

O prawidłowości statystycznej Yule'a (uwagi polemiczne)

MAREK RUSZKOWSKI

(Kielce)

Długość zdania jest jego najbardziej zewnętrzną formalną cechą, która bardzo często poddawana jest analizie. Długość mierzy się w wyrazach (słowoformach) lub składnikach.

Parametr ten ma różną wartość liczbową, która jest uzależniona od rozmaitych czynników: 1. przynależności rodzajowej, gatunkowej i stylistycznej tekstu (por. obliczenia dla języka ukraińskiego – Perebejnis 1967: 156), 2. czasu powstania utworu (np. Wierzbicka 1966), 3. form podawczych: narracja – dialog (np. Mikołajczak 1983), narracja – dialog – monolog wewnętrzny – mowa pozornie zależna (np. Goczołowa 1975), 4. indywidualnych upodobań stylistycznych autora.

Do innych celów omawiany wskaźnik wykorzystał statystyk angielski Georg Udny Yule, który badał długość zdania mierzoną liczbą wyrazów (słowoform) jako cechę charakteryzującą dany styl prozy. Wyniki swoich badań opublikował w artykule *On sentence-length as a statistical characteristic of style in prose with application to two cases of disputed authorship* (Yule 1939). Stwierdził w nim, że w tekstach danego autora długość zdania z czasem maleje, tzn. w tekstach pochodzących z późniejszych okresów jego twórczości zdania są z reguły krótsze niż w tekstach z okresów wcześniejszych. „Długość zdania (y) uznał więc Yule w tekstach jednego autora za malejącą funkcję czasu (t), stąd ze względu na tę cechę trzeba uznać teksty późniejsze danego autora za absolutnie odrębne od wcześniejszych” (Sambor 1972: 82).

Obserwacja Yule'a ma duże znaczenie dla sprawdzania hipotez o autorstwie tekstów. Jadwiga Sambor wymienia i omawia pięć zasad, które należy wziąć pod uwagę przy ustalaniu autorstwa metodami ilościowymi:

1. Jakość badania zależy od trafności wyboru elementów, które będziemy liczyć. Zanim użyjemy statystyki, musimy określić, co jest przedmiotem obserwacji, jaki typ elementów tekstu. Wszechstronne badanie statystyczne, uwzględniające wszystkie składniki tekstu, przeprowadza się rzadko, jest bowiem niezwykle czasochłonne, szczególnie w odniesieniu do obszernych utworów i całych korpusów tekstów;

2. Elementy tekstów danego autora, uznane przez badacza za charakterystyczne dla jego stylu, muszą mieć mniej więcej stałe częstości w tekstach pochodzących z różnych okresów twórczości tego autora;

3. Pożądane jest, aby jako podstawę porównania wybierano cechy tekstu, których autor sobie nie uświadamia. Tkwią one niejako głębiej w tekście, nie są wynikiem świadomego procesu twórczego, pisarz używa ich mechanicznie, dlatego też lepiej zdradzają autorstwo;

4. Statystyczne argumenty wspierające hipotezy o autorstwie mają tym większą wagę, im większą długość mają teksty objęte badaniami;

5. Przy rozstrzyganiu wątpliwości dotyczących autorstwa tekstów trzeba zawsze pamiętać o ograniczeniu metod statystycznych, które dostarczają dodatkowych argumentów wspierających argumenty historyczne, filologiczne i językoznawcze. Metody statystyczne nie są samowystarczalne i rozstrzygające (Sambor 1972: 80-100).

Prawidłowość statystyczna Yule'a odnosi się do zasady drugiej i wskazuje na jej nieprzydatność do badań nad ustalaniem autorstwa. Wiadomo, że styl osobniczy ewoluuje w czasie, więc znaczna część elementów językowych ma inne wartości statystyczne w różnych utworach tego samego autora. Dlatego też długości zdania nie należy brać pod uwagę przy ustalaniu autorstwa tekstów. Tezę Yule'a potwierdziły inne współczesne badania (Sambor 1972: 82-83).

