Measuring readability of Polish texts

Włodzimierz Gruszczyński[1], Bartosz Broda[2], Edyta Charzyńska[3], Łukasz Dębowski[2],
Milena Hadryan[4], Bartłomiej Nitoń[2] and Maciej Ogrodniczuk[2]

[1]University of Social Sciences and Humanities, ul. Chodakowska 19/31, 03-815, Warsaw, Poland.
[2]Institute of Computer Science, Polish Academy of Sciences, ul. Jana Kazimierza 5, 01-248 Warsaw, Poland
[3]Institute of Pedagogy, University of Silesia, ul. Grażyńskiego 53, 40-126 Katowice, Poland
[4]Institute of Linguistics, Adam Mickiewicz University, al. Niepodległości 4, 61-874 Poznań, Poland

Abstract

In this paper we present a series of new methods of measuring readability of Polish non-literary texts. Starting with a short discussion of previous approaches we attempt at identification of new factors influencing readability and propose a new formula taking into account various linguistic features of a text. We also implement two corpus-based methods of assessing readability and finally describe an application implementing the results of our research.

1. Introduction

The awareness that clarity of texts can be perceived differently by different recipients is probably as old as the invention of writing. In ancient Greece writers consulted orators to find out whether a given text is sufficiently understandable. Since 900 AD words in texts were counted to estimate their difficulty but only in 19 century linguists started to express interest in the topic, inventing various methods for calculating textual legibility. One of the first analytical formulae based on average word and sentence length in a text which met with big response has been proposed by Flesch (1948), but numerous other methods were put forward: cloze procedures (Taylor, 1953), multiple choice understandability tests (Royer et al., 1979), written free recall (Bernhardt, 1991), eye-tracking (Copeland et al., 2014) etc.

In Poland the forerunner of analytical methods was Pisarek (1966), proposing his own formula for calculating readability of Polish texts:

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

where T is text difficulty level, T_s is percentage of 4-or-more-syllable words (in lemmatized form) and T_w is average sentence length (in words).

In 1970s statistical methods of analysis of Polish texts were introduced based on frequency lists (Pisarek, 1972; Woronczak, 1976). Since then, several more detailed work on readability in Polish were published (Cygal-Krupa, 1986; Imiołczyk, 1987; Markowski, 1990; Ruszkowski, 2004), but the problem was frequently treated as a marginal issue in research on stylistics and normalization (Gajda, 1990; Gizbert-Studnicki, 1986; Wojtak, 1993; Malinowska, 1999) or language teaching (Seretny, 2006; Banach, 2011). Only recently the need of maintaining readability of official communication was emphasized by the Council for the Polish Language Rada Języka Polskiego (2012) and the research topic was taken up by two academic centres: University of Wrocław's Plain Language Studio (Pol. Pracownia Prostej Polszczyzny, `http://www.ppp.uni.wroc.pl/`) and University of Social Sciences and Humanities, implementing Jasnopis (`http://www.jasnopis.pl/`).

Measuring readability poses a series of questions: What does it mean 'to understand a text'? How to verify that the reader comprehended what has been written? Can a given text be interpreted differently by different readers? The rift about language properties influencing understandability is widening while at the same time most readers can easily, based on intuition and probably also language experience, estimate the difficulty of the text being investigated. However, much fewer people can reasonably justify their opinion.

Apart from certain properties of text related to its legibility (based on typographical or stylistic features, use of punctuation etc.) rather than readability (ability to understand it), many linguistic features can be investigated and are regarded as correlating with text (un)clarity. As for lexical features they are e.g.: presence of abstract nouns, use of professional jargon, foreign and archaic words, low-frequency words, abbreviations, phraseological or periphrastic expressions. Among syntactic features we can distinguish length and complexity of sentences, presence of long sequences of nouns in genitive, use of passive voice, impersonal constructions, negation, inversion etc.

Our original methodology (created particularly for measuring readability of Polish texts and not adapted from formulae defined for other languages) is analytical, i.e. assumes computation of readability score based on selected features of the text being analysed. We have applied experimental methods to calculate analytical factors (such as average sentence length or percentage of 'difficult' words) in order to maintain optimal readability with respect to perception of text. Since one of our aims was also application of psycholinguistic methods in the process of measuring readability, we have conducted a series of cloze tests, open question tests and expert evaluation surveys. The resulting formula, created in the process, takes into account textual frequency of lexical units, certain morphosyntactic structures and so-called subjective frequency of lexemes (Imiołczyk, 1987).

2. A questionnaire survey

To verify theoretical premises offered by existing analytical methods, identify new factors influencing readability and investigate potential correlation of responses on understandability with education of respondents we have conducted two surveys presented below.

Survey 1 was carried out among 1309 respondents who were asked to estimate readability of 3 texts randomly selected from a set of 15 texts. The samples were 300-word-long each and they featured varied predicted readability (textbooks, scientific texts, acts of Polish Parliament, official letters, manuals, newspaper articles). Each person was asked to respond to: (a) 4 single choice questions with 4 alternative answers, (b) cloze test with 50 gaps (created by removing each 5th word, starting from the second sentence), (c) 5 open questions, (d) subjective assessment of their understanding of the text on a scale of 1 (little understanding) to 5 (complete understanding). The profile of the person (sex, age, education, address, acquired and actual profession etc.) was recorded for further investigation.

