Jadwiga Sambor

"Статистические методы изучения лексики", Р. М. Фрумкина, Москва 1964, Издательство «Наука», Академия Наук СССР, Институт языкознания, s. 114, 2 nlb. : [recenzja]

Pamiętnik Literacki : czasopismo kwartalne poświęcone historii i krytyce literatury polskiej 56/4, 598-603

1965

leżności, wrażenia zmysłowe. U Fray Luis można — badając *Profecia del Tajo* — zauważyć przewagę elementów afektywnych (afektywna wariacja, operowanie strukturą klimaks i antyklimaks), natomiast w *Ode en Salinas* połączone są różne formy. San Juan de la Cruz nad elementami afektywnym i obrazowym daje przewagę konstrukcji logicznej. Lope de Vega w swojej wielorakości ukazuje aspekt mało u niego oczekiwany (poezja filozoficzna). U Queveda wreszcie, u którego element logiczny jest tak zagęszczony, a obrazowy tak rozwinięty, dokonuje się — jak mówi Alonso — „najazd afektywności", i jego twórczość jest „wulkanem emocji".

Rozróżnienie sześciu typów w zastosowaniu do poszczególnych utworów poetyckich prowadzi jednak do tak wielkiego zróżnicowania — pisarze bowiem ewoluują, u każdego mogą wystąpić wszystkie możliwe rodzaje — że w efekcie bardzo żmudnej typologii otrzymuje się właściwie to samo, co na początku: potwierdzenie jedyności i niepowtarzalności utworu. Doszedłszy do tego wniosku, Alonso powraca więc do punktu wyjściowego, do intuicji: każdy utwór literacki, każdy pisarz wymaga innej, jemu tylko właściwej metody.

Urszula Dąmbska-Prokop

Р. М. Фрумкина, СТАТИСТИЧЕСКИЕ МЕТОДЫ ИЗУЧЕНИЯ ЛЕКСИКИ. Москва 1964. Издательство „Наука", s. 114, 2 nlb. Академия Наук СССР, Институт языкознания.

Wydana niedawno w Moskwie praca R. M. Frumkiny należy do nielicznych publikacji z zakresu lingwistyki statystycznej. Przedmiotem zainteresowań autorki jest leksykologia, a w szczególności — zastosowanie metod statystycznych przy badaniu słownictwa. Prace tego typu przeprowadzane są w ZSRR przede wszystkim na materiale języka Puszkina, istnieje bowiem słownik języka tego poety, oparty na tekstach całej jego twórczości. Liczne dotychczas ogłaszane publikacje Frumkiny dotyczyły bądź problemów czysto metodologicznych, bądź zagadnień o charakterze teoretyczno-statystycznym. Wydana ostatnio praca jest więc w pewnym sensie podsumowaniem dotychczasowych wyników badań autorki, a także przedstawieniem ogólnoświatowego stanu badań w dziedzinie leksykografii statystycznej.

Stwierdzić trzeba, że problematyka teoretyczno-metodologiczna jest wciąż przedmiotem zainteresowania wielu innych, także i zachodnich badaczy (por. prace Mandelbrota, Somersa, Herdana, Millera) — trwa nadal etap doskonalenia narzędzi i dopasowywania ich możliwości do przedmiotu badań. Wynika to z nowego ujęcia faktów językowych, jako zdarzeń losowych przejawiających prawidłowości o charakterze ilościowym. Prawa statystyczne w języku zauważono najwcześniej w systemie fonologicznym — obecnie wiele prac zajmuje się prawami ilościowymi w słownictwie, a tematyka ta podejmowana jest współcześnie tym gorliwiej, że większość badań statystycznych stanowi podstawę do stosowania teorii informacji w badaniach językoznawczych.

Praca Frumkiny, jakkolwiek zawiera dość obfity aparat statystyczny, jest dla językoznawcy całkowicie czytelna; lekturę znakomicie ułatwia zamieszczony na końcu *Dodatek*, który stanowi doskonałe wprowadzenie w podstawową problematykę statystyczną badań językoznawczych. Autorka definiuje tu przede wszystkim pojęcie prawa statystycznego w odróżnieniu od prawa ścisłego i wykazuje na przykładzie szeregu zjawisk językowych, iż stanowią one prawidłowości statystyczne.

