstu *Wizerunku* (450000 liter). Na osi pionowej (Y) odkładamy odpowiadające ilości różnych wyrazów, występujących w danych grupach tekstu. Skala po lewej stronie osi oznacza ilość różnych wyrazów, po prawej stronie — względną ilość wyrazów, wyrażoną w procentach ilości wszystkich wyrazów występujących w *Wizerunku* (6013 wyrazów). W ten sposób każdej grupie tekstu został przyporządkowany punkt na płaszczyźnie. Łącząc odcinkami otrzymane punkty według wzrastającej długości tekstu, otrzymamy linię łamaną, charakteryzującą w pewien sposób omawianą zależność ilości wyrazów od długości tekstu. Prócz połączonych linią punktów odpowiadających poszczególnym grupom tekstu *Wizerunku*, umieszczono na wykresie, dla porównania, punkty dodatkowe, ilustrujące obliczenia ilości różnych wyrazów i liter *Figlików* Reja, *Fraszek* Kochanowskiego i *Biblii Zofii* (każdy utwór zbadano w całości i w 20%).

Linia otrzymana na wykresie I ma pewne nieregularności, na pierwszy rzut oka trudne do wytłumaczenia. Trzeba by bowiem oczekiwać linii stale, lecz coraz wolniej się wznoszącej. Wraz z przyrostem długości tekstu ilość użytych różnych wyrazów rośnie coraz wolniej w miarę, jak wyczerpuje się zasób wyrazów znanych przez autora. Tymczasem na wykresie I widzimy obniżenie się linii w punkcie C oraz wyraźną wklęsłość w punktach B i $B+C$. Przyczynę tych nieregularności znajdziemy jednak łatwo, biorąc pod uwagę nie tylko wielkość, ale i jakość odpowiednich grup tekstu. Przedstawione na wykresie punkty nie są jednorodne. Linia wykresu łączy punkty dwojakiego rodzaju. Punkty oznaczone literami łacińskimi odpowiadają zwartym odcinkom tekstu, podczas gdy litery greckie oznaczają teksty brane z co piątej strony. Wybieranie co piątej strony tekstu daje materiał bardziej różnorodny, o bogatszym słownictwie, i dlatego grupa C, obszerniejsza niż β, zawiera jednak mniej różnych wyrazów, a grupa $\alpha+\beta+\gamma$ daje niemal tyle samo wyrazów, co prawie dwukrotnie od niej większa



Wykres I. Zależność ilości różnych wyrazów w tekście od jego długości. (Punkty odpowiadające grupom tekstu *Wizerunku* połączono linią).



Wykres II. Zależność ilości różnych wyrazów w tekście od jego długości. (Linią ciągłą połączono punkty odpowiadające zwartym partiom tekstu *Wize-runku*, linią przerywaną — punkty odpowiadające grupom tekstu powstałym przez wybór co piątej strony).

grupa *B*. Przy badaniu zależności ilości różnych wyrazów od długości tekstu należy więc osobno traktować grupy tekstu obu rodzajów. Na wykresie II te same punkty odpowiadające różnym grupom tekstu *Wizerunku* połączone są dwiema liniami. Jedna z nich, ciągła, łączy punkty oznaczone literami łacińskimi i obrazuje ilość różnych wyrazów w zwartych partiach tekstu, druga, przerywana, łączy punkty odpowiadające grupom powstałym z wyboru co piątej strony tekstu.

Porównywając przebieg obu linii, przedstawionych na wykresie II, możemy stwierdzić, że:

1) Linia przerywana przebiega powyżej linii ciągłej. Oznacza to, że z dwu grup tekstu tej samej długości grupa uzyskana przez wybór co piątej strony tekstu zawiera więcej różnych wyrazów, niż zwarty fragment. Jeszcze lepszy wynik otrzymałoby się wybierając nie strony, lecz wiersze lub zdania, np. chcąc zbadać 20% całego utworu bierzemy co piąte zdanie. Pomimo indywidualnych różnic w długości zdań otrzymałoby się w wyniku prawie dokładnie 20% całości tekstu.

2) Obie linie mają przebieg podobny, tj. w odpowiadających sobie punktach tworzą analogiczne załamania (np. w punkcie $B+C$ linia ciągła i w odpowiadającym mu punkcie $\beta+\gamma$ linia przerywana załamują się ku górze). Wynika stąd ważny wniosek. 20% próbka tekstu, otrzymana przez wybór co piątej strony tekstu, dość dobrze reprezentuje całość utworu. Tu też wynik można by polepszyć uzyskując próbki reprezentacyjne przez wybór wierszy lub zdań.

