ACTA UNIVERSITATIS LODZIENSIS
KSZTAŁCENIE POLONISTYCZNE CUDZOZIEMCÓW 28, 2021

https://doi.org/10.18778/0860-6587.28.22

C O P E
Member since 2019
JM14480

*Tomasz Możdzierz**

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

PROPORCJA LICZBY WYRAZÓW FAKTYCZNYCH I PRZECIĘTNYCH TEKSTU W KONTEKŚCIE SIEDMIOSTOPNIOWEJ SKALI TRUDNOŚCI JASNOPIS.PL

Słowa kluczowe: stopień trudności tekstu, zrozumienie tekstu, miara stopnia trudności tekstu, przeciętne słowo

Streszczenie. Długość tekstu liczona w wyrazach faktycznych (WF) i przeciętnych (PPW, zob. Możdzierz 2020) nie jest identyczna, a stosunek WF do PPW zmienia się w zależności od stopnia trudności wypowiedzi. W badaniach opisanych w niniejszym artykule stosunek ten odniesiono do siedmiopunktowej skali trudności tekstu programu jasnopis.pl (jest to ogólnodostępny algorytm przeznaczony do szacowania stopnia czytelności polszczyzny pisanej). Analiza tekstów autentycznych o łącznej objętości dziesięciu tysięcy WF dla każdego poziomu trudności według w/w skali pozwoliła pokazać, jak stosunek ten zmienia się wraz z rosnącym poziomem trudności tekstu. Opracowany sposób analizy można wykorzystywać w procesie kształcenia językowego, by w sposób szybki i łatwy wstępnie szacować poziom trudności dowolne wypowiedzi pisemnej w języku polskim.

1. WPROWADZENIE

„Obywatele współczesnych społeczeństw, by odnieść sukces, muszą umieć dobrze czytać. Czytanie nikomu wprawdzie sukcesu nie gwarantuje, ale jego osiągnięcie bez opanowania tej umiejętności jest zdecydowanie trudniejsze”[1] (Grabe, 2009, s. 5). Proces czytania jest złożony (zob. Alderson 2000), a na jego efektywność wpływa wiele czynników (zob. Wolter 2017). Jednym z nich jest stopień trudność

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

[1] Wszystkie tłumaczenia, o ile nie zaznaczono inaczej, są autorskie.

Received: 4.05.2021; verified: 3.07.2021. Accepted: 8.07.2021.

tekstu. Xiaobin Chen i współpracownicy (2018, s. 487) twierdzą nawet, że jest ona najistotniejszym ze wszystkich parametrów, gdyż, ich zdaniem, jakość procesu czytania (a więc czas spędzony na lekturze, osiągnięty poziom rozumienia, motywacja towarzysząca lekturze) zależy od stopnia dostępności językowej tekstu, a więc od poziomu jego zrozumiałości. J. Charles Alderson (2000, s. 35) z kolei pisze, że „[k]oniecznością męczenia się z tekstem z powodu nieznanych słów w sposób oczywisty niekorzystnie wpływa na jego rozumienie, odbiera również całą przyjemność z czytania". Konsekwencją źle dobranego przez nauczyciela tekstu może więc być niewielki przyrost kompetencji językowych uczących się, frustracja i demotywacja do nauki. Zbyt wysoki stopień trudności tekstów może też stać się przyczyną słabej sprzedaży gazet czy też niewielkiej poczytności artykułów w witrynach internetowych.

Ocena poziomu trudności tekstu wymaga przeprowadzenia analizy jego struktury. Można to zrobić za pomocą testów takich jak Taylor cloze-procedure lub rozmaitych algorytmów, np. wskaźnika Flescha, indeksu FOG Gunninga czy też programu jansopis.pl. Pomocne w tym być może również wykorzystanie konstruktu 'średniego / przeciętnego wyrazu'[2], tj. kontrolowanej pod względem długości jednostki tekstu (zob. Carver 1972; Seidenberg 2017, s. 83; Brysbaert 2019, s. 25; Możdzierz 2020). Z badań wynika bowiem, że jeśli liczba wyrazów faktycznych (dalej: WF) w tekście jest mniejsza od obliczonej liczby jednostek standaryzowanych pod względem długości, tekst powinien być klasyfikowany jako „trudniejszy", natomiast w sytuacji odwrotnej jako „łatwiejszy" (zob. Carver 1977–78; Brysbaert 2019). W niniejszym artykule hipoteza ta zostanie zweryfikowana dla polszczyzny. Opisane w nim badania miały na celu sprawdzenie, czy i w jaki sposób stosunek liczby WF do przeciętnych (dalej: PPW[3]) w polskich tekstach koreluje z poziomem trudności tekstu mierzonym algorytmem programu jasnopis.pl.

2. STAN BADAŃ

Poziom trudności komunikatu można rozpatrywać na dwóch płaszczyznach. Pierwszą jest trudność zrozumienia tworzących go wyrazów, natomiast drugą – trudność całego tekstu, nazywana również poziomem zrozumiałości (ang. *readability*). Ten ostatni definiuje się na różne sposoby, na przykład jako:

[2] Badacze nazywali ten konstrukt w różny sposób. Ronald Carver (1976) proponował „wyraz standardowej długości" (ang. standard length word, W), Mark Seidenenberg (2017, s. 83) nie nadał konstruktowi konkretnej nazwy, lecz jedynie przyjmował, że większość wyrazów w tekstach mierzy 5 znaków, natomiast Marc Brysbaert (2019) pisał o średniej długości słowa (ang. the average word length for English). Ponieważ cel i sposób wykorzystania tych trzech jednostek jest właściwie identyczny, standaryzowana pod względem długości jednostka tekstu nazywana będzie w artykule 'wyrazem przeciętnym'.

[3] Od: przeciętny polski wyraz.

– „[s]uma wszystkich elementów, które wpływają na rozumienie, tempo czytania i poziom zainteresowania tekstem czytelnika" (Chen i in. 2018, s. 487);
– „[t]o, co czyni tekst łatwiejszym od innych" (DuBay 2004, s. 3);
– „stopień, w jakim ludzie uznają pewne teksty za wciągające i zrozumiałe" (ibid.);
– „suma elementów, które sprawiają, że jakaś grupa ludzi uznaje tekst za zrozumiały, interesujący i możliwy do lektury w optymalnym dla nich tempie" (ibid.).

