

A R T Y K U Ł Y I R O Z P R A W Y

Włodzimierz Gruszczyński

(Szkoła Wyższa Psychologii Społecznej, Warszawa),

Bartosz Broda

(Instytut Podstaw Informatyki PAN, Warszawa),

Bartłomiej Nitoń

(Instytut Podstaw Informatyki PAN, Warszawa),

Maciej Ogrodniczuk

(Instytut Podstaw Informatyki PAN, Warszawa)

W POSZUKIWANIU METODY AUTOMATYCZNEGO MIERZENIA ZROZUMIAŁOŚCI TEKSTÓW INFORMACYJNYCH¹

1. WPROWADZENIE

Wiele tekstów funkcjonujących w polskiej przestrzeni publicznej jest formułowanych tak, że część odbiorców ma poważne kłopoty z ich zrozumieniem. Są to zwłaszcza teksty urzędowe (w tym – prawne), teksty informacyjne (w tym – zamieszczane w prasie i mediach elektronicznych), instruktażowe (zwłaszcza instrukcje obsługi urządzeń, ulotki informacyjne dołączane do środków farmaceutycznych i kosmetyków, instrukcje wypełniania formularzy itp.). Autorzy tego typu tekstów zazwyczaj przerzucają odpowiedzialność za taki stan rzeczy na odbiorców. Przypisują im brak wystarczającego wykształcenia (np. w zakresie prawa), trudności ze skupieniem, nieznamość podstawowej terminologii itp. Wydaje się jednak, że w wielu wypadkach tym, co utrudnia odbiór, jest struktura samego tekstu – m.in. skomplikowane zdania, trudne słownictwo, zwłaszcza terminologia fachowa, metaforyka, używanie żargonu zawodowego itp.

W wielu krajach od dawna przywiązuje się wagę do tego, by teksty przeznaczone dla przeciętnego odbiorcy (czasem nawet dla każdego odbiorcy, który umie czytać) były formułowane w sposób możliwie jasny (czytelny, przejrzysty).² Dla niektórych języków, zwłaszcza dla języka angielskiego, od wielu lat istnieją metody określania (wyliczania) stopnia trudności sformułowanych w nich tekstów (por. niżej). Dla tekstów

¹ Tekst powstał w ramach projektu badawczego finansowanego przez Narodowe Centrum Nauki (2011/03/BHS2/05799), koordynowanego przez Szkołę Wyższą Psychologii Społecznej w Warszawie.

² Na przykład w Szwecji wdrożony został w ciągu ostatnich kilkunastu lat program *klarspråk*, czyli „jasny język”, dzięki któremu pisma urzędowe dawniej redagowane w tradycyjnym stylu kancelaryjnym stały się czytelne dla znakomitej większości obywateli. Píše o tym szczegółowo M. Hadryan [2012].

angielskich (i niektórych innych) proste narzędzie wyliczające parametr czytelności dołączane jest nawet do popularnego edytora tekstów.

W Polsce do niedawna niemal nikt nie interesował się tą kwestią. Co prawda, w latach sześćdziesiątych XX wieku L. Sokołowski [1967] zwrócił uwagę na wpływ struktury gramatycznej tekstu na jego zrozumiałość, a W. Pisarek [1966, 1969], wzorując się na metodach amerykańskich (por. niżej), zaproponował wzór służący do obliczania stopnia trudności tekstu polskiego i próbował uświadomić (przynajmniej dziennikarzom) wagę problemu, jednak z różnych powodów jego propozycja nie spotkała się z większym zainteresowaniem i nie dotarła do powszechnej świadomości. W latach siedemdziesiątych pojawiały się też prace zawierające wyniki badań statystycznych, mających dla omawianej kwestii zasadnicze znaczenie [m.in. Pisarek 1972; Woronczak 1976], a przede wszystkim *Listy frekwencyjne* 1974–1977. W latach późniejszych doszły do tego prace bardziej szczegółowe, które mogły stanowić podstawę głębszej analizy problemu zrozumiałości tekstu polskiego [m.in. Cygał-Krupa 1986; Imiołczyk 1987; Markowski 1990; Ruszkowski 2004]. Jednak problem (nie)zrozumiałości tekstów był w polskich pracach językoznawczych, czy ogólniej – naukowych, poruszany niezwykle rzadko, przeważnie na marginesie innych kwestii, zazwyczaj stylistycznych [por. np. Gajda 1990]. Brak było świadomości wagi problemu – w praktyce nie sprawdzano pod względem zrozumiałości (przystępności, czytelności, jasności) tekstów instrukcji, regulaminów, pism urzędowych, a nawet podręczników szkolnych. Ostatnio coraz więcej przedsiębiorstw (zwłaszcza wydawnictw prasowych), instytucji i urzędów zaczyna dostrzegać problem.³ Zaczyna zdawać sobie z niego sprawę przede wszystkim administracja państwowa, m.in. Ministerstwo Rozwoju Regionalnego, Agencja Rozwoju Przedsiębiorczości, Zakład Ubezpieczeń Społecznych, Ministerstwo Zdrowia,⁴ ale także specjaliści zajmujący się nauczaniem języka polskiego jako obcego [Seretny 2006; Banach 2011]. Wreszcie problemem zainteresował się Rzecznik Praw Obywatelskich, z którego inicjatywy odbył się I Kongres Języka Urzędowego i zainicjowana została kampania społeczna „Język urzędowy przyjazny obywatelom” [por. *Kongres* 2013]. Problemem zaczęło w znacznie większym stopniu interesować się także środowisko językoznawców. Na Uniwersytecie Wrocławskim powstała Pracownia Prostej Polszczyzny, w której językoznawcy wraz ze specjalistami z Politechniki Wrocławskiej opracowali i udostępnili aplikację automatycznie oceniającą mglistość tekstu (tzw. *fog index*, por. niżej), opracowali poradnik dla piszących o funduszach europejskich [Miodek i in. 2010] oraz

³ Dokładniej mówił o tym na I Kongresie Języka Urzędowego Tomasz Piekot [2013, 170].