3) Odchylenia i nieregularności w przebiegu każdej z linii przedstawiają nam zmienność tekstu. Na wykresie widać wyraźnie podwyższenie przebiegu linii ciągłej w punktach zawierających grupę *A* (analogicznie podwyższenie linii przerywanej w punktach zawierających grupę α). Wskazuje to na bogactwo słownikowe grupy *A*, co można łatwo wytłumaczyć różnorodnością grupy *A*, która obejmuje teksty wstępne i początek właściwego utworu.

4) Otrzymana na wykresie linia w istotny sposób zależy od podstawowego podziału tekstu na grupy *A*, *B* i *C*. Chcąc w obiektywny sposób przedstawić rozwój ilości różnych wyrazów wewnątrz pewnego utworu, należałoby przy badaniu całości uwzględniać na wykresie tylko coraz dłuższe partie tekstu od początku utworu do danego miejsca (np. kolejno do stron 10, 20, 30, 40 . . . itd.) lub odpowiadające im grupy przy badaniu częściowym (przy wybieraniu co którejś strony, zdania czy wiersza). Metoda pracy, stoso-

wana obecnie przez IBL przy opracowywaniu słownictwa *Postylli* Reja, pozwoli uzyskać taki wykres dla tego utworu.

3

Graficzne przedstawienie badanej zależności umożliwia nam poglądowe porównanie otrzymanych dla *Wizerunku* wyników z wynikami podobnych badań, wykonanych na innych utworach. Zwłaszcza interesujące jest porównanie z innym utworem Reja, które może nam wykazać, o ile jest słuszne przypuszczenie, że jego styl, a więc i słownictwo, nie ulega ostrym wahaniom. Na wykresach I i II dodano w tym celu dwa punkty dodatkowe z twórczości Reja. Mianowicie mamy do dyspozycji pełny indeks *Figlików* i również ich opracowanie w IBL z 20% tekstu. Cały utwór jubileuszowego wydania Wittyga z 1905 r. zawiera 239 nie związanych treściowo, luźnych „figlików" 8-wierszowych, 13-zgłoskowych z osobnymi tytułami; do tego we wstępie, prócz karty tytułowej, jest około 24 wierszy prozą, 22 wiersze 14-zgłoskowe skierowane *Do czytelnika* na początku utworu i 28 wierszy 13-zgłoskowych o podobnym tytule na końcu utworu. Licząc przeciętnie u Reja 36 liter na wiersz 13-zgłoskowy, otrzymujemy dla *Figlików* około 75000 liter. Pełny indeks do tego utworu obejmuje około 2600 wyrazów, podczas gdy 50 pierwszych „figlików" (wraz z tytułami), opracowanych przez IBL jako 20% całości, daje około 15000 liter tekstu i 1043 wyrazy. Na wykresie I punkt odpowiadający całości *Figlików* leży poniżej linii *Wizerunku*, punkt zaś odpowiadający 20% *Figlików* — ponad linią. Dopiero na wykresie II widać wyraźnie, że *Figliki* mają bogatsze słownictwo niż *Wizerunk*. Punkt przedstawiający całość *Figlików* leży powyżej linii ciągłej, punkt zaś odpowiadający 20% *Figlików* — powyżej linii przerywanej. Różnice są jednak nieznaczne, zwłaszcza gdy się uwzględni, że linie łamane na wykresie trzeba by wyrównać do linii gładkich wypukłych ku górze.

Podobnie jak z *Figlików* Reja, mamy opracowany pełny indeks wyrazów z *Fraszek* Kochanowskiego i również ich opracowanie w IBLu, w zakresie 20% tekstu (co piąta fraszka). Tu dopiero się pokazuje, że długość tekstu trzeba koniecznie zliczać literami, a nie stronami druku czy wierszami. Kochanowski ma fraszki najczęściej 13-zgłoskowe, ale również często 11-zgłoskowe, nierzadko 8-zgłoskowe lub 12-zgłoskowe, a wyjątkowo także 7-zgłoskowe,

5-zgłoskowe i 10-zgłoskowe. Razem z tytułami *Fraszki* zamykają się liczbą około 75000 liter, czyli tyle samo co *Figliki* Reja, a obejmują 2833 wyrazy. 20% *Fraszek* opracowanych przez IBL zawiera 15000 liter i 1262 różne wyrazy. Na wykresach I i II oba punkty ilustrujące *Fraszki* leżą ponad punktami *Figlików*. Wynika z tego, że słownik *Fraszek* Kochanowskiego jest ilościowo bogatszy niż słownik *Figlików* i *Wizerunku* Reja. Być może, większa ilość wyrazów we *Fraszkach* i *Figlikach* w stosunku do *Wizerunku* ma swoje uzasadnienie w tym, że są to teksty krótkie, dotyczą coraz to innego tematu, więc w sumie słownictwo ich jest bardziej urozmaicone.