Jak widać, każda z definicji uwzględnia wpływ stopnia trudności tekstu na rozumienie jego treści przez odbiorców. W większości zaznacza się przy tym, że jeśli treść ta jest dla czytelników interesująca, są zazwyczaj bardziej zmotywowani, by się z nią mierzyć (Grabe 2009, s. 181; Alderson 2000, s. 44; Charzyńska 2015). Teksty zbyt trudne i mało ciekawe są zaś dla nich frustrujące i demotywujące, co skłania do zarzucenia lektury, a przecież wiadomo, że czytelnicy, którzy mało czytają, nie rozwijają swoich kompetencji językowych i czytelniczych. Analogicznie, ci odbiorcy, którzy są zmotywowani i czytają dużo, dzięki swojej praktyce i rosnącemu doświadczeniu będą robić to coraz sprawniej.

2.1. PŁASZCZYZNA KWANTYFIKOWALNYCH ELEMENTÓW JĘZYKOWYCH

Mimo swojej istotności kryterium tematyki jest całkowicie subiektywne, gdyż dla każdego odbiorcy co innego jawić się może jako interesujące. Analizując poziom trudności tekstu, szuka się więc elementów obiektywnych, które oprą się zróżnicowaniu jednostkowemu. Dlatego też najczęściej badaniu poddaje się rozmaite kwantyfikowalne jednostki lingwistyczne. Takie podejście nie jest doskonałe, bez liczb jednak nie da się stworzyć rzetelnego, systemowego opisu parametru, jakim jest trudność tekstów.

Dotychczas w badaniach wykorzystywano następujące miary:

– długość słów liczoną liczbą sylab lub znaków (trudność leksykalna),
– długość zdań (trudność składniowa),
– przynależność słów do określonych przedziałów frekwencji (trudność leksykalna).

2.1.1. Długość słów lub średnia liczba znaków na słowo w tekście (tzw. trudność leksykalna) (Crossley i in. 2019, s. 542)

W środowisku naukowym panuje konsensus co do tego, że im wyraz jest dłuższy, tym trudniejszy do rozpoznania, zapamiętania i przywołania z pamięci. Z tego też względu krótsze słowa są na co dzień używane zdecydowanie częściej niż dłuższe (zob. Lado, 1955; Laufer 1990, s. 297–298; Ellis 2002; Reeves i in. 2005, s. 184; Sigurd i in. 2004; Seretny 2006, 2016; Broda i in. 2014; Charzyńska, Dębowski 2015; Dębowski i in. 2015; Gruszczyński i in. 2015). Sformułowania „krótsze" i „dłuższe" są jednak w dalszym ciągu nieprecyzyjne, gdy brakuje danych, jak długie jest słowo przeciętnej długości. W swojej pracy, Marc Brysbaert (2019, s. 25) stworzył konstrukt angielskiego wyrazu przeciętnej długości (4.6 znaku), tj. abstrakcyjnej jednostki, do której można porównywać słowa w języku faktycznie istniejące. Dzięki temu można precyzyjnie określić, które słowo 'dłuższe' lub 'krótsze'. Pracujący wcześniej nad tym zagadnieniem Mark Seidenberg (2017, s. 83) przyjął, że średni wyraz w języku angielskim ma 5 znaków. Prawie 50 lat wcześniej Ronald Carver (1972, 1976, 1977–78) ustalił zaś, dość jednak arbitralnie, iż przeciętny angielski wyraz składa się z 6 znaków. Konstrukt ten, który badacz ten wykorzystywał następnie z powodzeniem w badaniach tempa czytania, nazwał wyrazem standardowej długości ('W'), odróżniając go od tych faktycznie występujących w tekście ('w').

Dzięki wskazaniu długości średniego wyrazu ustalono obiektywny punkt odniesienia. Jeśli więc w tekście wyrazów faktycznych jest więcej niż by to wynikało z liczby znaków, większość jednostek musi być krótka, a więc, zgodnie z przyjętą wyżej zasadą – łatwa. W sytuacji odwrotnej, jeśli spodziewana liczba wyrazów średnich / standardowych jest wyższa od faktycznej, oznacza to, że większość słów w wypowiedzi jest długa, a więc zarazem trudna (zob. Carver 1977–78, s. 28).

Badacze polszczyzny zajmujący się tym problemem stosowali do tej pory miarę sylab, ustalając, że średnia długość wyrazu wynosi 3 sylaby (Pisarek 1972; Seretny 2006; Broda i in. 2014; Charzyńska, Dębowski 2015; Dębowski i in. 2015; Gruszczyński i in. 2015). Sylaba jest jednak jednostką nieprecyzyjną i mocno zróżnicowaną pod względem długości. Wzorując się więc na wspomnianych wyżej konstruktach wyrazu średniego / wyrazu standardowej długości, na podstawie analizy reprezentatywnego zbioru jednostek wyekscerpowanych z Narodowego Korpusu Języka Polskiego, obliczyłem, że przeciętny polski wyraz liczy 6 znaków (zob. Możdzierz 2020).

2.1.2. Średnia liczba wyrazów w zdaniu (tzw. trudność składniowa) (Crossley i in. 2019)

Średnia liczba wyrazów w zdaniu to parametr, który pozwala oszacować trudność poziom złożoności składni. Im zdania są dłuższe i bardziej złożone, tym trudniejsze, w odróżnieniu od łatwych do przetworzenia krótkich zdań pojedynczych. Wraz z rosnącym skomplikowaniem struktury tekstu pojawia się w nim coraz więcej spójników, wyrażeń przyimkowych oraz zaimków, które zwiększają trudność jego odbioru (Dale, Tyler 1935, s. 397; Sung i in. 2015, s. 377). Choć zaimki nie są długimi wyrazami, to ich obecność, zwłaszcza tych nieokreślonych, w tekście angielskim, znacząco podnosi jego trudność (Dale, Tyler 1934).