⁴ Por. *Rozporządzenie Ministra Zdrowia z 26 kwietnia 2010 r. w sprawie badania czytelności ulotki*, Dz.U. nr 84 z 18 maja 2010 r., poz. 551, <http://dokumenty.rcl.gov.pl/D2010084055101.pdf>.

zbadali język raportów ewaluacyjnych [Maziarz i in. 2012]. Co ważne, zainteresowała się tym prasa codzienna i dzięki temu informacja trafiła do masowego odbiorcy [por. Banaś 2012].

W dalszym ciągu jednak bardzo niewiele osób i instytucji w Polsce ma świadomość tego, że czytelność można obiektywnie mierzyć. Wydaje się, że stworzenie i wprowadzenie nowoczesnego narzędzia, które umożliwiałoby obiektywne ustalenie tej cechy tekstu, jest niezbędne. Prace nad takim narzędziem prowadzone są obecnie w dwóch ośrodkach akademickich:⁵ w Pracowni Prostej Polszczyzny kierowanej przez Tomasza Piekota na Uniwersytecie Wrocławskim [por. m.in. Broda i in. 2010; Piekot 2013] oraz w Szkole Wyższej Psychologii Społecznej w Warszawie przez zespół utworzony przez Włodzimierza Gruszczyńskiego w celu realizacji projektu badawczego „Mierzenie stopnia zrozumiałości polskich tekstów użytkowych (pozaliterackich)”.⁶ W niniejszym tekście przedstawimy założenia tego projektu oraz wyniki uzyskane w trakcie dotychczas zrealizowanych etapów.

2. HISTORIA PROBLEMU

Próby mierzenia czytelności tekstów rozpoczęły się we wczesnym średniowieczu, od kiedy to liczono różne słowa w tekstach, aby w przybliżeniu oddać poziom trudności dzieła [por. np.: Taylor, Wahlstrom 1999]. O naukowych metodach można mówić od połowy XIX w. Wtedy przy okazji prowadzenia badań nad historią języka angielskiego zauważono zależność między długością zdania a stopniem trudności tekstu. W Rosji w tym samym czasie N. Rubakin opublikował listę 1500 słów, które powinny być znane większości Rosjan, zestawioną na podstawie badań codziennych tekstów pisanych przez przeciętnych Rosjan [Choldin 1979].

Współczesne badania nad czytelnością tekstów zapoczątkowali amerykańscy naukowcy w latach dwudziestych XX wieku. Zauważono wówczas wartość list frekwencyjnych. W ciągu kilkudziesięciu lat powstały dziesiątki wzorów, za pomocą których wyliczano stopień zrozumiałości – w większości przeznaczone one były do badania tekstów pisanych w amerykańskiej odmianie języka angielskiego.

⁵ Można tu też wspomnieć o narzędziach stworzonych przez spółkę Utilitia (przedsiębiorstwo społeczne) do badania dostępności stron internetowych (www.utilitia.pl). Jednym z testów, które są stosowane do oceny stron internetowych za pomocą tych narzędzi, jest sprawdzanie zrozumiałości tekstu, ale niestety z informacji podanych na stronie spółki nie wynika, jakie metody zastosowano w tym teście.

⁶ W skład zespołu wchodzi: Bartosz Broda (IPI PAN), Edyta Charzyńska (SWPS), Łukasz Dębowski (IPI PAN), Milena Hadryan (UAM), Bartłomiej Nitoń (IPI PAN), Grażyna Oblas (UW), Maciej Ogrodniczuk (IPI PAN), Laura Polkowska (UKSW).

Spośród wielu sposobów obliczania czytelności tekstów angielskojęzycznych, największe uznanie zdobyło zaledwie kilka wzorów, wśród nich: formuła Flescha wraz z późniejszą modyfikacją znana pod nazwą formuły Flescha–Kincaida, wzór Dalea i Chall, „indeks mgły” Gunninga, graf czytelności Frya, formuła Bormutha. O ich wartości i uznaniu, które zdobyły sobie w Stanach Zjednoczonych, może świadczyć chociażby to, że niektóre z tych wzorów zostały zastosowane w amerykańskiej wersji powszechnie znanego edytora tekstów **Microsoft Word**, a wzór Flescha–Kincaida stosowany jest przez niektóre urzędy administracji centralnej Stanów Zjednoczonych do oceny czytelności tekstów tworzonych w tych urzędach.

R. Flesch [1948] opracował jeden z pierwszych szeroko uznanych wzorów zrozumiałości. Wzór opierał się na średniej liczbie sylab w wyrazie i średniej liczbie wyrazów w zdaniu. Wzór Flescha ma następującą postać:

$$(1) \quad T = 206,835 - 1,015 \times T_w - 84,6 \times T_s$$

gdzie: T – trudność tekstu, T_w – wskaźnik trudności składni, czyli średnia liczba wyrazów w zdaniu, a T_s – wskaźnik trudności słownictwa, czyli średnia liczba sylab w słowie.

