ACTA UNIVERSITATIS LODZIENSIS
KSZTAŁCENIE POLONISTYCZNE CUDZOZIEMCÓW 27, 2020

http://dx.doi.org/10.18778/0860-6587.27.09

C O P E
Member since 2019
JM14480

*Tomasz Możdzierz**

https://orcid.org/0000-0002-2037-0469

# DŁUGOŚĆ PRZECIĘTNEGO POLSKIEGO WYRAZU W TEKSTACH PISANYCH W ŚWIETLE ANALIZY KORPUSOWEJ

**Słowa kluczowe:** język polski, glottodydaktyka, słownictwo, przeciętne słowo, trudność tekstu

**Streszczenie.** W niniejszym artykule podjęto próbę znalezienia odpowiedzi na pytanie: *Jak długi jest przeciętny polski wyraz?* W tym celu w trakcie dwóch badań przeanalizowano 90 próbek tekstów zaczerpniętych z Narodowego Korpusu Języka Polskiego zawierających po 1000 słów W czasie badań kontrolowano średnią liczbę znaków w każdej próbce, jak również medianę liczby znaków na wyraz. Uzyskane wyniki pozwoliły na określenie długości *przeciętnego polskiego wyrazu* (PW), która wynosi 6 znaków. Przeliczono też liczbę znaków we wszystkich próbkach i ujęto ją za pomocą PW, porównując długość tekstu wyrażoną słowami faktycznymi i przeciętnymi. Zastosowane procedury statystyczne umożliwiły odrzucenie hipotezy zerowej o braku różnicy między użyciem wyrazów konkretnych, a PW jako miary długości tekstu. W podsumowaniu przedstawiono możliwe zastosowania tego modelu dla celów (glotto)dydaktycznych.

## 1. WPROWADZENIE

Język to system znaków, które na podstawie konwencji społecznej, w sposób całkowicie arbitralny oznaczają elementy naszego świata (de Saussure 1961; Reeves i in. 2005, s. 174; Ratner i in. 2005, s. 19). Znaki te zwyczajowo nazywamy słowami lub wyrazami. Postrzegamy je jako całostki, choć nie są monolitami. Stanowią tworzywo wypowiedzeń, pozostając do siebie w stosunku zależności. Ratner i współpracownicy (2005, s. 19) twierdzą, że struktura hierarchiczna, czyli możliwość rozbicia wypowiedzenia na elementy składowe, takie jak słowa, sylaby czy fonemy, to jedna z najistotniejszych cech odróżniających język ludzki od innych systemów komunikacji. To dzięki niej mamy możliwość przekazania

* tomasz.mozdzierz@doctoral.uj.edu.pl, Szkoła Doktorska Nauk Humanistycznych Uniwersytetu Jagiellońskiego (językoznawstwo); ul. Czapskich 4, 31-110 Kraków.

każdej treści na wiele sposobów i nie jesteśmy, jak np. pszczoły, ograniczeni do skończonej liczby komunikatów,

Wielu z nas traktuje „słowa jak cegiełki, z których zbudowany jest język" (Reeves i in. 2005, s. 173). One też często są przedmiotem analiz, przy czym bardziej obiektywnie można je badać w języku pisanym, gdyż, jak stwierdza Yeni -Komshian (2009, s. 138), choć wiemy już, że „analiza psychologiczna i językoznawcza pozwala na wyróżnienie segmentów czy fonemów w płynnej mowie, to sam sygnał akustyczny nie ma widocznych wskaźników, które odpowiadałyby tym segmentom." W celu obiektywnej analizy rozmaitych własności wyrazów lepiej więc posługiwać się skodyfikowaną formą języka, tj. pismem. Badaniom podlegają wówczas kombinacje grafemów[1]. Systemy pisma różnią się przystawalnością dźwięków do symboli graficznych. W logograficznym jeden grafem obrazuje całe pojęcie, w ideograficznym – znak odpowiada słowu. Są też sylabariusze ze znakami odpowiadającymi sylabom, czy wreszcie pisma alfabetyczne, w których istnieje wyraźna korespondencja między grafemami a fonemami. Grafemy często odpowiadają fonemom jak 1:1 (zob. Defrancis 1996; Ratner 2005, s. 50–51), jeden fonem może mieć jednak kilka odpowiadających mu grafemów, bywa też, że jest realizowany na różne sposoby, mimo jednego znaku kodującego[2]. Zależności te nie zmieniają faktu, że w piśmie alfabetycznym *słowo*/*wyraz* jest elementem znajdującym się między dwiema spacjami (Carver 1972, 1976; Seretny 2006).

Wyrazy różnią się między sobą: znaczeniem, wymową, przynależnością do części mowy, stopniem konkretności/abstrakcyjności, regularnością odmiany (w przypadku odmiennych), a także długością (Laufer 1990). Ostatni parametr jest nie tylko zauważalny, lecz także obiektywnie mierzalny[3]. Długość, co ciekawe, łączy się bezpośrednio z frekwencją użycia. Słowa najczęściej używane są zazwyczaj krótsze niż te, po które sięga się rzadziej, co jest zgodne z prawem Zipfa (Laufer 1990, s. 297–298; Ellis 2002; Reeves i in. 2005, s. 184; Seretny 2006, 2016; Sigurd i in. 2004). Są szybciej przetwarzane, łatwiej je więc zapamiętać. Mają też mniej 'zamienników', co wykazał w swoich badaniach Andersen (2002). Zadaniem respondentów była substytucja jak największej liczby wyrazów w tekstach prasowych przy zachowaniu tej samej liczby jednostek w wypowiedzi. Wyniki pozwoliły badaczowi sformułować wniosek, że krótsze słowa mają mniej 'zamienników'. Użytkownicy języka są więc poniekąd „skazani" na niektóre z nich.

Ellis (2002, s. 158) twierdzi, że frekwencja występowania słów w poszczególnych znaczeniach wpływa bezpośrednio na rozumienie treści wypowiedzi. Jego zdaniem użytkownicy języka dokonują czegoś na kształt nieświadomych obliczeń

[1] Tj. graficznych odpowiedników/reprezentacji fonemów (Rey i in. 2000).

[2] Rey i in. pokazują (2000, s. 8), jak dodatkowe litery w angielskim zapisie, mogą zmienić wymowę, a więc przystawalność grafem-fonem (np. różna wymowa *a* w słowie *stage* i *stag*).