Trudno zgłaszać wątpliwości do ogólnych założeń tezy Yule'a, tym bardziej że intuicyjnie można przyjąć, iż w stylu artystycznym każdy utwór powstaje jako wynik oddziaływania wielu czynników: zamierzeń artystycznych twórcy, poetyki dzieła, przynależności gatunkowej, czasu jego powstania, po części tematyki itp. Mimo że wszystkie dzieła danego autora są przejawem tej samej osobowości pisarskiej, osobowość ta ujawnia się w różnorodny sposób (por. Winogradow 1970). Można więc analizować nie cały dorobek pisarza, lecz jego poszczególne utwory, „stanowiące z jednej strony etap w jego twórczości, z drugiej zaś dające się traktować jako pewne samodzielne i wewnętrznie uformowane jedności, tak treściowe, jak i formalne” (Karaś 1974: 4). Spod tego samego pióra wychodzą teksty o bardzo różnym ukształtowaniu stylistycznym.

Poważne wątpliwości budzi natomiast stwierdzenie szczegółowe Yule'a, że tekst tego samego autora napisany w okresie późniejszym jego twórczości zawiera na ogół zdania krótsze niż tekst pochodzący z wcześniejszego okresu. Można więc wysunąć następującą hipotezę: jeżeli utwór jest splotem wielu indywidualnych cech językowych, jest tworem jednostkowym i niepowtarzalnym, omawiana prawidłowość jest dyskusyjna, a tym samym nie musi być prawidłowością.

Żeby tę hipotezę zweryfikować, zbadałem długość zdań liczoną w wyrazach (słowoformach) w 40 tekstach polskiej prozy artystycznej, napisanych przez 15 autorów (3, rzadziej 2 utwory jednego pisarza). Pierwsze wydania badanych tekstów zostały opublikowane między rokiem 1925 a 1997. Odstęp czasowy między utworami tego samego autora wynosi od kilku do kilkudziesięciu lat, z wyjątkiem opowiadań J. Iwaszkiewicza (*Panny z Wilka*, *Brzezina*), które ukazały się w tym samym roku.

Nie sposób zbadać całych testów, dlatego z konieczności opieram się na metodzie reprezentacyjnej, pobierając z utworów odpowiednie próby statystyczne. Żeby jednak na podstawie próby można było wyciągać wnioski na temat całego tekstu, musi ona spełniać określone warunki, a mianowicie musi być: a) jednorodna, b) dostatecznie liczna, c) pobrana losowo, d) jak najbardziej rozproszona (Kniaginina 1962: 95).

Aby podstawa analizy statystycznej była jednorodna, badam partie narracyjne współczesnej prozy artystycznej. Dzięki temu zachowana zostaje jednorodność stylu funkcjonalnego, rodzaju literackiego oraz formy podawczej (narracja).

Przeanalizowałem pod względem długości po 200 zdań z każdego utworu, co bez wątpienia jest próbą wystarczająco liczną (por. Kniaginina 1962: 95; Szulc 1968: 593; Wierzbicka 1966: 233; Wolińska 1987: 34, 50). Zdanie rozumiem jako ciąg wyrazów zaczynający się wielką literą i zakończony kropką, znakiem ekwiwalentnym lub kombinacją tych znaków.

Próba została pobrana losowo – pięć pierwszych zdań z wylosowanej strony.

Aby była jak najbardziej rozproszona, przeprowadziłem losowanie ograniczone warstwowe (por. Sobczyk 1991: 108; Zasępa 1962: 63), tzn. nie losowałem stron z całej populacji generalnej, tylko z mniejszych, równych pod względem długości odcinków tekstu, na które wcześniej został podzielony.

Tak więc wszystkie cztery warunki (jednorodność, liczebność, pobranie losowe, rozproszenie) zostały spełnione.