W roku 1975 powyższy wzór został przeformułowany przez J.P. Kincaida tak, by wartość wyliczonego wskaźnika odpowiadała liczbie lat, które trzeba spędzić w amerykańskiej szkole, by móc zrozumieć dany tekst. Wzór Flescha–Kincaida ma postać:

$$(2) \quad T = 0,39 \times T_w - 11,8 \times T_s - 15,59$$

Na podobnych podstawach utworzony jest także tzw. indeks mglistości, czyli *fog index* R. Gunninga [Miles 1990]:⁷

$$(3) \quad T = 0,4 \times (T_w + T_s)$$

gdzie zmienna T_w oznacza również średnią liczbę wyrazów w zdaniu, ale T_s – procent wyrazów trój sylabowych lub dłuższych.

Co istotne, wyniki otrzymywane zarówno za pomocą wzoru Flescha–Kincaida, jak i wzoru Gunninga zostały pozytywnie zweryfikowane za pomocą empirycznych badań psycholingwistycznych (*cloze procedure*, por. niżej). Współczynnik korelacji wyników obu badań wynosił 0,91 [DuBay 2006].

⁷ Wzór ten został w ostatnich latach dostosowany do tekstów pisanych po polsku i zastosowany w informatycznym narzędziu opracowanym i udostępnionym w Internecie przez Pracownię Prostej Języka na Uniwersytecie Wrocławskim [por. Broda i in. 2010; Piekot 2013].

W Polsce te same cechy tekstu wziął pod uwagę W. Pisarek [1969]. Zaproponowany przez niego wzór ma postać:

$$(4) \quad T = \frac{\sqrt{T_w^2 + T_s^2}}{2}$$

gdzie: T – stopień trudności tekstu, T_s – przeciętna długość zdania mierzona w wyrazach, a T_w – procent wyrazów czterosylabowych lub dłuższych.

Wzór Pisarka – w przeciwieństwie do wzorów Flescha–Kincaida i Gunninga – nie był nigdy weryfikowany na dużym korpusie tekstów za pomocą metod psycholingwistycznych.

Mimo stosunkowo dużej popularności wzorów takich jak Flesza–Kincaida czy Gunninga w ostatnich latach zwrócono uwagę, że formuły te odnoszą się tylko do struktury powierzchniowej tekstu, a problemy z rozumieniem mogą wynikać z innych jego cech. Zwrócili na to uwagę kognitywiści, a później językoznawcy związani z lingwistyką tekstu. Ci ostatni odwołali się do pojęć kohezji i koherencji tekstu [Graeser, Louwerse, McNamara 2002–2005]. Kohezja oznacza związki między składnikami tekstu: słowami, frazami, zdaniami. Koherencja natomiast wynika z oddziaływania między kohezją a odbiorcą. Określony poziom kohezji może wywołać inny poziom koherencji u jednego odbiorcy, inny u drugiego. Koherencja oznacza połączenia tworzone w wyniku oddziaływania elementów językowych na modele kognitywne odbiorcy. Innymi słowy kohezja i koherencja określają, w jakim stopniu tekst jest spójny syntaktycznie i tematycznie dla potencjalnych odbiorców. Kohezja i koherencja tekstu mają, zdaniem twórców teorii, znamienity wpływ na czytelność tekstu.⁸ Im większa jest kohezja i koherencja tekstu, tym tekst jest bardziej spójny, a przez to łatwiejszy w odbiorze. W Stanach Zjednoczonych prowadzi się prace nad zaawansowanym systemem komputerowym, służącym do określania kohezji i koherencji tekstów angielskojęzycznych, a tym samym do oznaczania ich czytelności. Program o nazwie **Coh-Metrix**⁹ jest interaktywnym systemem, na który składają się moduły obliczeniowe, syntaktyczne i leksykony. Aplikacja ma za zadanie analizować spójność danego tekstu na podstawie znaczników kohezji, a następnie dać odpowiednie miary czytelności tekstu dla różnych poziomów koherencji. Co więcej, intencją autorów jest również to, by program potrafił w taki sposób zanalizować strukturę syntaktyczną zdania, aby można było automatycznie wypełnić luki w kohezji danego tekstu i tym samym zwiększyć jego czytelność.

Poza metodami analitycznymi stosowane też bywają metody empiryczne, odwołujące się do doświadczeń psycholingwistyki. Prekursorem

⁸ Na gruncie polskim zwróciła na to uwagę M. Gaszyńska-Magiera [2010].

⁹ Por.: <http://www.cohmetrix.com>.

tego rodzaju metod był amerykański dziennikarz Wilson Taylor [1953]. Zaproponowany przez niego sposób badania zrozumiałości tekstu, nazywany *cloze procedure*, a odwołujący się do psychologii postaci (czyli naturalnej zdolności i potrzeby uzupełniania braków tak, by uzyskać poprawną całość), polega na tym, że w badanym tekście wykreśla się co piąty wyraz i zastępuje go pustym (linią zawsze tej samej długości). Respondentom poleca się uzupełnić brakujące wyrazy. Procent poprawnie uzupełnionych wyrazów traktowany jest jako wskaźnik czytelności badanego tekstu.¹⁰

„Test brakujących wyrazów” z samej swej istoty uwzględnia nie tylko wszystkie znane czynniki, ale także te jeszcze nieodkryte i niezbadane. Co więcej, nie pomija czynników pozajęzykowych, takich jak znajomość problematyki poruszanej w danym tekście czy zainteresowanie odbiorcy daną tematyką. Bierze też pod uwagę pewne czynniki semantyczne, takie jak: nonsensowne połączenia wyrazowe, dziwne struktury składniowe czy użycie zaimków wskazujących w funkcji anaforycznej.