[3] W jednych językach wyrazy są krótsze, w innych – dłuższe. Na przykład przeciętne angielskie słowo ma 2 sylaby, a polskie 3 (Seretny 2006, 2016). Takie różnice pomagają szacować poziom trudności słów.

prawdopodobieństwa znaczeń słów na podstawie frekwencji ich występowania. Na dowód przywołuje przykłady homonimów (np. angielskie *plane* i *tank*), których znaczenie zmienia się zależnie od kontekstu. Autor pisze „[p]sycholingwistyczne eksperymenty wskazują, że biegli czytelnicy rozwiązują takie niejasności [tj. sytuację, gdy wyraz ma kilka znaczeń – TM] przez szybkie przywołanie prawdopodobnych ograniczeń będących pochodną wcześniejszych doświadczeń. Informacją pierwszego rzędu jest frekwencja, *plane* znacznie częściej oznacza bowiem maszynę niż coś innego, a *left* jest znacznie częściej używane w stronie czynnej niż biernej”[4]. Odzwierciedleniem takiego ‘matematycznego’ ujęcia rozumienia jest metodologia stosowana w ramach językoznawstwa korpusowego, która umożliwia interpretacje danych statystycznych dotyczących użycia języka, np. frekwencji danej konstrukcji czy formy fleksyjnej (zob. Górski 2012, s. 291).

Posługiwanie się samą frekwencją i długością słów jako wyznacznikami stopnia ich trudności jest pewnym uproszczeniem. Dla zilustrowania tego problemu Lado (1955, s. 26) przytacza wiele przykładów, w których kontekst czyni niezrozumiałymi nawet najbardziej pospolite słowa. Pokazuje też, że przymiotnik *observational*, mimo niskiej frekwencji, nie sprawi problemu nawet początkującym adeptom języka angielskiego, jeśli tylko poznali już pierwsze dwa tysiące najczęściej używanych słów, wśród których znajduje się czasownik *observe*. Ogólnie jednak wielu badaczy przyjmuje, że słowa dłuższe są trudniejsze, a krótsze – łatwiejsze (zob. Lado 1955; Laufer 1990; Seretny 2006; Broda i in. 2014; Charzyńska, Dębowski 2015; Dębowski i in. 2015; Gruszczyński i in. 2015).

Jako rodzimi użytkownicy polszczyzny intuicyjnie potrafimy wskazać w niej wyrazy krótkie czy długie, a zatem potencjalnie łatwiejsze i trudniejsze. Nie wiemy jednak, *jakiej długości jest przeciętny polski wyraz*. Poszukiwanie odpowiedzi na to pytanie stało się więc przedmiotem podjętych przez mnie analiz.

## 2. STAN BADAŃ

W dotychczasowych badaniach nad powiązaniami między stopniem trudności wyrazu a jego długością posługiwano się miarą sylab. Dla języka polskiego przyjęto, że trudne są słowa liczące więcej niż 3 sylaby (zob. Seretny 2006; Broda i in. 2014; Charzyńska, Dębowski 2015; Dębowski i in. 2015; Gruszczyński i in. 2015). Nie jest to jednak miara do końca precyzyjna. Dla przykładu porównać można dwa słowa – *mama* i *chrzcielny*. Oba składają się z dwóch sylab, a drugie na pewno jest dłuższe i trudniejsze (nie tylko dla cudzoziemców). Alternatywną miarą mogłyby być więc nie sylaby, lecz użyte do zapisu znaki. Wyraz *mama* zapisujemy za pomocą 4 znaków, a przymiotnik *chrzcielny* – 9.

[4] Tłumaczenie własne.

Konstrukt modelowego wyrazu o określonej liczbie znaków został wprowadzony przez Carvera i wykorzystany w badaniach nad tempem czytania (zob. Carver 1972, 1976, 1977–1978). Badacz posługiwał się określeniem *wyraz standardowej długości (W)*[5], który według niego liczył W = 6 znaków-spacji[6]. Wynik taki otrzymał po analizie średniej liczby znaków-spacji na wyraz (która wahała się od 5,1 do 7) w czterech wykorzystywanych tekstach. Choć wartość W została przyjęta dość arbitralnie, stosowana była przez lata jako jednolita miara tempa czytania tekstów o różnych poziomach trudności (zob. Carver 1982). W swoich badaniach Carver udowodnił (1983), że interpretacja wyników tempa czytania zmienia się zależnie od wykorzystywanej jednostki miary szybkości. Gdy tempo wyrażane było w faktycznych wyrazach na minutę (*words per minute* = wpm), potwierdzało się intuicyjne przypuszczenie, że teksty coraz trudniejsze są czytane coraz wolniej. Gdy jednakże jednostką szybkości były wyrazy standardowej długości (W) na minutę (Wpm), tempo czytania okazywało się stałe na niemal wszystkich poziomach trudności, spadając dopiero wówczas, gdy tekst przekraczał umiejętności poznawcze badanych i/lub wymagał od nich wzmożonej koncentracji. Mimo swej arbitralności model zwracał uwagę na obserwowalne dysproporcje w długości wyrazów i miał na celu stworzenie jednostki nazywającej pole tekstowe obejmowane wzrokiem w trakcie jednej fiksacji. Wielkość tego pola w badaniach nad czytaniem wcześniej określana była jako ~2,5 cm (ok. 5 znaków ze spacjami[7]) w obie strony od punktu fiksacji (Huey 1908, s. 52; Taylor 1965, s. 188). Carver (1983) badał tempo czytania tekstów pisanych normalnie oraz z dodatkowymi s p a c j a m i w każdym możliwym miejscu. Zaobserwował wówczas umiejętność adaptacji ludzkiego oka do wydłużonej odległości między wyrazami, potwierdzając tym samym przydatność wyrazu standardowej długości jako miary tempa czytania.

*Przeciętny wyraz* (dalej: PW) jako miarę tempa czytania proponuje również Brysbaert (2019, s. 25), postulując jednocześnie konieczność stworzenia takiego konceptu dla każdego języka na podstawie analiz korpusowych. Dla angielskiego, wspólnie z Johnsem, ustalili długość PW równą 4.6 znaku. W swojej metaanalizie Brysbaert przywołuje też pracę Seidenberga, który wykorzystywał podobny model w badaniach tempa czytania. U tego ostatniego PW w angielskim było równe 5 literom (Seidenberg 2017, za: Brysbaert 2019, s. 8).