Możemy umieścić na wykresach jeszcze jeden punkt dodatkowy. Dotyczy on 20% *Biblii Zofii*. Ogólna ilość materiału *BZ* wynosi około 1220000 liter. Babiaczyk wynotował z tego 5566 haseł, ale opuścił prawie wszystkie nazwy osób i miejscowości. Ile ich może być razem? Obliczenie wyrazów z 20% materiału, tj. z co piątej strony (około 244000 liter), przyniosło liczbę około 3300 wyrazów pospolitych (w tym około 40 wyrazów opuszczonych przez Babiaczyka) i ponad 300 nazw. Nie biorąc nazw pod uwagę, w zakresie wyrazów pospolitych 20% tekstu obejmuje około 59% wszystkich wyrazów *BZ*, co godzi się ze wspomnianym wyżej obliczeniem M. R. Mayenowej. Można zatem przypuścić, że wszystkich nazw jest w *BZ* około 500, czyli ogólna ilość wyrazów *BZ* zamknie się liczbą około 6100, zatem niewiele więcej niż liczba wszystkich wyrazów *Wizerunku*. Na wykresach nie mieści się jednak punkt dodatkowy odpowiadający pełnemu tekstowi *BZ*, ponieważ długość *BZ* prawie trzykrotnie przewyższa długość *Wizerunku*, więc punkt ten musiałby w kierunku poziomym być 3 razy dalej niż ostatni punkt *Wizerunku*. Jednocześnie z tym ilość różnych wyrazów w *BZ* nieznacznie tylko przewyższa ilość wyrazów *Wizerunku*. Dowodzi to niewątpliwie dużego ubóstwa słownikowego *BZ* w porównaniu z *Wizerunkiem*. Przemawia za tym również położenie na wykresie punktu przedstawiającego 20% *BZ*, który leży zdecydowanie pod linią *Wizerunku*. W tabeli III podajemy zestawienie danych liczbowych dotyczących punktów dodatkowych.

W ten sposób można porównawczo oceniać objętość słownictwa w rozmaitych utworach. Nie przesądza to jeszcze ich bogactwa czy ubóstwa stylistycznego, bo w indeksach nie brano pod uwagę różnic znaczeniowych wyrazów, np. rozmaitych przenośni. Można jednak przypuszczać, że porównanie zwykłych przeliczeń wyrazowych, w szczególności w utworach literackich podobnych

TABELA III.

L. p.	Tekst	Długość tekstu		Ilość różnych wyrazów	
		Ilość liter (w tysiącach)	% całego tekstu *Wizerunku*	Liczba bezwzględna	% wszystkich wyrazów *Wizerunku*
15	*Figliki*	75	16,6%	2600	43,2%
16	20% *Figl.*	15	3,3%	1043	17,3%
17	*Fraszki*	75	16,6%	2833	47,1%
18	20% *Fr.*	15	3,3%	1262	20,9%
19	*BZ*	1220	271,0%	5600+ nazwy	ok. 101,5%
20	20% *BZ*	244	54,2%	3300+ 300 nazw	59,8%

pod względem treści, okaże się pożyteczne dla ich oceny stylistycznej. Należałoby tylko mieć jako podstawę porównania kilka wykresów dostatecznie dużych tekstów w zakresie różnych gatunków literackich. Wymaga to jednak jeszcze dużego wysiłku pracy zespołowej.

4

Na podstawie uwag o wykresach I i II możemy powiedzieć, że zdecydowawszy się na badanie tylko 20% utworu, IBL postępuje słusznie biorąc co piątą stronę tekstu, a nie odpowiedniej długości zwarty fragment. I tak w przypadku *Wizerunku* 20% tekstu z co piątej strony zawiera 53% wszystkich wyrazów, podczas gdy do uzyskania tej ilości wyrazów, badając zwarte fragmenty, trzeba by wziąć około 35% tekstu. Należałoby jeszcze zbadać możliwości bardziej drobiazgowego wyboru (zdań lub wierszy), co podniosłoby jeszcze wydajność próbki.