W zasadzie niemal wszystkie dotychczas formułowane reguły gramatyk opierano na intuicyjnym poczuciu częstości form w bliżej nie określonej liczbie tekstów — stosowanie metod statystyki matematycznej pozwala te reguły uściślić. Niemożność badania wszystkich tekstów zmusza do wyboru reprezentatywnej próby — określenie niezbędnej jej wielkości i ocena dokładności otrzymanych parametrów to dwa podstawowe zadania, z jakimi styka się językoznawca w swojej pracy. Rezygnując z wyjaśnień teoretycznych (odsyłacze do podręczników statystyki matematycznej), autorka ogranicza się do podawania podstawowych wzorów metody reprezentacyjnej i elementarnych przykładów językowych.

Pierwszym etapem badań ilościowych nad słownictwem jest budowa słowników frekwencyjnych. Zgodnie z tą chronologią problematyce słowników poświęcony jest rozdział 1: *Słowniki frekwencyjne, ich historia, teoria i metodologia budowy.*

Zasadnicze cele, jakie stawiali sobie autorzy słowników frekwencyjnych, to — zdaniem Frumkiny — 1) racjonalizacja procesu nauczania języka; 2) ulepszenie różnych systemów kodowania; 3) badanie języka poszczególnych autorów, gatunków literackich, itp. Najwięcej słowników powstało w związku z celem pierwszym; wykorzystywanie słowników frekwencyjnych w studiach nad językami autorów jest wciąż jeszcze bardzo niewielkie. Podając krótki zarys historii tych słowników, Frumkina zwraca uwagę na znamienną ewolucję ich struktury: słowniki najnowsze zawierają także częstości form morfologicznych, odmiennych znaczeń wyrazu, itp.

W zasadzie im bogatszą klasyfikację leksyki podaje dany słownik frekwencyjny, tym większa jego wartość, ponieważ niezwykle cenne są dane liczbowe dotyczące np. częstości poszczególnych kategorii gramatycznych, fleksyjnych, słowotwórczych, wyrazów nacechowanych (archaizmy, neologizmy, emocjonalizmy itp.) — jednakże pamiętać trzeba, że zbyt szczegółowy podział nie jest możliwy także ze względów technicznych.

Natomiast z reguły teksty stanowiące podstawę dla danego słownika frekwencyjnego należą do różnych stylów, rodzajów literackich. Ideałem byłby słownik podający przy danym haśle leksykalnym nie tylko jego częstość ogólną, ale częstości w poszczególnych stylach, rodzajach (nie mówiąc o częstościach w poszczególnych tekstach, bo to wydaje się czymś zgoła niemożliwym do zrealizowania).

O ile słownik nie dotyczy języka konkretnego autora lub dzieła (wtedy badaną populacją jest dany tekst), powstaje bardzo trudny problem wyboru dostatecznie reprezentatywnej próby, przy czym reprezentatywność ta musi być zarówno jakościowa, jak też ilościowa. O zestawie tekstów decydują względy literackie, językowe itp., wielkość próby zależy ściśle od przyjętego stopnia dokładności szacunku — jest to zależność funkcyjna i nie można jej pomijać.

Mankamentem niemal wszystkich dotychczasowych słowników było intuicyjne przyjmowanie danego zbioru tekstów jako „dużej próby" — chwalebnym wyjątkiem jest słownik języka angielskiego Torndike'a, słownik hiszpański Garcia Hoza oraz słowniki języka rosyjskiego Josselsona i Steinfeldta. Autorzy słowników języka angielskiego i hiszpańskiego próbują stosować bardziej ścisłe kryteria przy doborze wielkości próby, jednakże nie wykorzystują zależności między wielkością próby a dokładnością szacunku, uznać więc trzeba ich kryteria za dowolne. Ten błąd usunięto przy słownikach rosyjskich: przyjmując dopuszczalny błąd szacunku δ nie większy niż 30% dla wszystkich wyrazów, których częstości empiryczne są nie mniejsze niż $p_r = 0{,}00001$, ustalono wielkość próby N przy założeniu, że zawarte w słowniku wyrazy pokryją założony z góry procent tekstu C.