[5] Ang. *standard-length word* (W). W swoich poszukiwaniach zdecydowałem się na przymiotnik *przeciętny*, gdyż mowa będzie o pewnej abstrakcji, stanowiącej punkt odniesienia. Określenie *standardowy* implikuje, moim zdaniem, zasadę/normę.

[6] Carver używał terminu *character-spaces*. Skonstruował również koncept standardowej długości zdania (S), które składało się z 100 *character-spaces*, a więc 16,7 W (zob. Carver 1977–1978).

[7] Ang. *character-spaces.*

## 3. METODOLOGIA BADAŃ

Poszukiwanie odpowiedzi na pytanie, jak długi jest przeciętny polski wyraz, wymagało zebrania danych ilościowych, metodologie ilościowe dają bowiem „podstawy do wyciągania wniosków o naturze języka" (Lewandowska-Tomaszczyk 2011, s. 143). Do analizy zagadnienia zdecydowano się wykorzystać zasoby *Narodowego Korpusu Języka Polskiego* (dalej: NKJP lub *Korpus*), czyli komputerowego zbioru autentycznych tekstów językowych, mówionych i pisanych, reprezentujących różne odmiany, style i typy tekstów (Lewandowska-Tomaszczyk i in. 2012, s. 4). Na oficjalnej stronie *Korpusu* widnieje informacja, że zawiera on łącznie ponad 1,5 mld wyrazów. Jego źródła w 90% stanowią teksty pisane, 3% niesklasyfikowane, a 7% mówione (Górski 2012, s. 28–29). 80% wszystkich tekstów pochodzi z prasy lub książek (Górski 2012, s. 28–29). Jak piszą twórcy, *Korpus* to „największy, morfologicznie anotowany zbiór danych języka polskiego" (Górski 2012, s. 9), a zarazem narzędzie: (i) dostarczające danych statystycznych, (ii) umożliwiające ilościową analizę języka, a także (iii) pozwalające odkryć w nim pewne tendencje (Górski, Łaziński 2012, s. 291, 293).

Zbiór danych do badania długości przeciętnego polskiego wyrazu odbywał się na dwa sposoby z zastosowaniem podobnych procedur analizy. Każda wygenerowana z *Korpusu* **próbka** zawierała 1000 słów z zaznaczonym minimalnym kontekstem (+1 jednostka z obu stron wyrazu, która nie była brana pod uwagę w obliczeniach). Łącznie wyekscerpowano 90 list (40 w **próbie 1** i 50 w **próbie 2**) zawierających w sumie 90 000 pojedynczych słów. Wszystkie one stanowią przypadkowe wyimki z tekstów autentycznych reprezentujących polszczyznę w użyciu. W obu zbiorach liczba znaków w próbkach miała rozkład normalny (w teście Shapiro-Wilka: W = 0.98613 przy p-value = 0.8971 dla próby 1; W = 0.98589 przy p-value = 0.8091 dla próby 2)[8].

Badanie pierwsze (próba 1) polegało na zebraniu czterdziestu próbek zawierających tysiąc losowych słów z uwzględnieniem zróżnicowania gatunkowego tekstów. Pierwsze 10 próbek otrzymano przy pomocy zapytania „[orth="[bcćdfghjklłmnńprsśtwzźżaąeęioóuy]+" /i]", które generowało wyniki z całości zasobów (dalej: „teksty ogólne"). Następnie każde kolejne 5 próbek zawierało zapytanie z modyfikatorem typu (wymieniona wyżej komenda + „meta type=lit_proza/fakt/publ/nd/inf-por/urzed")[9], co pozwoliło wychwycić tylko teksty konkretnych gatunków. Przy ekscerpcji próbek uwzględniono kolejno: (1) wszystkie teksty zawarte w *Korpusie*, (2) tylko teksty prozy, (3) literaturę faktu, (4) publicystykę, (5) teksty naukowo-dydaktyczne, (6) informacyjno-poradnikowe i (7) urzędowe.

[8] Wszystkie obliczenia w artykule są autorskie.

[9] zob. http://nkjp.pl/poliqarp/help/plse3.html#x4-120003.7 [22.05.2020]

W pierwszej próbie wybór gatunków dokonany był ze względu na ich typowo pisany charakter. Typy takie jak wiersze, dramaty, książki niebeletrystyczne, niesklasyfikowane i inne nie zostały w badaniach uwzględnione ze względu na: nieostry charakter kategorii (teksty niesklasyfikowane), możliwe występowanie w nich neologizmów oraz form niestandardowych (w przypadku tekstów internetowych, dramatów, poezji), wpływy języka mówionego (teksty internetowe). Aby uzyskać średnią liczbę znaków na wyraz, a następnie średnie dla poszczególnych gatunków, każdorazowo podzielono liczbę znaków w próbce przez liczbę wyrazów. Następnie obliczono średnią ze średnich dla każdego gatunku. Uzyskano również średnią oraz medianę dla wszystkich 40 próbek. Wyniki przedstawione zostały w formie dokładnej, tj. liczby z miejscami po przecinku, jak również w formie liczby całkowitej. Zaokrąglenia dokonano ze względu na fakt, że litery stanowią najmniejszą część naszego pisma (Defrancis 1996; Wolf i in. 2005, s. 442) i nie dzieli się ich na mniejsze części (Juel, Minded-Cupp 2000). Choć wykorzystujemy charakterystyczne elementy graficzne liter, by szybciej dekodować znaczenie słów, zasadniczo jednak nie przetwarzamy znaków pojedynczo, lecz holistycznie (Huey 1908, s. 73–74; Cooper, Petrosky 1976, s. 186; Alderson 2000, s. 18–19; Perfetti 2007; Grabe 2009, s. 24; Wolf i in. 2005, s. 445–446; Kuhn i in. 2009, s. 232).