Posiadane materiały pozwalają dokładniej przedstawić wyniki stosowania metody wyboru co piątej strony tekstu. Rozpatrzonych 20 grup tekstu (tabele I, II, III) możemy podzielić na 10 par, w których obok pełnych tekstów występują ich 20% próbki. Zbadamy zależność względnej ilości wyrazów w próbce (tj. ilości wyrazów w próbce wyrażonej w procentach ilości wyrazów danego tekstu) od długości tekstu. W tym celu sporządzimy tabelę IV.

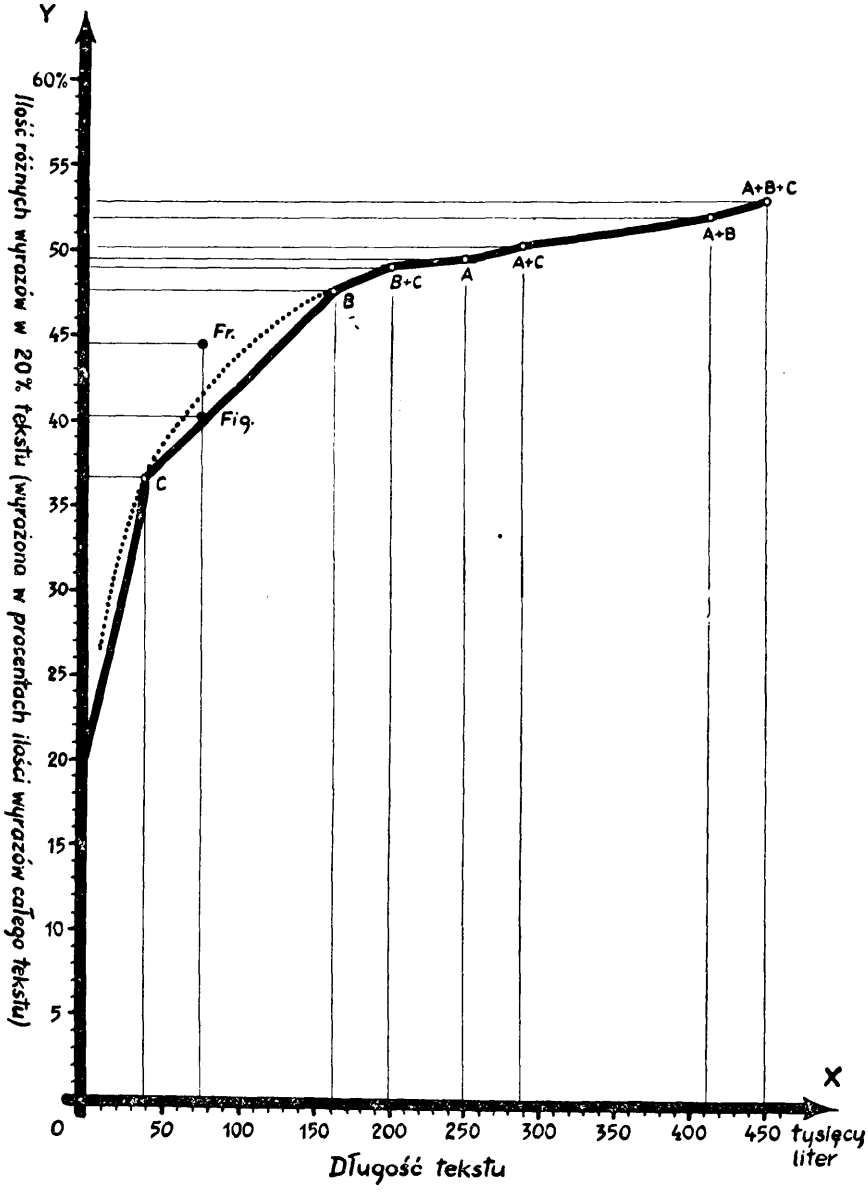
Względna ilość wyrazów w próbce charakteryzuje nam skuteczność próbkowego badania tekstu. Jej zależność od długości badanego tekstu przedstawia wykres III, sporządzony na podstawie

TABELA IV.