Frumkina wyprowadza zależności między N, C, p_r i δ, wykorzystując równanie Zipfa i empiryczną funkcję rozkładu częstości wyrazów $C = F(B)$, gdzie $F(B) = \sum_{r=1}^{B} p_r$. Rozważania teoretyczne doprowadzają autorkę do wniosku, iż do celów praktycznych wystarcza w zasadzie oszacowanie grupy 1500 najczęstszych wyrazów.

Wniosek ten wydaje się trochę zaskakujący — i na pierwszy rzut oka rozczarowujący, przede wszystkim dla tych, którzy znają strukturę statystyczną słownictwa badanych przez siebie tekstów i wiedzą, że w konkretnym tekście najbardziej interesującą, „charakterystyczną" warstwę leksyki stanowią wyrazy najrzadsze, tworzące „ogon" krzywej na wykresie częstości. Wobec tego wydawałoby się, że słownik frekwencyjny powinien zawierać także te wyrazy rzadkie, a więc szacunek miałby dotyczyć nie 1500, lecz np. 10 000 wyrazów najczęstszych.

Ze względów merytorycznych nic nie stoi temu na przeszkodzie, pamiętać tylko trzeba, że aby z dostateczną dokładnością δ oszacować częstość tak rzadkich wyrazów, należy wziąć próbę N o wiele wiekszą, a jest to często niemożliwe ze względów technicznych.

Proponowaną przez Frumkinę metodę zastosował E. Steinfeldt przy budowie najnowszego słownika języka rosyjskiego, pomyślanego jako „słownik-minimum" dla potrzeb szkolnych, który powstał w Instytucie Pedagogiki Estońskiej SRR (Tallin 1964). Słownik ten oparto przede wszystkim na tekstach współczesnej literatury młodzieżowej, a także na tekstach rosyjskiej literatury klasycznej, słuchowisk radiowych, prasy itp.

Z pojawieniem się znacznej liczby słowników frekwencyjnych podjęto też rozważania teoretyczne na temat zaobserwowanych prawidłowości ilościowych w leksyce. Jedną z pierwszych i, jak dotąd, najważniejszych było tzw. p r a w o Z i p f a. Prawo Zipfa, jego zastosowania i ograniczenia są tematem rozdziału 2 w pracy Frumkiny.

Omawiając historię odkryć tego prawa, autorka przyjmuje za ostateczną jego postać wzór $p_r = k \cdot r^{-\gamma}$, gdzie k i γ są parametrami stałymi, a więc uogólnioną postać równania Zipfa, który, jak wiadomo, przyjmował $\gamma = 1$.

Prawo Zipfa, określające liniową zależność między rangą wyrazu a jego częstością, stało się przedmiotem szczególnego zainteresowania w latach 50-ych w związku z rozwojem teorii informacji, pozwala ono bowiem określić prawdopodobieństwo pojawienia się wyrazu, a więc daje podstawy do obliczania entropii. Na gruncie teorii informacji wykazano (Mandelbrot), że język spełniający prawo Zipfa jest równocześnie optymalnym kodem informującym. Niestety, wykazano również, że jakkolwiek wzór $p_r = k \cdot r^{-\gamma}$ odpowiada optymalnemu kodowi, nie zawsze język zachowuje to prawo (Miller).