2.1.3. Odsetek słów z róznych przedziałów frekwencji w tekście (Sung i in. 2015; DuBay 2004)

Istotny parametr wpływający na czytelność tekstu stanowi relacja między stopniem trudności słów a ich miejscem na listach frekwencyjnych. Parametr ten znany jest już od lat 20. XX w. Wówczas to Edward Thorndike wydał *Teacher's Word Book*, tj. pierwszą listę frekwencyjną dla języka angielskiego, która stała się podstawą skalowania trudności tekstu (DuBay 2004, s. 12). Wyrazy najczęściej używane, tj. zajmujące najwyższe miejsca na listach frekwencyjnych (przedziały 1–1000, 1001–2000), znamy najlepiej ze względu na wzmożony recykling językowy. Bez problemu przychodzi nam więc ich przywoływanie; pokrywają one nawet 80% większości przeciętnych tekstów mówionych i pisanych (Seretny 2006, s. 19). Wypowiedzi specjalistyczne wymagają jednak obecności słów tematycznych, a te z reguły zajmują dalsze na listach frekwencyjnych (Chen i in. 2018, s. 491). Mimo dalekich miejsc, mogą mieć one jednak wysoką frekwencję 'dziedzinową'. Sama więc informacja na temat tego, jak często jakieś słowo występuje, na przykład na milion jednostek w korpusie językowym, może prowadzić do przypisania mu nadmiernego stopnia trudności lub niedoszacowania stopnia jego skomplikowania. Analizując poziom trudności tekstów tematycznych, należy więc prowadzić badania dwutorowo, tj. ustalać ich przystępność dla ogółu użytkowników języka oraz dla grup ekspertów. Chen i współpracownicy (ibid.) proponują na przykład, by sprawdzać, jak często np. dany termin używany w dyskursie prawniczym występuje na milion wziętych z korpusu ogólnego i na milion pochodzących z korpusu aktów prawnych. Taka metoda może dać lepszy ogląd poziomu trudności leksemu. Inne sposoby uwzględnienia kontekstu proponowali wcześniej Edgar Dale i Ralph Tyler (1934), licząc w badanych tekstach

słowa techniczne, skomplikowane słowa nietechniczne oraz zaimki. William Gray i Bernice Leary (1935, s. 115), natomiast celem ustalenia poziomu czytelności, obliczali odsetek słów nieznanych dla 90% populacji uczniów ówczesnej szóstej klasy amerykańskiego systemu edukacji[4].

2.2. ALGORYTMY STOPNIA TRUDNOŚCI TEKSTU

Wykorzystując wymienione wyżej policzalne komponenty lingwistyczne, badacze opracowali szereg sposobów, pozwalających oszacować, jak zrozumiały dla „przeciętnego odbiorcy" będzie konkretny tekst.

2.2.1 Formuła Rudolpha Flescha[5]

Wzór do obliczeń poziomu trudności:
206.835 – (1.015 x ASL) – (84.6 x ASW)
Gdzie: ASL = średnia długość zdania (liczba słów/liczba zdań)
ASW = średnia liczba sylab na wyraz (liczba sylab/liczba wyrazów)
wynik między 0 (trudny) a 100 (łatwy), gdzie 30 = bardzo trudny, a 70 = odpowiedni dla dorosłego odbiorcy

2.2.2. Druga formuła Rudolpha Flescha

Wzór do obliczeń poziomu łatwości czytania:
1.599NOSW – 1.015sl – 31.518
Gdzie: NOSW = liczba słów jednosylabowych/100 wyrazów
Sl = średnia długość zdania (w słowach)

2.2.3. Wskaźnik mglistości Roberta Gunninga

Wzór do obliczeń poziomu trudności –> 3.0680 + 0.877 (średnia długość zdania) + 0.984 (procent słów jednosylabowych)

[4] Informacja o znanych i nieznanych szóstoklasistom wyrazach pochodziła z niepublikowanego badania Dale'a, w którym sprawdził on znajomość 8000 najczęstszych angielskich wyrazów w grupie 7878 uczniów szóstej klasy (Gray i Leary 1935, s. 101).

[5] Przykłady 2.2.1–2.2.3 pochodzą z pracy DuBay'a (2004, s. 21).

2.2.4. Formuła Walerego Pisarka (Gruszczyński i in. 2015, s. 1)

Wzór do obliczeń:

$$T = \frac{\sqrt{T_s^2 + T_w^2}}{2}$$

Gdzie: T = poziom trudności tekstu;
Ts = procent słów zawierających 4 lub więcej sylab w lemmacie
Tw = średnia długość zdania (w słowach)

Formuła ta została wykorzystana w aplikacji Jasnopis. Powstała na jej podstawie skala trudności tekstów liczy siedem poziomów. Każdy z nich ma swoją charakterystykę opisową (zob. poniżej), koresponduje też z kolejnymi etapami procesu kształcenia sprzed reformy systemu edukacji z 2017 r.

2.2.5. Algorytm Pracowni Prostej Polszczyzny UWr (Piekot 2019, s. 209–210):

Algorytm PPP określa poziom trudności tekstu, łącząc wyniki indeksu mglistości, odsetka słów trudnych i średniej długości zdania. Stworzony przez Pracownię prestiżowy Certyfikat Prostej Polszczyzny jest uznawany przez polski rząd i, jak na swojej stronie wspominają autorzy, spełnia on „warunki światowego standardu prostego języka[6].

3. METODOLOGIA BADAŃ

Ścisłe, matematyczne podejście do tekstu uważane jest z reguły zbyt mechaniczne, gdyż opiera się wyłącznie na liczbach. Z tego względu często też podlega krytyce (zob. Chen i in. 2018, s. 487; Sung i in. 2015, s. 373; Crossley i in. 2019, s. 542). Algorytmy jednak dobrze korelują z wynikami testów rozumienia i, choć, jak pisze William DuBay (2004, s. 3): „można się kłócić o ich dobór", to w istocie rzeczy tylko one „dają rzetelny przewidywany poziom trudności tekstu". Często są one jednak skomplikowane i nie zawsze dostępne. Alternatywę dla nich stanowić może wykorzystanie stosunku liczby WF (wyrazów faktycznych) do liczby PW (przeciętnych wyrazów). Carver (1977–78, s. 28) twierdził, że powinien on spadać wraz ze wzrostem poziomu trudności tekstu.

[6] http://ppp.uni.wroc.pl/certyfikat.html (dostęp: 12.12.2020).