Oceny stopnia trudności tekstu można dokonywać także innymi sposobami. Po pierwsze, za pomocą testów sprawdzających rozumienie przeczytanego tekstu (zazwyczaj jest to test pytań otwartych, ale może być także test wyboru). Po drugie, za pomocą mniej lub bardziej subiektywnych osądów dokonywanych przez przeciętnych czytelników lub przez ekspertów.

3. JASNOPIS – INFORMATYCZNE NARZĘDZIE DO USTALANIA STOPNIA TRUDNOŚCI TEKSTU

Od dwóch lat realizowany jest wspomniany na wstępie projekt badawczy „Mierzenie stopnia zrozumiałości polskich tekstów użytkowych (pozaliterackich)”. Projekt ma charakter interdyscyplinarny (automatyczne przetwarzanie języka naturalnego, statystyka, psychologia i oczywiście językoznawstwo). Jednym z najistotniejszych jego celów jest stworzenie narzędzia informatycznego służącego do mierzenia stopnia zrozumiałości użytkowego tekstu pisanego w języku polskim (ang. *readability*).¹¹ Narzędzie to, nazwane przez nas ***jasnopisem***,¹² w znacznym stopniu jest już gotowe. Ma postać programu komputerowego umożliwiającego obliczenie wskaźników czytelności danego tekstu i ustalanie tych cech tekstu, które

¹⁰ Dokładniej metodę tę i jej podstawy teoretyczne charakteryzuje W. Pisarek [1969, 36–38 lub 2007, 247–249].

¹¹ Wymiennie stosujemy określenia *zrozumiałość* i *czytelność*, a także na zasadzie skrótu myślowego – *trudność*. Termin *czytelność* jest niezbyt dogodny, ponieważ odnosi się on w swym pierwszym dosłownym znaczeniu do cech fizycznych tekstu pisanego, a dopiero w drugim, przenośnym – do jasności, zrozumiałości tekstu (dowolnego).

¹² Prototypowa wersja aplikacji dostępna jest pod adresem: <http://jasnopis.pl/>.

odpowiadają za zbyt duży stopień niezrozumiałości. Teksty przeznaczone do oceny mogą być albo wprowadzane („wklejane”) bezpośrednio, albo poprzez wskazanie drogi dojścia do pliku tekstowego znajdującego się na nośniku pamięci, albo przez podanie adresu URL.

Docelowo wskaźniki obliczane przez program będą umieszczane na skali wypracowanej w ramach projektu. Przy ustalaniu skali bierzemy pod uwagę wyniki prowadzonych przez nasz zespół szeroko zakrojonych badań ankietowych.¹³ Badania zostały przeprowadzone dwukrotnie. W pierwszym badaniu próba wynosiła 1306 respondentów, w drugim – ponad 1700. Wyniki pierwszego etapu badań zostały przedstawione przez E. Charzyńską i Ł. Dębowskiego [2014]. Wyniki drugiego etapu badań są obecnie analizowane. W ankietach zastosowano przede wszystkim testy *cloze procedure* i testy pytań otwartych. Pytaliśmy także o subiektywną ocenę stopnia zrozumienia tekstu po jego przeczytaniu.

W pierwszym etapie realizacji projektu stworzone zostało narzędzie umożliwiające automatyczne obliczanie stopnia trudności tekstów za pomocą metod analitycznych znanych z literatury. Uwzględniliśmy wzór Gunninga, za pomocą którego oblicza się tzw. indeks mglistości (ang. *fog index*), oraz wzór Pisarka. Obydwa zastosowano w dwóch wersjach: w pierwszej jako „wyrazy trudne” traktowano czterosylabowe lub dłuższe po sprowadzeniu ich do formy bazowej (hasłowej), w drugiej – brano pod uwagę długość formy faktycznie występującej w tekstach (roboczo nazywanej „ortograficzną”). Prócz tego indeks Pisarka obliczano za pomocą dwóch wzorów. Pierwszy z nich, przywołany powyżej, opisuje funkcję nieliniową. Drugi wzór, opisujący funkcję liniową, zrekonstruowany został na potrzeby projektu na podstawie wykresu zamieszczonego w kilku tekstach W. Pi-sarka [1969, 47 lub 2007, 259; 2002, 200] przez Dębowskiego [2012]:

$$(5) \quad T = 1/3 T_s + 1/3 T_w + 1$$

Ogółem więc obliczane są dwa wskaźniki w sześciu wariantach:

- dwa warianty *fog index* wg wzoru (3): z uwzględnieniem form podstawowych i z uwzględnieniem form tekstowych wyrazów „trudnych”;
- cztery warianty indeksu Pisarka: według wzorów (4) i (5), a każdy z nich z uwzględnieniem form podstawowych i z uwzględnieniem form tekstowych wyrazów „trudnych”.

W wariantach, w których konieczne jest sprowadzanie form tekstowych do form podstawowych, czyli tzw. lematyzacja, zastosowany został do tego celu analizator morfologiczny **Morfeusz**, wykorzystujący dane SGJP.¹⁴ Ponieważ **Morfeusz** jest analizatorem bezkontekstowym, zastosowany został także tager WCRFT, który w wypadku alternatywnych

¹³ Badania te koordynuje Edyta Charzyńska. Ich metodologia i szczegółowe wyniki przedstawione zostaną w innym miejscu.