W badanym materiale rozkład średnich długości zdań mierzony liczbą wyrazów (słowoform) przedstawia się następująco (w nawiasach rok pierwszego wydania):

Jerzy Andrzejewski: *Ład serca* (1938) – 10,35
Popiół i diament (1948) – 11,25
Ciemności kryją ziemię (1956) – 14,10

M. Choromański: *Biali bracia* (1931) – 11,46
Zazdrość i medycyna (1933) – 12,83
Milosny atlas anatomiczny (1974) – 12,40

- Stanisław Dygat: *Jezioro Bodeńskie* (1946) – 13,00
Pożegnania (1948) – 13,92
Podróż (1958) – 13,10
- Pola Gojawiczyńska: *Powszedni dzień* (1933) – 12,78
Dziewczęta z Nowolipek (1935) – 11,68
Rajska jabłoń (1937) – 13,18
- Witold Gombrowicz: *Ferdydurke* (1937) – 13,00
Trans-Atlantyk (1953) – 13,08
Kosmos (1965) – 13,68
- S. Grabiński: *Cień Bafometa* (1926) – 13,35
Wyspa Itongo (1936) – 13,75
- Jarosław Iwaszkiewicz: *Panny z Wilka* (1933) – 13,68
Brzezina (1933) – 12,92
- Czesław Miłosz: *Zdobycie władzy* (1953) – 13,78
Dolina Issy (1955) – 13,93
- W. Myśliwski: *Palac* (1970) – 12,85
Kamień na kamieniu (1984) – 13,50
Widnokrąg (1997) – 14,95
- Zofia Nałkowska: *Dom nad łąkami* (1925) – 14,47
Granica (1935) – 13,97
Medaliony (1946) – 11, 73
- Tadeusz Nowak: *A jak królem, a jak katem będziesz* (1968) – 12,90
Diabły (1971) – 12,95
Dwunastu (1974) – 16,90
- A. Strug: *Pokolenia Marka Świdy* (1925) – 12,96
Miliardy (1937) – 17,38
- A. Szczypiorski: *Ucieczka Abła* (1963) – 10,06
Msza za miasto Arras (1971) – 10,12
Początek (1986) – 11,00
- K. Wierzyński: *Granice świata* (1933) – 12,92
Pobojowisko (1944) – 12,57
Moja prywatna Ameryka (1966) – 13,63
- Stanisław Ignacy Witkiewicz: *Pożegnanie jesieni* (1927) – 12,58
Nienasycenie (1930) – 14,50.

Jeśli pominiemy J. Iwaszkiewicza, którego oba opowiadania pochodzą z tego samego roku, pozostanie 14 autorów. Relacje między utworami tego samego pisarza ze względu na długość zdań mogą przybierać cztery postacie:

1. Długość zdania w kolejnych utworach maleje (prawidłowość Yule'a),
2. Długość zdania w kolejnych utworach rośnie,
3. Długość zdania (jeśli są trzy utwory) rośnie w drugim, a maleje w trzecim,
4. Długość zdania (jeśli są trzy utwory) maleje w drugim, a rośnie w trzecim.

Ewentualność pierwszą, potwierdzającą obliczenia Yule'a, reprezentują utwory tylko jednego pisarza (Nałkowska), co stanowi 7,1%. Relację drugą zaobserwować można aż u dziewięciu autorów (Andrzejewski, Gombrowicz, Grabiński, Miłosz, Myśliwski, Nowak, Strug, Szczypiorski, Witkiewicz), co daje 64,3%. Zależność trzecią widać u dwóch pisarzy (Choromański, Dygat) – 14,3%; podobnie zależność czwartą (Gojawiczyńska, Wierzyński) – również 14,3%. Tak więc spośród czternastu autorów teksty trzynastu, co stanowi 92,9%, nie potwierdzają prawidłowości Yule'a.