W badaniu drugim (próba 2) zebrano pięćdziesiąt próbek z *Korpusu*, każda po 1000 słów, tym razem z zastosowaniem modyfikatora kanału publikacji. Zapytanie brzmiało: „[orth='[a-zżźćńłśąóę]+' /i] meta kanał = ‚prasa|ksiazka'". Z zasobów NKJP wybrano więc losowo 50 próbek tekstów, które pochodziły z książek lub z prasy. Analiza polegała na podzieleniu liczby znaków w każdej próbce przez liczbę wyrazów, jak również ustaleniu mediany znaków na wyraz (dalej: znak/wyraz) w każdej próbce. Następnie uzyskano średnią wszystkich średnich i średnią medianę. Dane zostały przedstawione w formie wyników dokładnych, jak i w zaokrągleniu.

W zaokrągleniu średnia oraz mediana były takie same. Uzyskaną wartość wykorzystano do przeliczenia liczby znaków w próbkach na *wyrazy przeciętnej długości* (PW), po czym porównano różnice między liczbą wyrazów faktycznych i przeciętnych.

## 4. WYNIKI

### 4.1. PRÓBA 1

Analiza próbek w badaniu pierwszym dała uśrednione rezultaty liczby znaków w wyrazie oscylujące wokół 6 (zob. tabela 1).

Tabela 1. *Średnia liczba znaków na wyraz w zależności od analizowanego gatunku tekstu*

| LP. | TYP TEKSTÓW | ŚREDNIA ZNAKÓW NA WYRAZ | ZAOKRĄGLONA ŚREDNIA |
|---|---|---|---|
| 1 | proza | 5,3170 | 5 |
| 2 | literatura faktu | 5,6930 | 6 |
| 3 | informacyjno-poradnikowe | 5,9732 | 6 |
| **4** | **ogólne** | **6,0000** | **6** |
| 5 | publicystyka | 6,0004 | 6 |
| 6 | urzędowe | 6,1444 | 6 |
| 7 | naukowo-dydaktyczne | 6,4300 | 6 |
| 8 | wszystkie powyżej (1–7) | 5,937 | 6 |

Źródło: opracowanie własne

Wyniki wahały się od ~5,32 do 6,43 znaku na wyraz zależnie od analizowanego gatunku tekstu. Po zaokrągleniu, we wszystkich gatunkach, za wyjątkiem prozy, średni wyraz liczył 6 znaków. Parametry wszystkich próbek znajdują się w tabeli 2. W nawiasach ujęto wyniki po zaokrągleniu do liczb całkowitych. Uznając, że przeciętny wyraz (PW) zawiera 6 znaków, obliczono liczbę PW dla wszystkich próbek. Wyniki znajdują się w tabeli 3. Elementy wspólne dla danych zamieszczonych w tabeli 2 i 3 zostały wytłuszczone, zaznaczono w nich również informacje o parametrach przyjętych odgórnie.

Tabela 2. *Wyniki analizy całego zbioru 40 próbek z NKJP*

| **Liczba próbek** | **40** |
|---|---|
| Całkowita liczba wyrazów | 40 000 |
| Liczba wyrazów w 1 próbce | 1000 |
| **Całkowita liczba znaków** | **235 250** |
| Średnia liczba znaków na wyraz | 5,907 (6) |
| Mediana znaków na wyraz | 5,881 (6) |

Źródło: opracowanie własne

Tabela 3. *Spodziewana liczba przeciętnych wyrazów (PW) w próbie nr 1*

| **Liczba próbek** | **40** |
|---|---|
| Całkowita liczba PW | 39 206 |
| Liczba wyrazów PW na próbę | 980 |
| **Całkowita liczba znaków** | **235 250** |
| Liczba znaków na wyraz (narzucona) | 6 |

Źródło: opracowanie własne

Różnice między liczbą wyrazów faktycznych i liczbą PW w badanych próbkach były statystycznie średnie (d= 0.5120893 w teście Cohena[10]). Analizowane dwie grupy (liczba wyrazów faktycznych i przeciętnych) powstały jednak zgodnie z różnymi założeniami. W grupie wyrazów faktycznych liczba w każdej próbce była stała. W grupie PW każda próbka zawierała różną liczbę jednostek (od 858 w prozie do 1092 w tekście naukowo-dydaktycznym). Ponownie miały one rozkład normalny (W = 0.98629, p-value = 0.9014 w teście Shapiro--Wilka). Do porównania liczby wyrazów faktycznych i przeciętnych zdecydowano się wykorzystać test Wilcoxona. Każdy gatunek tekstu został przeanalizowany osobno. Wyniki przedstawiono w tabeli 4. Podkreślono w niej *teksty ogólne* jako próbkę najlepiej odzwierciedlającą zasoby polszczyzny.

Tabela 4. *Istotność statystyczna różnicy między liczbą wyrazów faktycznych a liczbą PW zależnie od gatunku tekstu*

| GATUNEK | CZY RÓŻNICA BYŁA ISTOTNA STATYSTYCZNIE? | WYNIK TESTU WILCOXONA |
|---|---|---|
| proza | TAK | W = 25, p = 0.007495 |
| literatura faktu | TAK | W = 25, p = 0.00729 |
| informacyjno-poradnikowe | NIE | W = 15, p= 0.6547 |
| **ogólne** | **TAK** | **W = 100, p = 0.0000634** |
| publicystyka | NIE | W = 7.5, p = 0.2896 |
| urzędowe | TAK | W = 0, p = 0.00729 |
| naukowo-dydaktyczne | TAK | W = 0, p = 0.007495 |

Źródło: opracowanie własne

Badania pokazały, że niektóre gatunki tekstów są bardziej zbliżone do „średniego" poziomu od innych. Może to wynikać z docelowej grupy odbiorców – teksty publicystyczne i informacyjno-poradnikowe z założenia mają dotrzeć do szerokiego grona, są więc prostsze. Co istotne, różnica w wyborze jednostki dla przedstawienia długości *tekstów ogólnych* jest istotna statystycznie, a to właśnie ta kategoria, moim zdaniem, najlepiej odzwierciedla całe zasoby polszczyzny. Możemy więc przyjąć, że w całym języku istotne jest, czy długość tekstu przedstawiona będzie w liczbie wyrazów faktycznych, czy w liczbie PW.

[10] Test Cohena powinien powiedzieć, czy w badanych grupach (w moim przypadku, w liczbie wyrazów konkretnych i przeciętnych) różnice są małe/średnie/duże. Dzięki temu testowi możemy wnioskować na temat całej populacji, co oznacza, że wskazana w teście różnica będzie wyglądać podobnie w całym języku polskim.