¹⁴ Por.: <http://sgjp.pl/morfeusz/index.html>.

interpretacji uzyskanych za pomocą **Morfeusza** wybiera na podstawie kontekstu najprawdopodobniejszą z nich [Radziszewski 2013].

Przed przystąpieniem do obliczania wskaźników trudności tekstu wykonywane są więc następujące operacje:

- 1) Nadesłany plik, w zależności od formatu, jest odpowiednio konwertowany do formatu akceptowanego przez wykorzystywany dalej tager. Jeśli format pliku jest jednym z obsługiwanych przez program **Open Office**, poza podziałem na akapity w tekście zostają oznaczone również tytuły i wypunktowania (pod warunkiem, że zostały one odpowiednio wyszczególnione w dokumencie wejściowym). Tytuły i wypunktowania będą ignorowane przy zliczaniu statystyk tekstu.
- 2) Odpowiednio przygotowany plik jest znakowany przez tager WCRFT. Wykrywa on końce zdań oraz nadaje poszczególnym słowom tagi zgodne z opisami na stronie Narodowego Korpusu Języka Polskiego [<http://nkjp.pl/poliqarp/help/pl.html>].
- 3) Oznakowana wersja pliku jest analizowana. Zliczane są kolejne zdania i akapity. Słowa oznaczone tagami: *fin*, *praet*, *bedzie*, *inf*, *imps*, *impt*, *winiem*, *pred* zliczane są jako czasowniki. Słowa oznaczone tagami *adj*, *pact*, *ppas* zliczane są jako przymiotniki. Słowa oznaczane jako *subst* i *ger* zliczane są jako rzeczowniki. Znaki interpunkcyjne (tag *interp*) są ignorowane.
- 4) W zależności od wybranego algorytmu przy klasyfikowaniu wyrazów jako trudnych brana jest pod uwagę ich forma podstawowa (BASE) lub tekstowa (ORTH). Za słowa trudne uznawane są te, które odpowiednio w formie podstawowej lub tekstowej mają cztery sylaby lub więcej. Jeśli wyraz uznany za trudny należy do jednej z wymienionych w punkcie trzecim części mowy (nie licząc elementów ignorowanych), jest on odpowiednio zliczany jako trudny czasownik, trudny przymiotnik lub trudny rzeczownik.
- 5) Jeśli wybrano indeks Pisarka, ignorowane ponadto są liczebniki (tagi: *num*, *numcol*) oraz liczby zapisane za pomocą cyfr. Ponadto dla indeksu Pisarka ignorowane są nazwy własne, przyjęto tutaj prosty algorytm pomijający wszystkie słowa rozpoczynające się dużą literą, nieznajdujące się na początku zdania.
- 6) Ostatecznie na podstawie zliczonych wartości ilościowych wyliczane są statystyki tekstu:
 - średnia długość zdania – jako iloraz liczby wszystkich znalezionych w tekście słów (bez interpunkcji) i liczby wykrytych zdań;
 - średnia długość akapitu – jako iloraz liczby wszystkich słów w tekście i liczby akapitów;
 - procent słów trudnych – jako iloraz liczby słów uznanych w tekście za trudne (4 lub więcej sylab) i całkowitej liczby słów w tekście razy 100%;
 - procent rzeczowników – jako iloraz liczby wykrytych w tekście rzeczowników i liczby wszystkich słów razy 100%;

- procent rzeczowników trudnych – jako iloraz liczby rzeczowników uznanych za trudne i liczby wszystkich słów w tekście razy 100%;
- procent czasowników – jako iloraz liczby wykrytych w tekście czasowników i liczby wszystkich słów razy 100%;
- procent czasowników trudnych – jako iloraz liczby czasowników uznanych za trudne i całkowitej liczby słów znalezionych w tekście razy 100%;
- procent przymiotników – jako iloraz liczby wykrytych w tekście przymiotników i wszystkich słów razy 100%;
- procent przymiotników trudnych – jako stosunek przymiotników uznanych za trudne do całkowitej liczby słów razy 100%;
- stosunek rzeczowników do czasowników liczony jest jako iloraz liczby wszystkich znalezionych w tekście wystąpień rzeczowników i liczby wszystkich znalezionych w tekście wystąpień czasowników.

W drugim etapie prac do **Jasnopisu** wprowadzony został kolejnym wskaźnik. Jest to automatyczny test Taylora, który bez wątpienia stanowi istotną innowację metodologiczną. Ten moduł aplikacji, uzupełnia co *n*-te słowo w badanym tekście na podstawie „treningu” na korpusach referencyjnych.¹⁵ Program „odgaduje” pominięte słowa na podstawie prawdopodobieństwa wystąpienia słowa w tzw. bigramie, czyli dwuelementowej sekwencji słów [dokładniejsze omówienie algorytmu – por. Broda i in. 2014, 576]. Automatyczny test Taylora ma również dwa warianty. W pierwszym wyniki generowane są z wykorzystaniem miary „zakłopotania” (ang. *perplexity*), drugi jest dokładnym odtworzeniem testu Taylora (narzędzie usuwa co piąty wyraz w tekście, następnie stara się te luki prawidłowo uzupełnić). W obydwu wariantach sprawdza się, do którego ze zdefiniowanych przez nas typów stylistyczno-gatunkowych (literatura dla dzieci,¹⁶ teksty popularnonaukowe, teksty prasowe, teksty ustaw, teksty artykułów w polskiej Wikipedii) dostarczony tekst jest najbardziej podobny.