Witold Mańczak zauważa:

Problemy roztrząsane przez językoznawców rozpadają się przede wszystkim na dwie kategorie: 1) te, które można rozstrzygać za pomocą statystyki, oraz 2) te, których za pomocą statystyki rozstrzygnąć się nie da. Językoznawstwo pojmowane jako nauka ścisła zajmuje się tylko problemami pierwszej kategorii. Innymi słowy chodzi o to, aby problemy tak formułować, żeby były rozwiązywalne za pomocą statystyki. Jeśli to nie jest możliwe, nie warto się nimi zajmować, jak nie warto się zajmować żadnymi badaniami, o których z góry wiadomo, że do sprawdzalnych wyników doprowadzić nie mogą (Mańczak 1996: 8).

Omawiany problem bez wątpienia należy tych, które można rozwiązać za pomocą statystyki. Jednak wyniki uzyskane przeze mnie nie potwierdzają ustaleń angielskiego statystyka. Nawet jeśli uznamy, że korpus tekstów poddanych obserwacji nie jest zbyt liczny, to i tak prawidłowość Yule'a należy podać w wątpliwość. Być może znajduje ona zastosowanie w tekstach nieartystycznych, np. naukowych, ale weryfikacja tej hipotezy wymagałaby oddzielnych badań.

Można natomiast przyjąć inną hipotezę – różnica przeciętnej długości zdania między utworami danego autora zależy od odstępów czasowych między tymi utworami: im odstęp ten jest mniejszy, tym różnice w długości zdań są mniejsze i na odwrót. Aby potwierdzić to założenie lub je podważyć, badane teksty podzieliłem na 4 grupy:

1. Utwory (tego samego pisarza), które dzieli okres do 5 lat,
2. Utwory, które dzieli okres 6-10 lat,

3. Utwory, które dzieli okres 11-20 lat,

4. Utwory, które dzieli okres powyżej 20 lat.

Różnice w średniej długości zdań dla poszczególnych grup wynoszą: 1. – 1,31; 2. – 1,37; 3. – 1,45; 4. – 1,20. Widać więc, że najmniejsze różnice występują między utworami, których powstanie dzieli okres powyżej 20 lat., czyli postawioną hipotezę należy odrzucić. Na przykład rozbieżność w długości zdań między powieściami M. Choromańskiego *Zazdrość i medycyna* i *Miłosny atlas anatomiczny*, które zostały wydane w odstępie 41 lat, jest mniejsza (0,43 słowoformy) niż między opowiadaniem J. Iwaszkiewicza, opublikowanym w tym samym roku – 0,76.

Jeśli jednak wzięlibyśmy pod uwagę tylko trzy pierwsze grupy, widać, że postawiona hipoteza się sprawdza, choć różnice nie są statystycznie istotne, nie przekraczają bowiem 0,14 słowoformy.

Nieistotna jest także różnica między przeciętną długością zdań w utworach dwudziestolecia międzywojennego (13,21) i w tekstach powojennych (tu wliczam *Pobojowisko* Wierzyńskiego z 1944 r.) – 12,97.

Średnia długość zdań w przeanalizowanych 40 utworach wynosi 13,08 słowoformy, jest więc nieco większa niż w na przykład w ukraińskiej prozie artystycznej – 12,43 (Perebejnis, red., 1967: 156; cyt. za: Hammerl, Sambor 1990: 201).

Przedstawione fakty skłaniają do wniosku, że tezy Yule'a, która mówi o zmniejszaniu się długości zdań w tekstach tego samego autora pochodzących z późniejszych okresów jego twórczości, nie można uznać za prawidłowość statystyczną dotyczącą prozy artystycznej. Nie podważa to zasadności badania długości zdań w różnych stylach funkcjonalnych i typach tekstów, gdyż dzięki temu można ustalić wiele statystycznych zależności i lepiej scharakteryzować stylistycznie tekst.