Podjęte przez mnie badania miały na celu sprawdzenie, czy i w jaki sposób proporcja PPW/WF oraz WF/PPW korelują z algorytmem szacującym poziom czytelności, jaki został wykorzystany w programie jasnopis.pl[7]. Za wyborem Jasnopisu jako punktu odniesień przemawia kilka ważkich argumentów:

- jest to algorytm stworzony specjalnie dla języka polskiego,
- metodologia jego tworzenia została przez autorów jasno wyłożona,
- program jest łatwo dostępny w Internecie.

Tabela 1. *Poziom trudności tekstu według programu jasnopis.pl i jego opis*

POZIOM TRUDNOŚCI	OPIS	ORIENTACYJNE WYMAGANE WYKSZTAŁCENIE ODBIORCY
1	tekst dziecinnie łatwy	klasy 1–3 szkoły podstawowej
2	tekst bardzo łatwy	klasy 3–6 szkoły podstawowej
3	tekst łatwy, zrozumiały dla przeciętnego Polaka	gimnazjum (obecnie klasy 7–8 szkoły podstawowe)
4	tekst nieco trudniejszy, zrozumiały dla osób z wykształceniem średnim lub mających duże doświadczenie życiowe	liceum
5	tekst trudniejszy, zrozumiały dla ludzi wykształconych	studia licencjackie/inżynierskie
6	tekst trudny w odbiorze dla przeciętnego Polaka	studia magisterskie
7	tekst bardzo skomplikowany, fachowy, którego zrozumienie może wymagać wiedzy specjalistycznej	doktorat lub specjalizacja w dziedzinie, której dotyczy tekst

Źródło: opracowanie własne na podst. jasnopis.pl

Dzięki uprzejmości autorów programu Jasnopis możliwe było wykorzystanie jego pełnej wersji (tj. bez limitu znaków), dzięki czemu w przystępny sposób analizowane mogły być obszerne teksty. Siedmiostopniowa skala trudności programu jasnopis.pl (zob. tabela 1) pozwoliła zaś pokazać, jak zmienia się stosunek liczby WF (wyrazów faktycznych) do liczby PPW (przeciętnych polskich wyrazów) oraz stosunek liczby PPW do liczby WF w polskich tekstach.

Działań matematycznych dokonywano w programie R; wykorzystano też funkcję *statystyka wyrazów* programu Microsoft Word, która dostarczyła infor-

[7] Podobne badania można by przeprowadzić z wykorzystaniem dowolnej innej formuły. Jasnopis wydał się jednak najbardziej przystępny, zwłaszcza w przypadku obliczeń dokonywanych komputerowo.

macji o liczbie wyrazów oraz liczbie znaków bez spacji w każdym tekście[8]. Przyjmując, że 1 PPW = 6 znaków (zob. Moździerz 2020), policzono liczbę PPW dla każdego tekstu. Następnie dla każdego obliczono także stosunek liczby PPW do WF oraz WF do PPW. Policzono również średnie uzyskanych wyników oraz zsumowano liczbę WF i PPW dla każdego poziomu. Na koniec zestawiono współzależność obliczonych wartości z poziomem trudności tekstu na skali Jasnopisu, wyprowadzając korelację Pearsona.

Ze względu na to, że nie da się ustalić stopnia czytelności tekstu przed jego analizą, doboru tekstów przynależnych do poszczególnych poziomów dokonywano metodą prób i błędów, kierując się wytycznymi opisu zamieszczonego w instrukcji programu Jasnopis (zob. tabela 1). Łącznie przeanalizowano 7 korpusów tekstów (dla każdego poziomu programu Jasnopis) o objętości ~10 tysięcy WF **każdy**.

Teksty z poziomu 1. zostały w Jasnopisie zdefiniowane jako „dziecinnie łatwe", zaś z 2. jako „bardzo łatwe", toteż w badaniach wykorzystane zostały bajki dla najmłodszych. Do poziomu 1. zakwalifikowały się takie pozycje jak: wybrane rozdziały „Kubusia Puchatka", „Chatki Puchatka" czy przygód „Mikołajka" oraz wiersze dla dzieci Juliana Tuwima, natomiast w poziomie 2 znalazły się baśnie braci Grimm. Opis poziomów 3–5 sugeruje, że wypowiedzi o tym stopniu trudności są zrozumiałe dla Polaków z wykształceniem, odpowiednio, podstawowym (autorzy mieli na myśli wykształcenie uzyskiwane po ukończeniu edukacji w istniejącym jeszcze wówczas gimnazjum), średnim oraz na poziomie studiów pierwszego stopnia. Mimo użytego w opisach przymiotnika „wykształcony" w odniesieniu do czytelnika, wyszedłem z założenia, że teksty o tym stopniu trudności nie powinny być wysoce specjalistyczne. Najlepszym źródłem ekscerpcji wydały mi się więc artykuły prasowe, ze względu na ich szeroką obecność w codziennym życiu. Większość materiałów (7 z 11), którym narzędzie Jasnopis przypisało poziom od 3 do 4, pochodzi z wydania specjalnego „Tygodnika Powszechnego" (z tzw. Kanonu „Tygodnika Powszechnego" zbierającego najlepsze teksty z lat 1945–2015). W materiałach, którym przypisany został poziom 5. znalazły się zarówno teksty publicystyczne, jak i niektóre abstrakty artykułów naukowych. Teksty z poziomu 6. i 7. zostały w programie opisane jako wymagające od czytelnika wiedzy specjalistycznej z poziomu studiów magisterskich i doktorskich, toteż analizie poddano głównie artykuły naukowe, ich abstrakty

[8] Trzeba tu zaznaczyć, że liczba znaków bez spacji w funkcji „statystyka wyrazów" uwzględnia znaki interpunkcyjne, tym samym zawyżając nieco uzyskaną liczbę PPW. Należy jednak pamiętać, że niezależnie od poziomu trudności w języku polskim zdanie zawsze kończy pojedynczy znak (kropka, pytajnik, wykrzyknik), jest to więc pewna stała. Liczba innych znaków interpunkcyjnych w tekście zawsze będzie stanowić ułamek wszystkich znaków, a jej wzrost łączy się ze zwiększoną trudnością składniową tekstu, w związku z czym, proporcja WF/PPW zostanie jedynie uwydatniona. Nie jest to więc wada, lecz zaleta PPW. Różnica złożoności zdań na różnych poziomach trudności tekstu jest tematem ciekawym, wymagającym jednak inaczej ukierunkowanych badań.

o objętości przynajmniej 100 wyrazów oraz artykuły o maksymalnej całkowitej objętości 5 stron z pominięciem streszczeń, przypisów oraz wymienionej bibliografii. Jako źródło tekstów najbardziej wymagających posłużyło repozytorium prac naukowych UJ[9].