¹⁵ W obecnej chwili wielkość i wartość korpusów jest jednym z największych kłopotów zespołu. Brakuje nam odpowiednio dużej liczby tekstów, które byłyby podzielone na klasy trudności. Uwidacznia się tu problem metodologiczny: aby wytrenować projektowany system tak, by oceniał tekst pod względem trudności, musimy dysponować dużą liczbą tekstów podzielonych na klasy trudności. Podziału takiego można dokonać albo na podstawie intuicji, albo na podstawie wcześniej istniejących metod obliczania wskaźnika trudności tekstu, albo na podstawie wyników badań ankietowych. Bez wątpienia najbardziej wiarygodną i poprawną metodologicznie podstawą są właśnie wyniki badań ankietowych. Jednak korpus tekstów, który można dzięki takim badaniom uzyskać, jest zdecydowanie za mały.

¹⁶ Oczywiście nie traktujemy literatury dla dzieci jako „tekstów pozaliterackich”. Odwołujemy się do nich, ponieważ trudno znaleźć teksty nieliterackie o takim samym stopniu trudności jak literatura dla dzieci, czyli zrozumiałe niemal dla wszystkich odbiorców.

Do **Jasnopisu** włączony został także moduł wyliczający stopień podobieństwa tekstu poddanego badaniu do tekstów w korpusie referencyjnym. Zakładamy, że jeśli jakiś tekst wykazuje dużo podobieństw (na różnych poziomach organizacji) do tekstów w jakimś korpusie referencyjnym określonego typu, to również poziom trudności tekstu badanego będzie podobny do poziomu trudności tekstów z tego korpusu. Na przykład, jeśli okaże się, że badany tekst wykazuje duże podobieństwa do tekstów ustaw, to można zasadnie przypuszczać, że również stopień trudności tego tekstu będzie podobny do stopnia trudności tekstów ustaw. W module ustalającym stopień podobieństwa zastosowano model wektorowy (ang. *vector space model*) [por. Salton i in. 1975]. Dokładniejsza charakterystyka tego modułu przedstawiona jest w tekście B. Brody i in. [2014].

W trakcie dotychczasowych prac nad systemem **Jasnopis** przeprowadzono szereg eksperymentów [Broda i in. 2014, 578–579]. Polegały one między innymi na sprawdzeniu stopnia zgodności wyników otrzymywanych za pomocą różnych metod. Sprawdzone korelacje zachodzące między wynikami uzyskiwanymi za pomocą różnych wariantów tego samego wzoru (np. *fog index*, wzory Pisarka) oraz wynikami uzyskiwanymi za pomocą różnych wzorów. Badania przeprowadzono na milionowym podkorpusie Narodowego Korpusu Języka Polskiego. Korelacje okazały się bardzo wysokie. Największy stopień korelacji (0,99) wykazują wyniki uzyskiwane za pomocą dwóch wzorów Pisarka – nieliniowego i liniowego, jeśli przy obliczeniach prowadzonych za ich pomocą jako „wyrazy trudne” traktuje się czterosylabowe lub dłuższe formy podstawowe wyrazów występujących w tekście. Najniższy stopień (choć też bardzo wysoki – 0,94) wykazują wyniki uzyskiwane za pomocą dwóch wariantów formuły *fog index*.¹⁷

Próbowaliśmy też sprawdzić, na ile obecność wymienianych w literaturze przedmiotu cech tekstu wpływających na stopień jego trudności oddziałuje na wyniki uzyskiwane za pomocą poszczególnych wzorów. W tym celu w piętnastu próbkach różnych tekstów oznakowane zostały te elementy leksykalne, których wpływ na stopień trudności wydaje się niezaprzeczalny, m.in.: rzeczowniki abstrakcyjne, terminy naukowe, wyrazy należące do stylu urzędowego, archaizmy, wyrazy synchronicznie obce, wyrazy o frekwencji tekstowej niższej niż ich synonimy, peryfrazy. Niektóre z tych typów jednostek leksykalnych występowały w naszych tekstach na tyle rzadko (np. archaizmy), że zbadanie ich wpływu na zrozumiałość okazało się niemożliwe. Liczebność pozostałych silnie korelowała zarówno z wynikami wskaźników obliczanych za pomocą stosowanych przez nas wzorów, jak i z wynikami badań psycholingwistycznych.

Jasnopis udostępniany będzie *online* pod adresem <http://jasnopis.pl> (obecnie jest tam dostępna wersja testowa, ciągle udoskonalana i uzu-

¹⁷ Szczegółowe wyniki eksperymentów przedstawione zostały w tekście B. Brody i in. [2014, 578].

pełniana o nowe moduły i funkcje). Z technicznego punktu widzenia udostępnienie działającego online systemu do sprawdzania trudności tekstów i proponowania w nich poprawek już jest wielkim ułatwieniem dla autorów i redaktorów. Wydaje się jednak, że narzędzie dostępne w takiej formie nie jest optymalne, jeśli chce się wprowadzać sugerowane poprawki na bieżąco. Kopiowanie fragmentów tworzonego właśnie dokumentu do przeglądarki w celu sprawdzenia ich zrozumiałości nie wydaje się najwygodniejszym rozwiązaniem. Tym samym istotną częścią projektu stało się stworzenie wtyczek (ang. *plug-in*) do najbardziej popularnych edytorów tekstowych **Microsoft Word** i **Open Office**.