## 4.2. PRÓBA 2

W próbie 2 wszystkie teksty były jednolite pod względem kanału publikacji, tj. pochodziły z prasy bądź książek. Średnie wyniki dla całego zbioru podane zostały w tabeli 5. Wartości w nawiasach to zaokrąglenia wyników do liczb całkowitych.

Tabela 5. *Średnie wyniki dla 50 prób z tekstów prasowych i książkowych*

| **Liczba próbek** | **50** |
|---|---|
| Całkowita liczba wyrazów | 50 000 |
| Liczba wyrazów/próba | 1000 (narzucona) |
| **Całkowita liczba znaków** | **292 480** |
| Średnia liczba znaków/próba | 5850 |
| Średnia liczba znaków/wyraz | 5,85 (6) |
| Mediana znaków/wyraz | 5,82 (6) |

Źródło: opracowanie własne

Ponownie wyniki wskazały, że w zaokrągleniu do liczby całkowitej średnia liczba znaków na wyraz wynosi 6. Wykorzystując tę wartość, obliczono liczbę PW w każdej próbce. Wyniki zaprezentowano w tabeli 6. W obu tabelach zaznaczono informacje o parametrach narzuconych. Elementy wspólne dla danych zamieszczonych w tabeli 5 i 6 pogrubiono.

Tabela 6. *Spodziewana liczba wyrazów przeciętnych (PW) w próbie nr 2*

| **Liczba próbek** | **50** |
|---|---|
| Całkowita liczba wyrazów przeciętnych (W) | 48 748 |
| Średnia słów przeciętnych/próba | 974,96 (975) |
| **Całkowita liczba znaków** | **292 480** |
| Liczba znaków/wyraz | 6 (narzucona) |

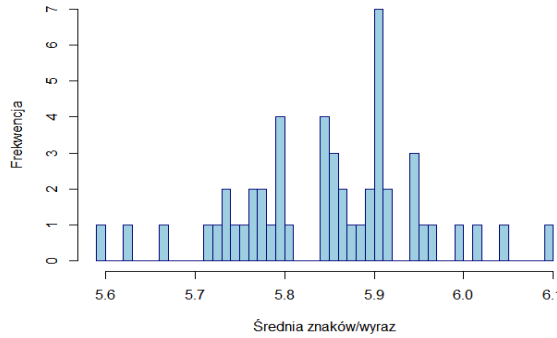
Źródło: opracowanie własne

Wykres 1 prezentuje graficznie średnie znaków w wyrazach w pięćdziesięciu analizowanych próbkach.

Różnice między liczbą wyrazów faktycznych a przeciętnych są znaczące (d= 2.096501 w teście Cohena). Analizowane grupy (liczba wyrazów faktycznych i przeciętnych) różniły się więc między sobą. W grupie wyrazów faktycznych każda próbka liczyła 1000 słów. Wariancja w grupie wynosiła 0. Grupa PW w każdej próbie zawierała różną liczę wyrazów, które rozkładały się normalnie (W = 0.98515, p-value = 0.7782 w teście Shapiro -Wilka). Do porównania liczby wyrazów i liczby PW zastosowano test Wil-

coxona. Różnica w wyborze jednostki okazała się istotna statystycznie, co przemawia za przyjęciem PW jako jednostki długości tekstów (W = 2325, p = 0.000000000000001793).

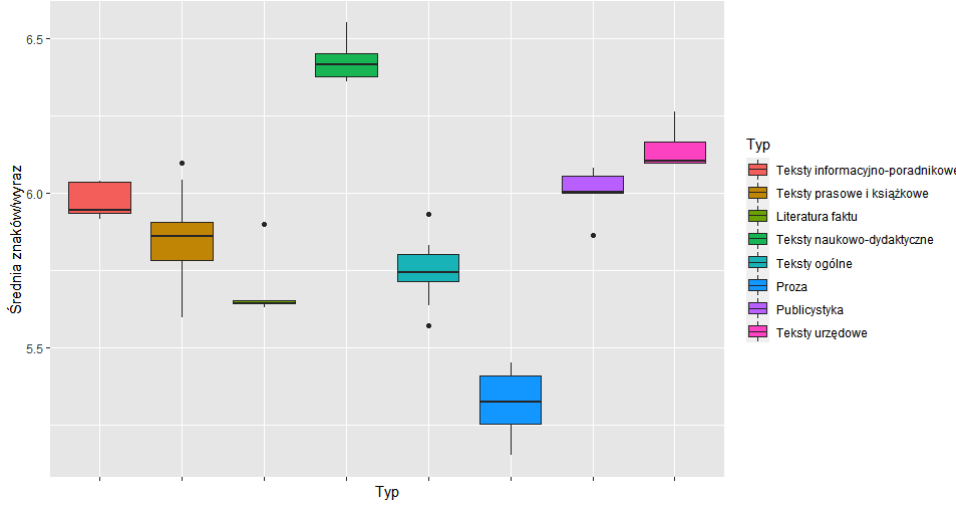
Wykres 1. *Średnia znaków w wyrazach w 50 próbkach tekstów prasowych i książkowych z NKJP*




Źródło: opracowanie własne

## 4.3. PRÓBY 1 I 2 – ŁĄCZNIE

Wykres 2. *Średnia liczba znaków na wyraz we wszystkich 90 próbach tekstów z NKJP*




Źródło: opracowanie własne

Wykres 2 prezentuje średnie wartości znaków/słowo dla wszystkich 90 próbek korpusowych, z uwzględnieniem podziału na gatunki tekstów dla próby 1 i kanału publikacji dla próby 2.