4. WYNIKI I ICH INTERPRETACJA

W tabeli 2 widnieją dane liczbowe dotyczące zebranych materiałów dla każdego z siedmiu poziomów skali jasnopis.pl.

Tabela 2. *Stosunek liczby WF do liczby PPW i WF/PPW zależnie od poziomu tekstu na skali jasnopis.pl*

POZIOM TEKSTÓW NA SKALI JASNOPIS.PL	LICZBA WF	LICZBA PPW	STOSUNEK LICZBY WF DO LICZBY PPW	STOSUNEK LICZBY PPW DO LICZBY WF
1	10038	8472	1.208156	0.8315218
2	10050	8733	1.1553	0.8658834
3	10035	9605	1.034514	0.9672754
4	10354	9890	1.04314	0.9610405
5	10067	11009	0.9155522	1.093398
6	10062	10924	0.922103	1.08915
7	10115	11205	0.8719228	1.15185

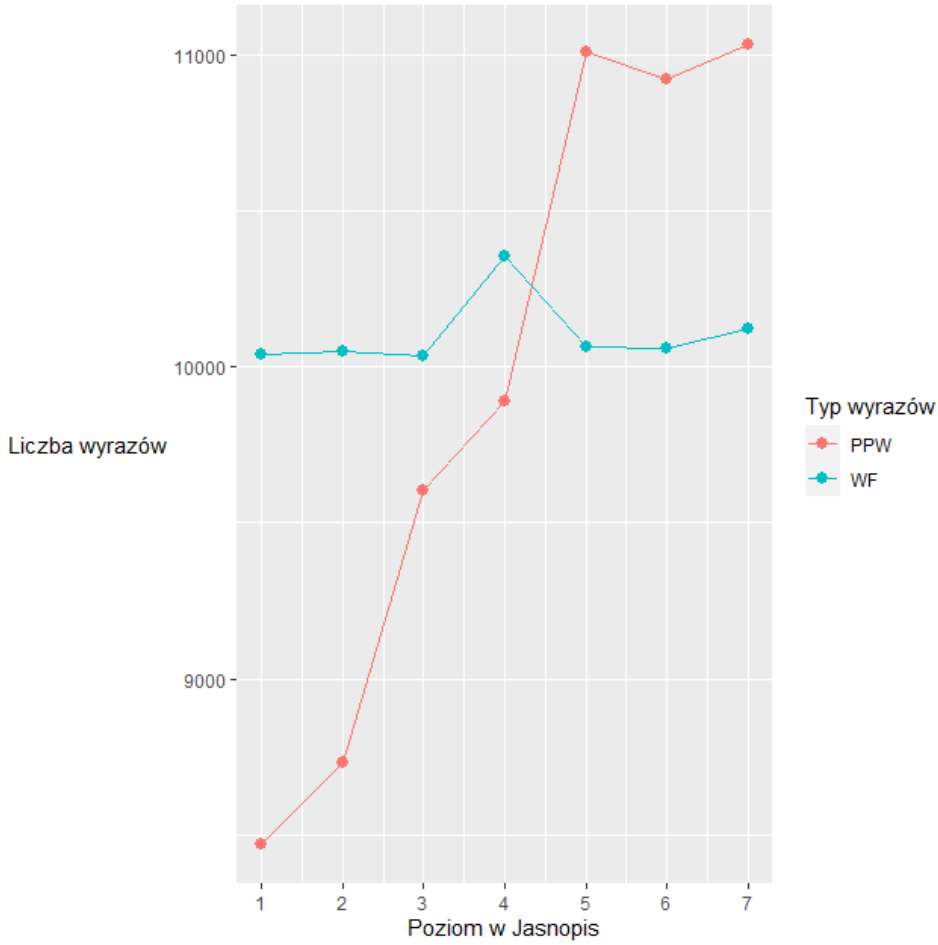
Źródło: opracowanie własne

Dla określenia stosunku WF/PPW i stopnia trudności tekstu na skali Jasnopis została wyprowadzona korelacja Pearsona. Wyniosła ona -0.8913494 przy poziomie istotności $p < 0.00000000000000022$. Korelacja dla stosunku PPW/WF i stopnia trudności tekstu według skali jasnopis.pl wyniosła zaś 0.8967218, przy poziomie istotności $p < 0.00000000000000022$.

Zmianę liczby PPW przy względnie stałej liczbie WF zależnie od poziomu trudności tekstu ilustruje wykres 1.

[9] Pełna lista wykorzystanych materiałów dostępna jest u autora artykułu. Nie została zawarta w niniejszym tekście ze względu na ograniczoną objętość publikacji.

Wykres 1. *Liczba wyrazów faktycznych, a liczba PPW na siedmiu poziomach trudności tekstu w skali jasnopis.pl*