Dzięki zastosowaniu wtyczek autorzy nie będą musieli znacząco modyfikować swojego sposobu pracy, ponieważ nie trzeba będzie kopiować pisanego tekstu do okna przeglądarki. Dodatkowo, ze względu na prostotę wykorzystania, wtyczka pozwala zintegrować mierzenie zrozumiałości z procesem pisania, a nie traktować je jako dodatkowy etap, który jest wykonywany już po napisaniu tekstu.

Jak już powiedzieliśmy wyżej, w ramach projektu prowadzone są na szeroką skalę badania ankietowe zrozumiałości tekstów. Pomaga to w wielu aspektach pracy nad mierzeniem zrozumiałości metodami analitycznymi. Przede wszystkim umożliwia zobiektywizowaną analizę czynników wpływających na rozumienie tekstów w języku polskim. Dodatkowo dzięki wynikom badań ankietowych jesteśmy w stanie wyskalować narzędzie i zmierzyć jakość działania automatycznych metod mierzenia zrozumiałości zaimplementowanych w programie **Jasnopis**.

4. WNIOSKI I PERSPEKTYWY

Już na obecnym etapie prac **Jasnopis** stanowi dużą pomoc dla piszących i redagujących teksty. Daje on możliwość bardzo szybkiego uzyskania względnie obiektywnej oceny stopnia trudności tekstu wyliczanej różnymi sposobami. Dodatkowo wskazywane są akapity najtrudniejsze i najłatwiejsze, dane o stopniu nominalności tekstu (stosunek liczby czasowników do rzeczowników).

Wydaje się, że osiągnięcie dokładniejszych (bliższych intuicji badaczy?) wyników za pomocą prostych wzorów nie jest możliwe. Dlatego podejmujemy próby zastosowania znacznie bardziej wyrafinowanych metod. Próbowujemy w nich uwzględnić różne cechy tekstu, m.in. długość łańcuchów koreferencyjnych (odpowiedzialnych m.in. za spójność tekstu), a także wyniki badań prowadzonych wcześniej w innych celach. Wydaje się, że istotne będzie uwzględnienie stopnia pokrycia tekstu przez wyrazy ogólnie znane, choć nie zawsze mające dużą frekwencję tekstową. Wykorzystujemy do eksperymentów listy frekwencyjne J. Imiołczyka [1987], listy wyrazów współnoodmianowych A. Markowskiego [1990], a także listy rangowe Z. Cygal-Krupy [1986]. Największe nadzieje wią-

żemy jednak ze wspomnianym powyżej wyliczaniem stopnia podobieństwa do wyodrębnionych wcześniej typów tekstów.

Bibliografia

- M. Banach, 2011, *Tekst trudny, czyli... jaki? O czynnikach wpływających na trudność tekstu*, „Polonica” 31, s. 27–35.
- J. Banaś, 2012, *Mów do mnie prosto [rozmowa z dr. Tomaszem Piekotem]*, „Gazeta Wyborcza” 24.10.2012.
- B. Broda, M. Maziarz, T. Piekot, A. Radziszewski, 2010, *Trudność tekstów o funduszach europejskich w świetle miar statystycznych*, „Rozprawy Komisji Językowej” XXXVII, s. 23–40.
- B. Broda, M. Ogrodniczuk, B. Nitoń, W. Gruszczyński, 2014, *Measuring Readability of Polish Texts: Baseline Experiments* [w:] *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC14)*, Reykjavik, s. 573–580, http://www.lrec-conf.org/proceedings/lrec2014/pdf/427_Paper.pdf (dostęp: 1.08.2014).
- E. Charzyńska, Ł. Dębowski, 2014, *Factors of Readability of Polish Texts: A Psycholinguistic Study*, referat przedstawiony na 20. edycji QUALICO 2014 (Quantitative Linguistic Conference), 29 maja–1 lipca 2014, Olomuniec.
- M.T. Choldin, N.A. Rubakin, 1979 [w:] A. Kent, H. Lancour, W.Z. Nasri i in., *Encyclopedia of library and information science* 26 (illustrated ed.), s. 178–79.
- Z. Cygal-Krupa, 1986, *Słownictwo tematyczne języka polskiego – zbiór wyrazów w układzie rangowym, alfabetycznym i tematycznym*, Kraków.
- Ł. Dębowski, 2012, *Koncepcja algorytmu wyliczającego czytelność tekstu na podstawie wskaźników stylistyczno-statystycznych*, niepublikowany raport sporządzony w ramach projektu „Mierzenie stopnia zrozumiałości polskich tekstów użytkowych (pozaliterackich)”, s. 8.
- W.H. DuBay, 2004, *The Principles of Readability*, <http://www.nald.ca/library/research/readab/cover.htm> [dostęp 10.11.2013].
- W.H. DuBay, 2006, *Smart Language: Readers, Readability, and the Grading of Text*, Costa Mesa.
- R. Flesch, 1948, *A new readability yardstick*, „Journal of Applied Psychology” 32, s. 221–233.
- S. Gajda, 1990, *Współczesna polszczyzna naukowa. Język czy żargon?*, Opole.
- M. Gaszyńska-Magiera, 2010, *Spójność tekstu a ocena stopnia jego trudności*, „Kwartalnik Polonicum” 10, s. 52–59.
- D. Gorczyca, *Tekst w Internecie*, „Zarządzanie Zmianami. Biuletyn POU” 5 (39), http://www.wsz-pou.edu.pl/biuletyn/?strona=biul_int&nr=3&p= [dostęp: 10.05.2014].
- A. Graeser, M. Louwerse, D. McNamara, 2002–2005, *Cox-Matrix: Automated Cohesion and Coherence Scores to Predict Text Readability and Facilitate Comprehension*, <ftp://129.219.222.66/pdf/IESproposal.pdf> [dostęp: 2.07.2014].
- M. Hadryan, 2012, *Demokratyzacja języka urzędowego. Współczesne tendencje polityki językowej w Szwecji i w Polsce*, rozprawa doktorska pod opieką naukową W. Maciejewskiego, Poznań, <http://hdl.handle.net/10593/2689> [dostęp: 10.05.2014].