L. p.	Grupa tekstu	Ilość liter (w tysiącach)	Ilość wyrazów w tekście	Ilość wyrazów w próbce 20%	Ilość wyrazów w próbce 20% w % ilości wyrazów tekstu
1	*C*	37	1550	568	36,7%
2	*B*	162	3343	1595	47,7%
3	*B*+*C*	199	3710	1819	49,0%
4	*A*	251	4672	2315	49,5%
5	*A*+*C*	288	4976	2506	50,3%
6	*A*+*B*	413	5822	3033	52,0%
7	*A*+*B*+*C*	450	6013	3183	53,0%
15	*Figliki*	75	2600	1043	40,1%
17	*Fraszki*	75	2833	1262	44,5%
19	*BZ*	1220	5600+ nazwy	3300+ 300 nazw	59,0%

tabeli IV. Na osi poziomej (X) odkładamy, jak poprzednio, długość danej grupy tekstu (wyrażoną w ilości liter), natomiast na osi pionowej (Y) względną ilość wyrazów w próbce (ilość wyrazów w próbce 20%, wyrażoną w procentach ilości wyrazów, występujących w całej grupie tekstu).

Punkty odpowiadające grupom tekstu *Wizerunku* połączono na wykresie odcinkami. Otrzymana linia łamana obrazuje nam wspomnianą zależność. Widzimy, że procent wyrazów wynotowanych w 20% próbce rośnie z początku bardzo szybko (mniej więcej do punktu *B*), a później już bardzo powoli. Szkoda, że pierwsza część tej linii jest bardzo słabo reprezentowana w posiadanym materiale, co nie pozwala stwierdzić, jak daleko odbiegają od linii *Wizerunku* punkty *Figlików* i *Fraszek*. Rozpatrywanie względnej ilości wyrazów, a więc stosunku ilości wyrazów w próbce do ogólnej ilości wyrazów w tekście, powinno tu wyeliminować w pewnym stopniu indywidualne cechy badanych utworów i ich autorów. Wyrównanie łamanej do linii gładkiej wypukłej ku górze (na wykresie linia kropkowana) pozostawia punkt, przedstawiający *Fraszki* nad linią, a *Figliki* pod linią. Jest to prawdopodobnie spowodowane tym, że przy wyborze 20% *Figlików* wzięto 50 pierwszych utworów. Mimo iż mamy tu do czynienia z oderwanymi całościami, dało to wyniki gorsze niż wybieranie co piątej strony tekstu, a to dlatego, że *Figliki* muszą być widocznie uporządkowane według pokrewnej tematyki. *Fraszki* leżą powyżej linii *Wizerunku* dlatego, że przy



Wykres III. Zależność względnej ilości wyrazów, w 20% próbce, od długości tekstu.

wybieraniu do próbki co piątej fraszki stosowano drobniejszy podział niż przy wybieraniu co piątej strony w tekstach *Wizerunku*. Jedna fraszka jest krótsza niż strona *Wizerunku*. Jest to ważny argument za stosowaniem drobniejszych podziałów, niż strona, przy pobieraniu próbek tekstu. Punkt obrazujący *Biblię Zofii* nie mieści się znów na wykresie, lecz bardzo wolny wzrost łamanej *Wizerunku* (w dalszych jej punktach) pozwala przypuszczać, że punkt *BZ* leży blisko jej przedłużenia. To potwierdza przypuszczenie o niezależności względnej ilości wyrazów w próbce od badanego utworu. Jeśli przy opracowywaniu tekstu rezygnujemy z badania całości utworu i jako warunek przyjmiemy uzyskanie przynajmniej 50% wyrazów, to wykres III gwarantuje pożądany wynik przy stosowaniu wyboru co piątej strony do tekstów dłuższych niż 300000 liter. Badając teksty mniejsze, należy odpowiednio zmodyfikować metodę pobierania próbek (brać obszerniejszą próbkę lub stosować drobniejszy podział na zdania czy wiersze).

Wobec tego obecnie w pracowniach leksykograficznych IBLu zdecydowano następujący zakres badania: 1) utwory małe, do 100000 liter, bada się w całości, 2) utwory od 100000 do 200000 liter bada się w 50% tekstu, tj. co drugą stronę, 3) utwory od 200000 do 300000 liter bada się w 33%, tj. co trzecią stronę, 4) utwory ponad 300000 liter bada się w 20%, tj. co piątą stronę.

Po zbadaniu wszystkich ważniejszych tekstów XVI-wiecznych w tym zakresie można się spodziewać w wyniku ostatecznym plonu słownikowego prawie zupełnego.

*

Obaj autorzy podzielili pracę przy tym referacie następująco: Władysław Kuraszkiewicz przy pomocy członków seminarium języka polskiego we Wrocławiu opracował omówione indeksy i pierwszą redakcję rękopisu; Józef Łukaszewicz, asystent Grupy Ogólnej Zastosowań Państwowego Instytutu Matematycznego, przygotował wykresy i przeredagował ich interpretację.