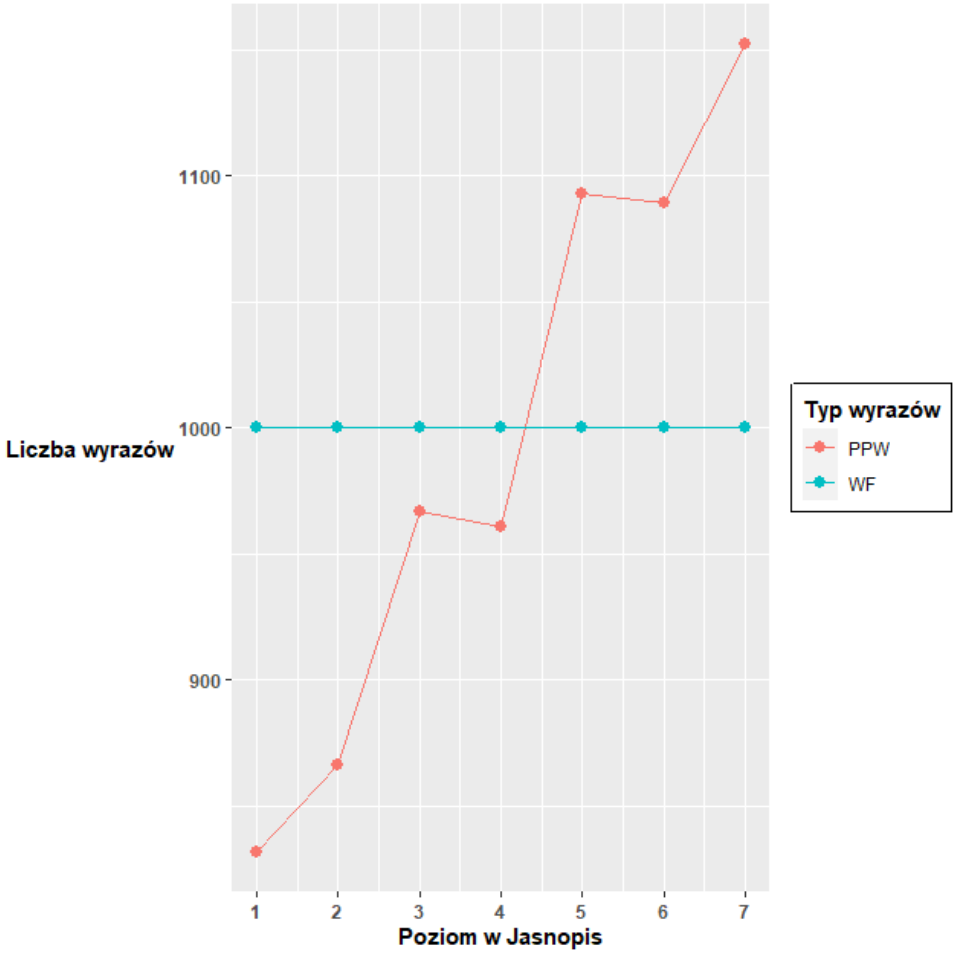


Źródło: opracowanie własne

Jak wynika z danych zamieszczonych w tabeli i wykresu, zgodnie z przytoczonymi wcześniej przewidywaniami badaczy, przy tej samej liczbie znaków wraz z rosnącym poziomem trudności tekstu rośnie liczba PPW i stosunek PPW do WF, a maleje WF do PPW. Liczba WF w zbiorach tekstów na siedmiu poziomach trudności pozostawała względnie stała, widoczne jest jedynie niewielkie odchylenie na poziomie 4. Całkowita liczba WF była tam o ~300 większa niż w innych próbkach, co wynikało z mniejszej liczby długich tekstów poddanych analizie. Ten właśnie poziom jest też najbardziej „przeciętny", tj. stosunek PPW /WF i WF/PPW wynosi w ich przypadku niemal 1:1.

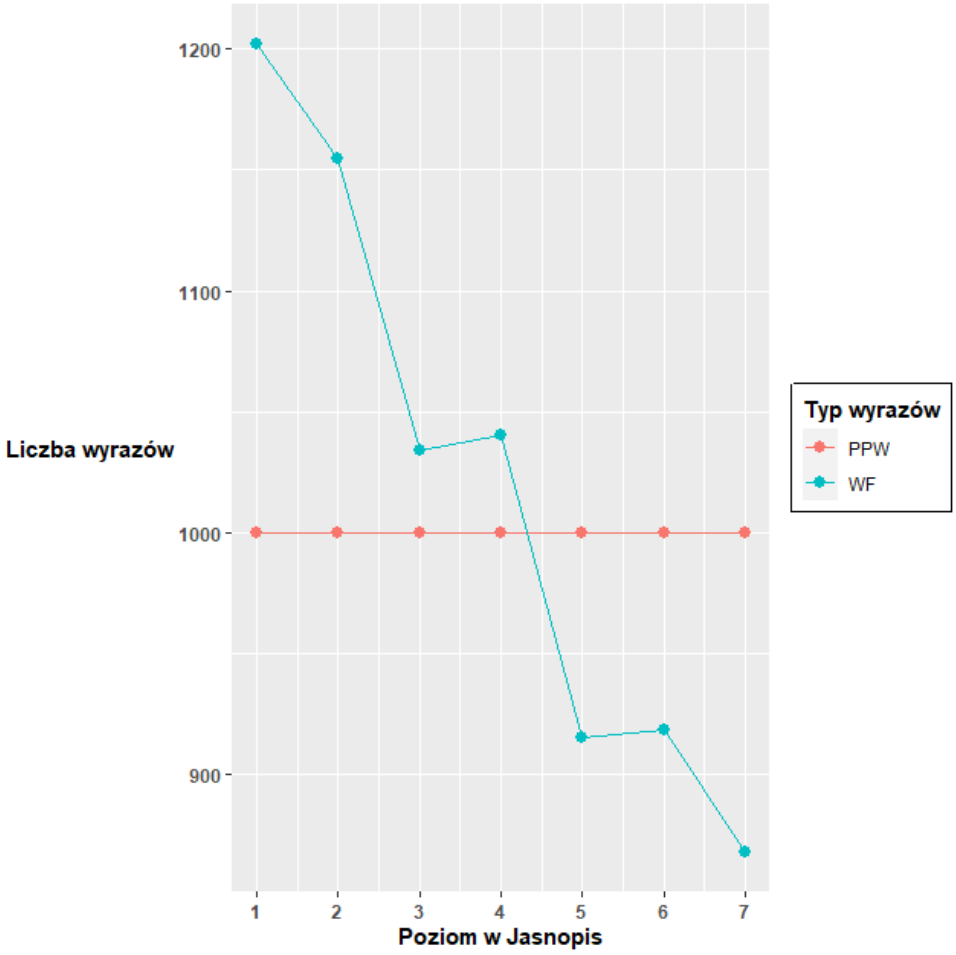
Wykres 1, przedstawiający **faktyczne** wyniki eksperymentu, wskazuje na istnienie trzech ‘zakresów trudności tekstu”, tj. tekstów łatwych (poziomy 1 i 2), średnich/przeciętnych (poziomy 3 i 4) oraz trudnych (poziomy 5–7). Wykres 2 pokazuje natomiast, jak będzie się zmieniać liczba PPW dla **dowolnego** 1000 WF na każdym poziomie trudności, a wykres 3 zmieniającą się liczbę WF dla dowolnego 1000 PPW. Te dwa ostatnie to wyabstrahowane z kontekstu modele. Poziomy 4 i 6 na obu tych wykresach łamią trend rosnący dla liczby PPW i WF (zależnie od wykresu). Trzeba tu jednak pamiętać, że korelacja stosunku PPW/WF i WF/PPW do skali Jasnopis nie była idealna. Jednocześnie, dokładne różnicowanie poziomów to zadanie właśnie dla zaawansowanych algorytmów, które są w stanie uchwycić w tekstach najdrobniejsze szczegóły. Dla człowieka, „ręczne” rozróżnianie poziomów 1 i 2, 3 i 4 czy 5 i 6 stanowiłoby mozolne zadanie, wymagające dużych nakładów pracy, najprawdopodobniej niewspółmiernych do zysków. O ile więc uzyskane w eksperymencie wyniki nie korelują idealnie z Jasnopisem, o tyle, jak stwierdzono wyżej, dobrze wyznaczają ‘zakresy trudności tekstu’. Na podstawie wykresu 2 i 3 można pokusić się o uznanie nawet czterech takich zakresów: teksty łatwe (poziom 1 i 2), średnie (poziom 3 i 4), trudne (poziom 5 i 6) i bardzo trudne (poziom 7). Wykorzystanie jako wskaźnika stosunku PPW/WF i WF/PPW umożliwia więc szybkie umiejscowienie tekstu na takiej skali trudności, co może być przydatne w trakcie poszukiwania odpowiedniego tekstu do wykorzystania w trakcie lekcji języka polskiego, przy adaptacji tego tekstu dla uczniów na niższych poziomach znajomości polszczyzny, lub po prostu, gdy nasz tekst jest zbyt długi, by zmieścić się w Jasnopisie.