- J. Imiołczyk, 1987, *Prawdopodobieństwo subiektywne wyrazów: podstawowy słownik frekwencyjny języka polskiego*, Warszawa–Poznań.
- Kongres, 2013, *I Kongres Języka Urzędowego, 30-31 października 2012 r.*, opr. M. Lipińska, Warszawa.
- Listy frekwencyjne, 1974–1977, Słownictwo współczesnego języka polskiego. Listy frekwencyjne*, t. 1–5, Warszawa.
- A. Markowski, 1990, *Leksyka wspólna różnym odmianom polszczyzny*, Warszawa.
- M. Maziarz, T. Piekot, M. Poprawa, B. Broda, A. Radziszewski, G. Zarzeczny, 2012, *Język raportów ewaluacyjnych*, Warszawa.
- T.H. Miles, 1990, *Critical thinking and writing for science and technology*, San Diego.
- J. Miodek, M. Maziarz, T. Piekot, M. Poprawa, G. Zarzeczny, 2010, *Jak pisać o funduszach europejskich*, Warszawa.
- T. Piekot, 2013, *Ruch prostego języka – korzyści i zagrożenia (Standard „plain language” jako język przyjazny obywatelom)* [w:] Kongres, 2013, s. 169–197.
- W. Pisarek, 1966, *Recepty na zrozumiałość wypowiedzi*, „Zeszyty Prasoznawcze” 2/3, s. 44–53.
- W. Pisarek, 1969, *Jak mierzyć zrozumiałość tekstu?*, „Zeszyty Prasoznawcze” 4, s. 35–48 [cyt. wg przedruku w: Pisarek 2007, s. 245–262].
- W. Pisarek, 1972, *Frekwencja wyrazów w prasie*, Kraków.
- W. Pisarek, 2002, *Nowa retoryka dziennikarska*, Kraków.
- W. Pisarek, 2007, *O mediach i języku*, Kraków.
- A. Przepiórkowski, M. Bańko, R.L. Górski, B. Lewandowska-Tomaszczyk (red.), 2012, *Narodowy Korpus Języka Polskiego*, Warszawa.
- A. Radziszewski, 2013, *A tiered CRF tagger for Polish* [w:] R. Bembenik, Ł. Skonieczny, H. Rybiński, M. Kryszkiewicz, M. Niezgódka (red.), *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*, Heidelberg.
- M. Ruszkowski, 2004, *Statystyka w badaniach stylistyczno-składniowych*, Kielce.
- G.M. Salton, A.K.C. Wong, C.S. Yang, 1975, *A vector space model for automatic indexing*, „Communications of the ACM” 18(11), s. 613–620.
- A. Seretny, 2006, *Wskaźnik czytelności tekstu jako pomoc przy określaniu stopnia jego trudności*, „LingVaria” 2, s. 87–98.
- L. Sokołowski, 1967, *Wpływ struktury gramatycznej tekstu na jego zrozumienie*, „Zeszyty Prasoznawcze” 4, s. 11–27.
- SGJP: Z. Saloni, M. Woliński, R. Wołosz, W. Gruszczyński, D. Skowrońska, 2012, *Słownik gramatyczny języka polskiego*, wyd. II, Warszawa.
- M.C. Taylor, M.W. Wahlstrom, 1999, *Readability as Applied to An ABE Assessment Instrument*, <http://www.nald.ca/fulltext/report4/rep31-35/rep34-01.htm> [dostęp: 12.10.2013].
- W.L. Taylor, *Cloze Procedure: A New Tool for Measuring Readability*, „Journalism Quarterly” 30, s. 415–433.
- J. Woronczak, 1976, *O statystycznym określeniu spójności tekstu* [w:] M.R. Mayenowa (red.), *Semantyka tekstu i języka*, Wrocław, s. 165–173.

***In search of a method of automatic measurement of readability
of informational texts***

Summary

Numerous texts that are incomprehensible for some of their intended recipients can be found in the Polish public space. In many cases, the problem is related to the structure of the text, which can be measured objectively. Methods of determining the degree of readability of texts have been developed for many languages. The importance of this problem has also been noticed by certain Polish institutions, publishing houses and offices in recent years. In the Polish academia, the problem is being investigated primarily by two centres: Pracownia Prostej Polszczyzny (Plain Polish Laboratory) at the University of Wrocław and the University of Social Sciences and Humanities (SWPS) in Warsaw. This paper presents the state of research on this problem within the project "Mierzenie stopnia zrozumiałości polskich tekstów użytkowych (pozaliterackich)" ("Measuring the degree of readability of nonliterary Polish texts") carried out at SWPS. The problem and the state of research thereon in Poland and abroad are described in the first place. A tool for measuring readability of Polish texts called *Jasnopis*, which is currently under construction and which uses both known methods, such as Pisarek index, as well as new ones, such as the automatic Taylor test, is presented afterwards. This paper is finished with a presentation of conclusions from the works performed so far and ideas for further research.

Adj. Monika Czarnecka