## 5. KOMENTARZ

Przeprowadzona analiza wskazuje, że niezależnie od typu tekstu w pisanej polszczyźnie, wyłączając prozę, średnia liczba znaków na wyraz oscyluje w okolicy 6. Na wynik ten trzeba patrzeć jako na pewien wyznacznik, nie zaś standard. W badaniach korzystano z zasobów NKJP. Jego twórcy, pisząc o problemach i wyzwaniach towarzyszących tworzeniu zasobów, wskazują szereg czynników (np.: różną popularność określonych typów tekstów, różną popularność określonych kanałów publikacji, poczytność literatury fikcjonalnej i niefikcjonalnej[11]), na które trzeba było zwrócić uwagę, by zapewnić NKJP zrównoważenie i reprezentatywność (zob. Górski, Łaziński 2012). PW również powinien być zrównoważony i reprezentatywny[12]. Na razie jego wartość to wynik obliczenia średniej długości wyrazu z całego polskiego słownika. W przyszłości, w szerzej zakrojonych badaniach należy wziąć pod uwagę te same czynniki, o których mówili twórcy *Korpusu*, a także i inne. Pierwszym jest frekwencja. Słowa krótkie są, jak już wspomniano, częstsze (Wyllys 1981), a plasując się najwyżej na listach frekwencyjnych, pokrywają większość przeciętnych tekstów (Seretny 2016). Ich ‘moc’ wpływania na długość PW jest przez to większa niż innych słów. W dalszych badaniach należy też (czynnik drugi) uwzględnić stosunek słów synsemantycznych do autosemantycznych w typowym tekście, te pierwsze są bowiem w polszczyźnie krótsze. Kolejną kwestią (czynnik trzeci) jest fakt, że słowa atematyczne, krótsze, pojawiają się znacznie częściej niż specjalistyczne (Seretny, 2011, s. 140–141), dłuższe, co także powinno proporcjonalnie zwiększyć ich wagę. Warto również w przyszłości, przykładem Carvera, spróbować uwzględnić w badaniach spacje i znaki interpunkcyjne. Na obecnym etapie PW jest więc, moim zdaniem, reprezentatywne, ale nie w pełni zrównoważone. Stworzenie modelu wynikało z konieczności pochylenia się nad kwestią różnic w długości słów jako jednostek do przetwarzania w procesie recepcji. Traktowanie na równych zasadach leksemów jedno- i dziewięcioznakowych nie wydawało mi się zasadne.

Analiza pozyskanych próbek wykazała różnicę między przedstawianiem długości tekstu w wyrazach faktycznych i przeciętnych. Posługiwanie się jednostką

[11] Terminy stosowane przez twórców NKJP.

[12] W pracach nad NKJP uwzględniano fakt, że np. romanse czytane są częściej niż artykuły naukowe, więc udział romansów w *Korpusie* jest większy.

PW pozwala porównywać długość tekstów i zagadnienia z nią powiązane według jednej, bardziej wystandaryzowanej miary. Choć sama praca jest ledwie zarysem problemu, a zagadnienie wymaga szerszych badań, uzyskany model PW można już na tym etapie wykorzystać na kilka sposobów.

Pierwszym zastosowaniem jest odniesienie PW do miary stopnia trudności tekstu. Obecne formuły zazwyczaj posługują się długością słów wyrażoną w sylabach, za trudne uznając, dla języka polskiego, te leksemy, które zawierają cztery sylaby lub więcej (Seretny 2006; Broda i in. 2014; Charzyńska, Dębowski 2015). W Polsce do ustalenia poziomu trudności tekstu można stosować, m.in, aplikację jasnopis.pl, opartą na formule Pisarka i Gunnnga (Dębowski i in., 2015; Gruszczyński i in. 2015) lub algorytm Pracowni Prostej Polszczyzny, wykorzystujący indeks mglistości Gunninga, średnią długość zdania i odsetek słów trudnych (Piekot i in. 2019, s. 209–210). Choć metody zasadniczo sprawdzają się w praktyce, sylaby, na co wcześniej zwrócono uwagę, nie są najbardziej obiektywną miarą długości wyrazu. Przyjmując, że średni wyraz liczy 6 znaków, można założyć, że leksemy krótsze od niego są łatwiejsze, a dłuższe trudniejsze tak w procesie zapamiętywania, odbioru, jak i produkcji (zob. Carver 1976; Sigurd i in. 2004; DuBay 2014). Przy przyjęciu tego złożenia porównanie długości tekstu wyrażonej w wyrazach faktycznych i przeciętnych może stanowić wskazówkę na temat jego trudności. Jeśli tekst zawiera, przykładowo 300 wyrazów, a powinien, biorąc pod uwagę liczbę znaków, mieć ich 400, oznacza to, że większość tworzących go wyrazów musi być dłuższa niż 6 znaków. Uwzględniając dodatkowo fakt, iż słowa funkcyjne są z reguły krótkie (Laufer 1990), słowa autosemantyczne w tym tekście muszą być bardzo długie. Jeśli zaś tekst ma 300 słów, a powinien mieć 200, ponownie biorąc pod uwagę liczbę znaków, to większość jednostek leksykalnych jest w nim krótka[13]. Za „przeciętne" można by wówczas uznać teksty, w których liczba PW i faktycznych byłaby podobna lub, w przypadku analizy naukowej, w których różnica między liczbą wyrazów faktycznych a przeciętnych nie byłaby istotna statystycznie. Dokładne powiązanie konstruktu PW z istniejącymi skalami trudności i gatunkami tekstu wymaga niewątpliwie dodatkowych i szerzej zakrojonych badań, uzyskane wyniki wskazują jednak, że pewne gatunki są statystycznie 'bardziej przeciętne' od innych. Wszystkie te informacje mogą stanowić wskazówki dla nauczycieli języka polskiego przy doborze tekstów. Posługiwanie się konceptem PW może być również użyteczne w adaptacji tekstów dla niższych poziomów zaawansowania.

Drugim zastosowaniem PW jest miara tempa czytania. Carver, którego wynikami posługujemy się obecnie, prowadził swoje badania dla angielskiego, ich przystawalność do polszczyzny ze względu na typologiczną odległość obu kodów jest więc wątpliwa. Wartość 6 znaków na wyraz poparta jest obliczeniami, wyniki wyrażone w $\frac{PW}{min.}$ w planowanych przeze mnie badaniach powinny być

[13] Prawidłowość tę stwierdził też Carver (1977–1978, s. 28), posługując się swoją jednostką W.

bardziej reprezentatywne dla procesu czytania niż obliczone w wyrazach faktycznych czy sylabach.