Wykres 2. *Liczba PPW przypadających na każdy 1000 WF na siedmiu poziomach trudności tekstu skali jasnopis.pl*



Źródło: opracowanie własne

Wykres 3. *Liczba WF przypadających na każdy 1000 PPW na siedmiu poziomach trudności tekstu skali jasnopis.pl*



Źródło: opracowanie własne

5. OGRANICZENIA I PERSPEKTYWY BADAWCZE PROJEKTU

Wykorzystanie PPW w ustalaniu poziomu czytelności tekstu jest działaniem mechanicznym i całkowicie ignorującym treść. Jest to jednak mankament większości formuł. Kierując się kryterium długości wyrazu, niezależnie od jednostki jaką jest wyrażane (znaki, sylaby), czy też analizując listy frekwencyjne, łatwo

o przeoczenia. Dla obcokrajowców uczących się naszego języka internacjonalizmy, zazwyczaj długie, często są bardziej zrozumiałe od ich polskich odpowiedników. Przykładami mogą tu być chociażby takie pary wyrazów jak: *ogólnie – generalnie*; *inteligentny – mądry*; *element – część*. Trudno też zgodzić się, że słowa takie jak *dziwaczka*, *tłuczony*, *tresowany* czy *gronostajowy* są wyrazami dziecinnie prostymi, mimo że znajdują się w tekstach, które Jasnopis przypisuje do poziomu 1. Dla polskiego dziecka jednak słowo *tresowany* może być bardziej zrozumiałe niż przymiotnik *charakterystyczny*, łatwy z kolei w odbiorze dla ucznia znającego język angielski. Tego typu problemów nie sposób pokonać inaczej niż na drodze wnikliwej lektury każdego analizowanego tekstu, dokonywanej z myślą o profilu odbiorcy.

Krytyce można też poddać wielkość próby oraz dobór analizowanych materiałów Zapewnienie podobnego rozmiaru próbie dla każdego poziomu wymagało wyznaczenia limitu. Wartość 10 000 jednostek faktycznych (WF), choć została przyjęta arbitralnie, wydawała mi się wystarczająca, by zapewnić rzetelność pomiarom.

Lektura analizowanych tekstów pokazała mi też, jak trudno uchwycić różnicę w poziomie trudności niektórych tekstów, na przykład bajek (np. o *Śpiącej królewnie* i *O wilku i siedmiu koźlątkach*, oba utwory autorstwa braci Grimm), czy też artykułów naukowych. Dopiero szczegółowa analiza algorytmem Jasnopisu pozwala dostrzec różnice w parametrach lingwistycznych[10].

Ciekawie też rysuje się perspektywa przyszłych badań, w których można byłoby porównać dane uzyskane w ramach powyższych analiz oraz dostarczone przez algorytm jasnopis.pl z poziomem trudności tekstów subiektywnie ocenianym przez rodzimych użytkowników (zob. Sung i in. 2015). Z tekstów trzeba by jednak koniecznie usunąć tytuły oraz informacje o autorze ze względu na możliwe skojarzenia. Wszak nazwiska takie jak Jerzy Turowicz, Leszek Kołakowski, Andrzej Stasiuk czy ks, Adam Boniecki przywodzą na myśl treści podniosłe i intelektualnie wysublimowane, co mogłoby zniekształcić subiektywną ocenę poziomu niekoniecznie trudnego tekstu.

[10] W wykazie źródeł (dostępnym u autora) wymienione zostały tylko te teksty, które wykorzystano w obliczeniach, nie zaś wszystkie poddane ocenie programem jasnopis.pl. Wynika to z przyjętego limitu ~10 tysięcy wyrazów na poziom oraz faktu, że dobór tekstów mógł odbywać się jedynie metodą prób i błędów. Podczas poszukiwań bardzo wiele materiałów plasowało się na poziomach 3 i 4. Również różnice między poziomami 6 i 7 oraz 1 i 2 okazały się niezwykle trudne do intuicyjnego uchwycenia. Celem uniknięcia nadreprezentacji po osiągnięciu progu 10 tysięcy wyrazów dla poziomu, każdy zbiór był więc zamykany, a żaden tekst powyżej tego limitu nie został wykorzystany do obliczeń ani też uwzględniony w liście źródeł.