Zależność liniowa między rangą wyrazu a jego prawdopodobieństwem (czyli częstością względną) nie jest spełniona przede wszystkim dla wyrazów o małych częstościach, gdyż w każdym tekście istnieją bardzo liczne grupy wyrazów o identycznych częstościach, równych 1, 2, 3, ... Jeśli częstość traktować jako zmienną losową, a liczbę względną wyrazów o danej częstości jako prawdopodobieństwo tej zmiennej, to zmienna losowa ma rozkład Poissona. Frumkina zwraca uwagę, że wynik ten, otrzymany po raz pierwszy przez Yule'a w r. 1944, potwierdza się we wszystkich zbadanych dotąd tekstach, jakkolwiek Yule sprawdzał ten rozkład na stosunkowo niewielkim materiale ($N = 20\,000$ wyrazów). Tak więc prawo Zipfa nie zachodzi w wypadku wyrazów rzadkich, za które uważa się na ogół wyrazy o rangach większych od 1500 ($r > 1500$).

Mandelbrot wykazał jednak, że liniowa zależność między rangą a częstością wyrazu nie jest spełniona także dla wyrazów najczęstszych, których rangi są mniejsze od 50. W sumie prawu temu podlega niewielka grupa wyrazów, których rangi spełniają nierówność $50 < r < 1500$ — stwierdzić jednakże trzeba, że jakkolwiek stanowią one nieznaczny procent słownictwa, wskutek wysokich częstości pokrywają bardzo dużą część tekstu (około 75—80%). Frumkina wykazuje także, że jedynie w tym przedziale (granice dla r podane powyżej traktować trzeba jako dane orientacyjne) przyjąć można parametr γ za stały: korzystając z danych liczbowych różnych słowników frekwencyjnych, autorka udowadnia dla $r > 1500$ zależność funkcyjną $\gamma = f(r)$, gdzie γ jest niemalejącą funkcją r.

Frumkina przyjmuje wielkość $\sum_{r=1}^{50} p_r$ za stałą dla różnych języków. Podkreślić trzeba, że wypowiada się ona bardzo sceptycznie o zmodyfikowanym równaniu Zipfa zaproponowanym przez Mandelbrota. Jak wiadomo, Mandelbrot podał ogólny wzór Zipfa w postaci $p_r = k(r + \varrho)^{-\gamma}$, gdzie ϱ jest parametrem zmiennym dla $r < 50$, zaś dla $r > 50$ $\varrho = 0$. Wzór Zipfa-Mandelbrota spełniony jest dla wyrazów z przedziału $1 < r < 1500$, a więc także dla wyrazów najczęstszych. Mandelbrot nie podał wprawdzie metody oszacowania parametrów γ i ϱ (jest to dość trudne ze względu na ich uwikłanie), niemniej jednak należałoby przyjmować raczej tę postać równania jako wyrażającą zależność liniową między rangą wyrazu a jego częstością. Wspomnieć tu trzeba o bardzo interesującej propozycji oszacowania parametrów γ i ϱ z równania Zipfa-Mandelbrota, podanej przez Jerzego Woronczaka [1]. Zdaniem Frumkiny, poprawka Mandelbrota polegająca na wprowadzeniu ϱ do równania Zipfa nie wnosi nic nowego, nie można bowiem nadać temu parametrowi żadnej interpretacji lingwistycznej.

Rozdział 3 poświęcony jest zależnościom ilościowym między słownikiem (V) a długością tekstu (N). Autorka omawia przede wszystkim statystyczną strukturę tekstu, czyli rozkład Poissona i jego własności: w dotychczas badanych językach i tekstach najliczniejsze są wyrazy o częstości $f = 1$, a więc wyrazy najrzadsze. Dla $f > 1$ prawdopodobieństwa szybko maleją, do tej pory nie natrafiono na tekst, w którym rozkład wyrazów dawałby krzywą z dwoma maksimami. Wydaje się, że rozkład Poissona jest istotą struktury ilościowej słownictwa każdego tekstu, że jest to cecha słownictwa wszystkich języków naturalnych.