Ostatnie z proponowanych tu zastosowań konstruktu PW dotyczy dydaktyki. W dzisiejszych czasach, gdy coraz większa liczba zadań zlecana jest w formie komputerowej, organizatorom egzaminów i testów, prowadzącym oraz uczącym się zajęcia miara PW mogłaby pomóc orientacyjnie określić, jaka przestrzeń przeznaczona do zapisania powinna być zajęta. W dydaktyce akademickiej już teraz nierzadko wyznacza się limity prac pisemnych w znakach (w setkach czy tysiącach). Jest to wprawdzie precyzyjne, lecz niezbyt obrazowe. Miara słów przeciętnych mogłaby uplastycznić obraz polecenia. Wyobrażenie sobie około 1 strony A4 jest łatwiejsze niż 1800 znaków. Limity słów faktycznych, w których musi się zmieścić uczeń, pojawiają się też w pisemnych częściach sprawdzianów i egzaminów z języka polskiego jako obcego[14]. Tu, gdyby praca była pisana na komputerze lub transkrybowana, miara PW mogłaby pomóc stwierdzić, które wypowiedzi ewidentnie nie spełniają wymogów poziomu. Jeśli bowiem uczeń np. na B2 pisze tekst **graficznie** o połowę krótszy niż się to szacuje według PW, oznacza to, że jest w nim zbyt duża liczba słów krótkich, tj. o niższej frekwencji lub że konstrukcje, którymi się posługuje, są prostsze niż oczekiwane. Oczywiście byłaby to jedynie wskazówka, niemniej na pewno wspomogłaby nauczycieli w holistycznej ocenie tekstów, zwłaszcza na etapie kształtowania u uczniów umiejętności tworzenia wypowiedzi pisemnej.

## 6. PODSUMOWANIE

Niniejszy artykuł miał na celu przybliżyć konstrukt przeciętnego polskiego wyrazu i możliwe obszary jego zastosowania. Inspiracją do podjęcia badań opisanych w niniejszym tekście były prace Carvera. Carverowska koncepcja wykorzystana została w analizie próbek słów wyekscerpowanych z NKJP, dzięki czemu liczba 6 znaków przypadająca na przeciętny wyraz nie jest arbitralna, lecz poparta danymi ilościowymi.

Przeciętne polski wyraz to pewne abstrakcyjne uogólnienie, które może być jednak bardzo użyteczne na płaszczyźnie praktycznej. Przede wszystkim PW oferuje jednolitą miarę dla analiz parametrów tekstów. Konstrukt przeciętnego polskiego wyrazu może znaleźć zastosowanie przy pomiarach stopnia trudności i/lub ich czytelności. Może być użyty do szacowania objętości konkretnego tekstu, czy też do porównywania i wyznaczania standardów długości różnych wypowiedzi pisemnych. Może także okazać się użyteczny dla nauczycieli języka polskiego, zwłaszcza jako nierodzimego, kształtujących kompetencję leksykalną i umiejęt-

[14] http://certyfikatpolski.pl/o-egzaminie/przykladowe-testy-zbiory-zadan/ [08.05.2020]

ności wzrokowo-manualne cudzoziemców. PW przydatny będzie również jako obiektywna jednostka pomiaru tempa/płynności czytania w języku polskim. Choć koncepcja niewątpliwie wymaga dodatkowych badań z poszerzoną metodologią – ich kierunek został w tekście wskazany – przyjęcie jednostki PW równej 6 znakom stanowi model gotowy do użycia w praktyce.

## BIBLIOGRAFIA


Alderson J., 2000, *Assessing Reading*, Cambridge.

Andersen S., 2002, *Speakers´s information content: length-frequency correlation as partial correlation*, "Glottometrica", nr 3, s. 90–109.

Broda B., Ogrodniczuk M., Nitoń B., Gruszczyński W., 2014, *Measuring Readibility of Polish Texts: Baseline Experiments*, w: N. Calzolari i in. (red.), *Materiały z konferencji LREC 2014 (9th International Conference on Language Resources and Evaluation, Rejkiawik, 26–31 maja 2014)*, Rejkiavik, s. 573–580.

Brysbaert M., 2019, *How many words do we read per minute? A review and meta-analysis of reading rate*, [online], https://psyarxiv.com/xynwg/ [29.05.2020].

Carver R.P., 1972, *Evidence for the invalidity of the Miller-Coleman Readability Scale*, „Journal of Reading Behavior", nr 4(3), s. 42–47.

Carver R.P., 1976, *Word Length, Prose Difficulty and Reading Rate*, „Journal of Reading Behavior", nr 8(2), s. 193–204.

Carver R.P., 1977–1978, *Toward a Theory of Reading Comprehension and Rauding*, „Reading Research Quarterly", nr 13(1), s. 8–63.

Carver R.P., 1982, *Optimal rate of reading prose*, „Reading Research Quarterly", nr 18(1), s. 56–88.

Carver R.P., 1983, *Is reading rate constant or flexible?*, „Reading Research Quarterly", nr 18(2), s. 190–215.

Charzyńska E., Dębowski Ł., 2015, *Empirical verification of the Polish formula of text difficulty*, „Cognitive Studies", nr 15, s. 125–132, https://doi.org/10.11649/cs.2015.010

Cooper C.R., Petrosky A.R., 1976, *A Psycholinguistic View of the Fluent Reading*, „Journal of Reading", nr 20(3), s. 184–207.

Defrancis J., 1996, *Graphemic indeterminacy in writing systems*, „Word", nr 477(3), s. 365–377, https://doi.org/10.1080/00437956.1996.11432455

De Saussure F., 1961, *Kurs językoznawstwa ogólnego*, Warszawa.

Dębowski Ł., Nitoń B., Broda B., Charzyńska E., 2015, *Jasnopis – A Program to Compute Readibility of Texts in Polish based on Psycholinguistic Research*, w: B. Sharp, W. Lubaszewski, R. Delmonte (red.), *Natural Language Processing and Cognitive Science, Proceedings 2015*, Kraków, s. 51–61.

DuBay W.H., 2004, *The Principles of Readability* [online], https://eric.ed.gov/?id=ED490073 [11.05.2020]

Ellis N.C., 2002, *Frequency Effects in Language Processing: A Review with Implications for Theories of Implicit and Explicit Language Acquisition*, „Studies in Second Language Acquisition", nr 24(2), s. 143–188, https://doi.org/10.1017/S0272263102002024

Górski R.L., 2012, *Zastosowanie korpusów w badaniu gramatyki*, w: A. Przepiórkowski, M. Bańko, R.L. Górski, B. Lewandowska-Tomaszczyk (red.), *Narodowy Korpus Języka Polskiego*, Warszawa, s. 291–301.