6. PODSUMOWANIE

Analiza ~10 tysięcy wyrazów na każdym z 7 poziomów skali jasnopis.pl pozwoliła uchwycić zmianę stosunku WF/PPW oraz PPW/WF wraz ze zmieniającym się poziomem trudności tekstów, umożliwiając równocześnie ostrożne przypisanie ich do trzech lub czterech sfer trudności. Choć badanie nie było wolne od ograniczeń, stanowi novum w językoznawstwie polskim, a opracowana metoda pozwala na szybkie, wstępne szacowanie stopnia trudności dowolnego tekstu. Cel badania został więc osiągnięty, a wyniki okazały się względnie satysfakcjonujące. Nowy w językoznawstwie polskim konstrukt PPW zyskał empiryczne potwierdzenie swojej użyteczności. Opracowana też została łatwa i szeroko dostępna metoda szacowania poziomu trudności tekstów. Zgodnie z nią należy:

1. wynotować z tekstu liczbę znaków bez spacji oraz liczbę wyrazów (WF)
2. obliczyć w tekście liczbę PPW, tj. podzielić liczbę znaków bez spacji przez 6: liczba PPW = $\frac{Liczba\ znaków\ bez\ spacji}{6}$
3. obliczyć stosunek liczby WF do liczby PPW lub liczby PPW do liczby WF: $\frac{PPW}{WF}\ lub\ \frac{WF}{PPW}$
4. porównać uzyskany wynik (ułamek dziesiętny) z danymi w tabeli (3).

Tabela 3. *Przybliżony stosunek liczby WF do liczby PPW i liczby WF do liczby PPW w odniesieniu do poziomu tekstu na skali jasnopis.pl*

STOSUNEK LICZBY PPW DO WF	STOSUNEK LICZBY WF DO PPW	POZIOM TEKSTÓW NA SKALI JASNOPIS.PL
0.832	1.208	1 – najłatwiejszy
0.866	1.155	2
0.967	1.035	3
0.961	1.043	4
1.093	0.916	5
1.089	0.922	6
1.152	0.872	7 – najtrudniejszy

Źródło: opracowanie własne

Sprawdzanie stopnia trudności tekstu przez analizę stosunku liczby WF do liczby PPW nie będzie z pewnością równie precyzyjne jak algorytmy, w oparciu, o które działa aplikacja Jasnopis. Może to być jednak pomocne, gdy tekst jest obszerniejszy, niż akceptuje to podstawowa wersja programu lub gdy precyzyjne wyliczenia nie są istotne (np. w przypadku różnic między poziomem 1 i 2).

BIBLIOGRAFIA

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

Broda B., Ogrodniczuk M., Nitoń B., Gruszczyński W., 2014, *Measuring Readibility of*

Brysbaert M., 2019, *How many words do we read per minute? A review and meta-analysis of reading rate*, https://psyarxiv.com/xynwg/ (dostęp: 29.05.2020), DOI: 10.31234/osf.io/xynwg

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., 1977–78, *Toward a Theory of Reading Comprehension and Rauding*, „Reading Research Quarterly", nr 13(1), s. 8–63.

Charzyńska E., 2015, *Text topic interest, the willingness to read and the level of reading comprehension among adults — the role of gender and education level*, „The New Educational Review", vol. 39, nr 1, s. 84–95.

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 (dostęp: 29.05.2020).

Chen X., Meurers D., 2018, *Word frequency and readability: Predicting the text-level readability with a lexical-level attribute*, „Journal Research in Reading", t. 41, nr 3, s. 486–510. DOI:10.1111/1467-9817.12121

Crossley S., Skalicky S., Dascalu M., 2019, *Moving beyond classic readability formulas: new methods and new models*, „Jorunal of Research in Reading", t. 42, nr 3–4, s. 541–561. DOI:10.1111/1467-9817.12283

Dale E., Tyler R.W., 1934, *A Study of the Factors Influencing the Difficulty of Reading Materials for Adults of Limited Reading Ability*, „The Library Quarterly: Information, Community, Policy", nr 4(3), s. 384–412.

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*, https://www.researchgate.net/publication/228965813_The_Principles_of_Readability (dostęp: 15.09.2020).

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

Grabe W., 2009, *Reading in a Second Language*, Cambridge.

Gray W., Leary S., 1935, *What makes a book readable*, Chicago.

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.

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

Na L., Nation, I.S.P. (1985). *Factors Affecting Guessing Vocabulary in Context*. "RELC Journal", nr 16(1), s. 33–42. https://doi.org/10.1177/003368828501600103

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.

Pisarek, W., 1972. *Frekwencja wyrazów w prasie: wiadomości, komentarze, reportaże*, „Biblioteka wiedzy o prasie, seria B", t. 16, Kraków.

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.

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.

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.

Sung Y.T. i in., 2015, *Leveling L2 Texts Through Readability: Combining Multilevel Linguistic Features with the CEFR*, „The Modern Language Journal", t. 99, nr 2, s. 371–379.

Wolter D., 2017, *Moving readers from struggling to proficient*, „The Phi Delta Kappan", nr 99, s. 37–39.

Yeni-Koshian G., 2005, *Percepcja mowy*, w: J. Berko Gleason, N.B. Ratner (red.), *Psycholingwistyka*, Gdańsk, s. 121–173.

Tomasz Możdzierz

THE NUMBER OF ACTUAL WORDS AND THE AVERAGE POLISH WORDS IN TEXTS AND THE RELATION TO THE READABILITY SCALE OF JASNOPIS.PL. COMPARATIVE ANALYSIS AND PRACTICAL IMPLICATIONS

Keywords: Readability, text difficulty, average word

Summary. The length of a text will be different depending on the choice of unit, if that should be the number of Actual Words (AW) or Average Words (APW, Możdzierz 2020). The ratio of AW to APW in a text changes along with this text's difficulty. An analysis of the ratio's change across seven levels of readability on jasnopis.pl (publicly accessible program to asses the difficulty of Polish texts) scale was done. Ten thousand words from authentic Polish texts were collected for each level on jasnopis.pl, and their examination allowed the capture of the ratio's change. As a result, a table of ratio proportional to different readability levels was created, which allows the easy assessment of any Polish text's difficulty, by comparison of the text's ratio to the table. Such an easily accessible assessment tool can be used in the process of Polish language education.