Na ogół zwiększanie próby prowadzi do lepszych, dokładniejszych rezultatów — Yule wykazał jednak, że parametry rozkładu Poissona rosną wraz ze zwiększeniem N, stąd porównywanie statystycznej struktury tekstów jest możliwe tylko przy tekstach tej samej długości. Zaproponowaną przez Yule'a charakterystykę K, pozwalającą porównywać teksty o różnych długościach, Frumkina uznaje za przydatną jedynie w wypadkach wątpliwego autorstwa tekstów itp. Zdaniem autorki, w większości wypadków poprzestawać należy na ujęciu danych w tabele i rysunki oraz na interpretacji czysto lingwistycznej.

Rozporządzając *Słownikiem języka Puszkina*, Frumkina podaje pewne dane dotyczące struktury statystycznej wszystkich jego tekstów. Cytując funkcję $F(B) = \sum_{r=1}^{B} f_r$, gdzie f_r jest częstością wyrazu o randze r, wskazuje autorka na

---

[1] Zob. J. Woronczak, *Metody obliczania wskaźników bogactwa słownikowego tekstów*. W: *Poetyka i matematyka*. Praca zbiorowa pod redakcją M. R. Mayenowej. Warszawa 1965.

znamienny wynik, że 1000 najczęstszych wyrazów pokrywa 70% tekstu — jest to regularność obserwowana nie tylko w języku Puszkina, ale we wszystkich zbadanych dotychczas językach. Z kolei wyrazy o częstościach $f = 1$, $f = 2$ stanowią 48% leksyki, i one to właśnie wpływają na jakościowy charakter leksyki, one decydują o jej cechach charakterystycznych. Tak więc o „bogactwie słownika" autora decyduje ilość i jakość wyrazów rzadkich. Tezę tę ilustruje Frumkina, podając próbę klasyfikacji jakościowej wyrazów rzadkich u Puszkina: wyróżnia ona wśród nich wyrazy stylistycznie nacechowane, imiona i nazwy lub pochodne od nich, wyrazy „charakterystyczne" dla danego autora (Frumkina cytuje słownictwo wyraźnie „puszkinowskie"), archaizmy, neologizmy i inne.

Analizując z kolei 200 najczęstszych wyrazów, dochodzi autorka do wniosku, że 1) nie ma wśród nich ani jednego stylistycznie nacechowanego (z punktu widzenia rosyjskiego języka literackiego lat 30-tych XIX w.); 2) prawie wszystkie te wyrazy są najczęstsze także i we współczesnym języku literackim, wyjąwszy grupę około 50 wyrazów, których częstości uległy przesunięciom wskutek przemian obyczajowo-historycznych, zmiany znaczeń, itp.

Kolejny problem, będący przedmiotem ostatniego rozdziału, to metody porównywania słownictwa tekstów. Porównywanie to dotyczyć ma wydzielonej grupy leksyki, uznanej za charakterystyczną dla badanych tekstów; winno ono polegać nie tylko na wyszukiwaniu wyrazów wspólnych, ale także na zestawieniu ich częstości.

Frumkina omawia metodę stosowaną przez autora frekwencyjnego słownika języka hiszpańskiego, Hoza. Wyszukiwał on słownictwo wspólne czterem wydzielonym przez siebie grupom tekstów (książki, dokumenty oficjalne, gazety, korespondencja prywatna) i zestawiając częstości wyrazów wspólnych w kolejnych parach grup (książki — dokumenty, książki — gazety, itd.) w tzw. tablice korelacji, badał zależności między słownictwem tekstów przy pomocy współczynnika korelacji r.

Tę metodykę Frumkina w zupełności akceptuje, proponując jedynie inny typ współczynnika korelacji. Hoz posługiwał się częstościami absolutnymi, co stwarza trudności rachunkowe i nie pozwala na porównywanie tekstów o różnej długości, stąd Frumkina wybiera współczynnik korelacji rangowej ϱ, używany przy porównywaniu cech niemierzalnych. Proponowany przez siebie tryb postępowania polegający na: 1) wydzieleniu w tekstach leksyki wspólnej; 2) ustaleniu jej częstości w obu tekstach; 3) uszeregowaniu wyrazów w obu tekstach według ich rang; 4) obliczeniu współczynnika korelacji — stosuje autorka, porównując cztery podstawowe grupy tekstów Puszkina: wiersze, prozę artystyczną, publicystykę, przekłady. W każdej z tych czterech grup wydzielono leksykę najczęstszą, a zarazem wspólną i częstą w pozostałych grupach — w ten sposób obliczano korelację dla nielicznej grupy 51 wyrazów, wyłączając z badań wszystkie wyrazy gramatyczne.