Górski R.L., Łaziński M., 2012, *Reprezentatywność i zrównoważenie korpusu*, w: A. Przepiórkowski, M. Bańko, R.L. Górski, B. Lewandowska-Tomaszczyk (red.), *Narodowy Korpus Języka Polskiego*, Warszawa, s. 25–37.

Grabe W., 1991, *Current Developments in Second Language Reading Research*, „TESOL Quarterly", nr 25(3), s. 357–406.
Grabe W., 2009, *Reading in a Second Language*, Cambridge.
Gruszczyński W. i in., 2015, *Measuring Readability of Polish Texts*, w: Z. Vetulani, J. Mariani (red.), *Materiały Konferencji LTC 2015 (7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznań, 27–29 listopada 2015)*, Poznań, s. 445–449.
Huey E.B., 1908, *The Psychology and Pedagogy of Reading*, New York.
Juel C., Minden-Cupp C., 2000, *Learning to read words: Linguistic units and instructional strategies*, „Reading Research Quarterly", nr 35(4), s. 458–492.
Kuhn M., Schwanenflugel P., Meisinger E., 2010, *Aligning Theory and Assessment of Reading Fluency: Automaticity, Prosody, and Definitions of Fluency*, „Reading Research Quarterly", nr 45(2), s. 230–251.
Lado, R., 1955, *Patterns of Difficulty in Vocabulary*, „Language Learning", nr 6(1), s. 23–41, https://doi.org/10.1111/j.1467-1770.1955.tb00829.x
Laufer B., 1990, *Why Some Words are More Difficult Than Others*, „IRAL", nr 28(4), s. 293–307, https://doi.org/10.1515/iral.1990.28.4.293
Lewandowska-Tomaszczyk B., 2011, *Nowe wyzwania w jakościowej i ilościowej metodologii analizy języka*, „Biuletyn Polskiego Towarzystwa Językoznawczego", nr 67, s. 141–165.
Lewandowska-Tomaszczyk B. i in., 2012, *Narodowy Korpus Języka Polskiego: geneza i dzień dzisiejszy*, w: A. Przepiórkowski, M. Bańko, R.L. Górski, B. Lewandowska-Tomaszczyk (red.), *Narodowy Korpus Języka Polskiego*, Warszawa, s. 3–9.
Perfetti C., 2007, *Reading Ability: Lexical Quality to Comprehension*, „Scientific Studies of Reading", nr 11(4), s. 357–383, https://doi.org/10.1080/10888430701530730
Piekot T., Zarzeczny G., Moroń E., 2019, *Standard „plain language" w polskiej sferze publicznej*, w: M. Zaśko-Zielińska, K. Kredens (red.), *Lingwistyka kryminalistyczna. Teoria i praktyka*, Wrocław, s. 197–214.
Ratner N.B., Berko Gleason J., Narasimhan B., 2005, *Wprowadzenie do psycholingwistyki – wiedza użytkowników języka*, w: J. Berko Gleason, N.B. Ratner (red.), *Psycholingwistyka*, Gdańsk, s. 15–60.
Reeves L.M., Hirsch-Pasek K., Golinkoff R., 2005, *Słowa i znaczenia – od pojęć pierwotnych do złożonych struktur*, w: J. Berko Gleason, N.B. Ratner (red.), *Psycholingwistyka*, Gdańsk, s. 173–240.
Rey A., Ziegler J.C., Jacobs A.M., 2000, *Graphemes are perceptual reading units*, „Cognition", nr 75(1), s. 1–12, https://doi.org/10.1016/S0010-0277(99)00078-5
Seidenberg M., 2017, *Language at the speed of sight: How we read, why so many can't, and what can be done about it*, Nowy Jork.
Seretny A., 2006, *Wskaźnik czytelności tekstu jako pomoc w określaniu stopnia jego trudności*, „LingVaria", nr 2(2), s. 87–98.
Seretny A., 2011, *Kompetencja leksykalna uczących się języka polskiego jako obcego w świetle badań ilościowych*, Kraków.
Seretny A., 2016, *Stopień trudności słowa w perspektywie glottodydaktycznej*, „Języki Obce w Szkole", nr 60(1), s. 18–25.
Sigurd B., Eeg-Olofsson M., van de Weijer J., 2004, *Word Length, Sentence Length and Frequency – Zipf Revisited*, „Studia Linguistica", nr 58(1), s. 37–52.
Taylor S.E., 1965*, Eye Movements in Reading: Facts and Fallacies*, „American Educational Research Journal", nr 2(4), s. 187–202.
Wolf M., Vellutino F., Berko Gleason J., 2005, *Psycholingwistyczna analiza czynności czytania*, w: J. Berko Gleason, N.B. Ratner (red.), *Psycholingwistyka*, Gdańsk, s. 439–477.
Wyllys R., 1981, *Empirical and Theoretical Bases of Zipf's Law*, „Library Trends", nr 30(1), s. 53–64.
Yeni-Koshian G., 2005, *Percepcja mowy*, w: J. Berko Gleason, N.B. Ratner (red.), *Psycholingwistyka*, Gdańsk, s. 121–173.

**Netografia**

http://certyfikatpolski.pl/o-egzaminie/przykladowe-testy-zbiory-zadan/ [08.05.2020].
https://cke.gov.pl/egzamin-osmoklasisty/arkusze/2019-2 [09.05.2020].
http://nkjp.pl/ [07.05.2020].

*Tomasz Moździerz*

## AVERAGE LENGTH POLISH WORD

**Keywords:** Polish language, applied linguistics, language learning/teaching, average length Polish word

**Abstract.** Considering the fact that words vary in length within a language and between different languages the author has conducted the research, inspired by Carver's model of a standard-length word, to answer the question: *What is the average length of a Polish word?* To achieve that goal, ninety samples, a 1000-words-long each, drawn from the National Corpora of Polish Language have been examined. The texts belonged to different genres and had different publication channel. The average length of a Polish word was established to be 6 characters. Using that value, the number of *average words* (AV) has been calculated for each analyzed sample. The number of actual words has been juxtaposed with the number of AV. The procedure made possible to reject the null hypothesis and validate the new unit of text measurement. In the last part of the article a number of possible uses of the new concept have been enumerated.

Data wpłynięcia tekstu: 4.06.2020