Literatura

- Goczołowa Z., 1975, *Składnia powieści Stanisława Przybyszewskiego*, Lublin.
Hammerl R., Sambor J., 1990, *Statystyka dla językoznawców*, Warszawa.
Karaś M., 1974, *O polskim języku artystycznym (1944-1974)*, Kraków.
Kniaginina M., 1962, *Próba zastosowania metod statystycznych w badaniach stylistyczno-składniowych*, „Język Polski” XLII, z. 2, s. 92-116.
Mańczak W., 1996, *Problemy językoznawstwa ogólnego*, Wrocław.
Mikołajczak S., 1983, *Składnia wybranych utworów Bolesława Prusa i Stefana Żeromskiego*, Poznań.
Perebejnis B. (red.), 1967, *Statisticzni parametri stiliv*, Kiev.
Sambor J., 1972, *Słowa i liczby. Zagadnienia językoznawstwa statystycznego*, Wrocław.

- Sobczyk M., 1991, *Statystyka*, Warszawa.
- Szulc S., 1968, *Metody statystyczne*, Warszawa.
- Wierzbicka A., 1966, *System składniowo-stylistyczny prozy polskiego renesansu*, Warszawa.
- Winogradow W., 1970, *Język artystyczny utworu literackiego*. – Rosyjska szkoła stylistyki, wybór i oprac. M. R. Mayenowa, Z. Saloni, Warszawa, s. 355-412.
- Wolińska O., 1987, *Język XIX-wiecznych wiadomości prasowych*, Katowice.
- Yule G. U., 1939, *On sentence-length as a statistical characteristic of style in prose with application to two cases of disputed authorship*, „*Biometrika*” vol. 30, nr 3-4.
- Zasępa R., 1962, *Badania statystyczne metodą reprezentacyjną. Zarys teorii i praktyki*, Warszawa.

On Yule's Statistical Regularity (Polemical Remarks)

British statistician G. U. Yule examined sentence lengths measured with a number words (word forms) as a characteristic of a particular prose style. He found out that in a particular author's works sentences tend to become shorter with time, i.e. in texts from later period of creative activity sentences as a rule are shorter than in texts written earlier. In order to verify this claim I have examined sentence lengths in 40 texts of contemporary Polish artistic prose authored by 15 writers (3 or, in few cases, 2 works by one author). I have analysed 200 sentences from each text. Yule's statistical regularity was corroborated only by one author's works, which makes up for 7,1%. However, the examination has proved another relationship – the shorter period between works of a particular author, the slighter differences in sentence lengths and vice versa.

Wieloletni podległy „państwa molierowskiego, który się widzi w otoczeniu Monsieura Jourdain, nieoceniony Albalat, próbujący jeszcze uwiecznić styl Prospera Merimee (Parandowski 1976: 156). Nawet jeśli zgodzimy się z oceną autora *Alchymista* Nowa, iż – „Każda «sztuka poetycka» składa się z przepisów, zaczerpniętych z wydziałów poprzednich epok, omawiając rzeczy przeszłe lub istniejące, niepodobne znaleźć w niej zarzów nowych gatunków literackich lub nieznanych form wiersza. Poetyki poruszają się w świecie zamkniętym. Ich przykazania i przestrogi opierają się na autorytetach, na wypróbowanej tradycji, odwołują się do pewnych zasad estetycznych” (Parandowski 1976: 156-157), to jednak w czasach renesansu poetyki i sztuki i pisania, warto się przyrzeć tym tekstom jako źródłom inspiracji pisarskich, zaś uwagi w nich zawarte zestawiać z wybrzanymi ogólnymi wskazówkami.

Kompendia autorstwa Albalata swym zakresem obejmowały następujące zagadnienia: talent pisarski, sztuka tekstu, reguły tworzenia powieści, istota stylu i metody jego rozwijania, wdrażanie stylu i środków stylistycznych. Zauważmy, iż au-