Uzyskane wyniki dla wszystkich grup tekstów wykazują korelację nieoczekiwanie małą: niemal żaden współczynnik korelacji ϱ nie przekracza wartości 0,5. Wykluczenie grupy wierszy nieco poprawia wynik, ale korelacja nadal jest dość słaba. Badania Hoza w tekstach hiszpańskich wskazywały na bardzo wysoką współzależność częstości leksyki (współczynnik r rzędu 0,7 — 0,9), jednakże wyniki takie otrzymano wskutek wliczenia do porównywanej leksyki także grupy najczęstszych wyrazów gramatycznych, co Frumkina uważa za zabieg całkowicie niesłuszny: wyrazy gramatyczne mają częstości najwyższe we wszystkich badanych tekstach i sztucznie zwiększają stopień wzajemnej współzależności tekstów.

W sumie należy więc uznać wyniki Hoza i Frumkiny za zbliżone, mimo rozbieżności danych liczbowych.

Jakkolwiek metoda proponowana przez oboje autorów i ich wyniki są interesujące ze statystycznego punktu widzenia, trudno jednak językoznawcy ukryć rozczarowanie: stwierdzenie małej korelacji między wierszami a prozą artystyczną Puszkina niewiele mówi, parametrom tym nie nadaje się tu żadnej interpretacji lingwistycznej.

Jest jeszcze inna wątpliwość, która nie dotyczy tekstów Puszkina, ale może być bardzo istotna np. przy porównywaniu tekstów o nierozstrzygniętym autorstwie: o niewątpliwym „zbliżeniu" dwóch tekstów świadczy właśnie między innymi występowanie w nich nawet nielicznej grupy wspólnych *hapax legomenon* lub *dislegomenon* o bardzo oryginalnych znaczeniach, etymologii itp. — współwystępowanie tak rzadkich wyrazów może być faktem bardzo znamiennym, stąd ograniczanie się do analizy wyłącznie wyrazów najczęstszych nie wydaje się wystarczające. Interesujących przykładów tego typu współzależności dostarcza np. praca R. Morgenthalera *Statistik des Neutestamentlichen Wortschatzes* (Zurich 1958), gdzie współwystępowanie niezwykle rzadkich wyrazów wykorzystywane jest jako jeszcze jeden argument przemawiający za pokrewieństwem poszczególnych ksiąg *Nowego Testamentu.*

Porównując teksty ze względu na ich leksykę, pamiętać również trzeba, że porównanie to dotyczyć powinno przede wszystkim struktury leksyki: różnice między tekstami to także różnice struktur statystycznych ich słownictwa. Porównywanie treści elementów językowych jest często bezowocne ze względu na różnice w tematyce, zestawienie form prowadzi do daleko ciekawszych rezultatów. Wydaje się np., że dla stylu młodopolskiego szczególnie charakterystyczny jest przymiotnik, a styl naukowy znamionuje częste użycie dopełniacza (w związku z przydawką dopełniaczową).

Oprócz załączonego w *Dodatku II* słownika frekwencyjnego języka Puszkina (1000 najczęstszych wyrazów) praca Frumkiny zawiera jeszcze niezwykle cenną i obszerną bibliografię dotyczącą prac z zakresu metod ilościowych w językoznawstwie radzieckim i światowym. Tematyka tych prac — podobnie jak i omawianej tu książki — dotyczy w znacznej mierze zagadnień bądź metodologicznych, bądź teoretyczno-statystycznych, nie zaś czysto językowych. Jest to chyba przejaw ogólnoświatowego stanu badań w tej dziedzinie.

Jadwiga Sambor