The results showed that a readability index calculated with Pisarek's equation correlates with understandability indicated by respondents, yet the correlation is lower than expected and component variables of the equation (i.e., average sentence length and percentage of 'difficult' words) correlate with just a few of the survey tests. It could have been influenced by the low number of texts being investigated, so we decided to conduct a second survey over a larger number of samples.

Survey 2 was carried out among 1759 respondents. Before the texts were selected, a 7-point scale was proposed to represent difficult readability levels corresponding to Gunning (1952) FOG-like blocks of education in Polish schools necessary to understand a text of a given class:

1. primary school, grades 1–3 (age 6–9)
2. primary school, grades 4–6 (age 9–12)
3. secondary school[1] (age 12–15)
4. high school[2] (age 15–18)
5. undergraduate studies[3] (age 18–21)
6. graduate studies[4] (age 21–23)
7. postgraduate studies or expert knowledge expected.

The initial set of survey texts was compiled from proposals of the team members consisting of 300-word-long texts representing (in their opinion) each readability class. This method resulted in selecting about 30 texts per class, 200 texts in total. In the next step the texts in each class were ordered by their FOG readability and 5 texts with average difficulty within each class were selected to be used in the survey.

Each respondent received 2 texts to evaluate and one of the two understandability tests: 50-gap cloze or 5 open questions. Apart from that they were asked to assess their interest in the topic of the text and willingness to complete the task on a scale of 1 (completely uninterested/unwilling) to 7 (excited/enthusiastic). For texts in classes 4–7 a question about knowledge in the topic described in the text was additionally asked. After the test was completed a question about subjective assessment of respondent's understanding of the text on a scale of 1 (little understanding) to 7 (complete understanding) was asked.

The results of the second survey confirmed correctness of Pisarek's equation and showed strong correlation between text difficulty measured analytically and based both on test answers and subjective evaluation. This also confirms correspondence of results obtained for Polish and English which may indicate language-independence of such readability predictors as lexical/semantic and syntactic features of a text. What is more, the survey confirmed validity of former theoretical deliberations on influence of certain linguistic variables on text readability. Use of short sentences and words, prevalence of verbs over nouns and reduction of foreign words, gerunds and genitive chains used in the text can greatly improve its comprehension.

3. New analytical readability formula

In this section we would like to propose a new analytical formula based on results of our linguistic and psycholinguistic research. Taking into account the values of the following variables:

- CLASS — readability class set by experts
- CLOZE — average points from cloze test
- OPEN — average points from open question test

obtained for each of our sampled 35 texts (5 texts for each class in survey 2, see previous section), we can try to create formulae predicting their values based on surface features of a given text. To achieve that, values of 17 linguistic variables (presented in Table 1) expected to be correlated with text readability were computed.

Due to limited data available (35 texts) for which a relatively high number of variables was calculated, we restricted our procedure to construction of linear formula:

$$Y_i \approx A_0 + \sum_{j=1}^{K} A_j X_{ji},$$

where Y_i is a selected response variable for i-th text, X_{ji} is a j-th linguistic variable for i-th text, A_j are certain constants, $i = 1, 2, ..., N$, $N = 35$ is number of texts, and $K = 17$ is the number of linguistic (dependent) variables.

To calculate A_j we minimized the following expression:

$$\sum_{i=1}^{N} \left[Y_i - A_0 - \sum_{j=1}^{K} A_j X_{ji} \right]^2 + \lambda \sum_{j=1}^{K} |A_j|^{\alpha}$$

where λ and α depended on the method investigated:

1. BASELINE: predicted readability is independent on text and equals to the average result from all samples, i.e. $\lambda = 0$ and $A_j = 0$ for $j = 1, 2, .., K$.
2. PISAREK: predicted readability depends on ASL and PHW only (as in Pisarek's formula); A_j coefficients are calculated using the method of least squares, i.e. $\lambda = 0$ and $A_j = 0$ for $j = 3, 4, .., K$.

[1] Pol. gimnazjum.

[2] Pol. liceum.

[3] Pol. studia licencjackie.

[4] Pol. studia magisterskie.

Table 1: Linguistic variables used in the new analytic formula

Variable name	**Explanation**
ASL	average sentence length (in words)
PHW	4-or-more-syllable words to all words ratio (in %)
PSUBST	nouns to all words ratio (in %)
PHSUBST	4-or-more-syllable nouns to all words ratio (in %)
PVERB	verbs to all words ratio (in %)
PHVERB	4-or-more-syllable verbs to all words ratio (in %)
PADJ	adjectives to all words ratio (in %)
PHADJ	4-or-more-syllable adjectives to all words ratio (in %)
SUBSTVERB	noun to verb ratio
APL	average paragraph length (in words)
AWL	average word length (in syllables)
AGEN	average genitive chain length (in words)
PIMPERS	impersonal verbs to all words ratio (in %)
PGER	gerunds to all words ratio (in %)
POSC	nouns ending with *-ość* to all words ratio (in %)
PHWIMIOL	4-or-more-syllable words from Imiołczyk's list to all words ratio (in %)
PGEN	genitives to all words ratio (in %)

3. NMK: least squares regression, takes into account all dependent variables; A_j coefficients are calculated using the method of least squares, i.e. $\lambda = 0$.
4. LASSO; takes into account all dependent variables; A_j coefficients are calculated using Lasso regression (Tibshirani, 1996), i.e. $\alpha = 1$ and λ is calculated in the process of cross-validation.
5. RIDGE: takes into account all dependent variables; A_j coefficients are calculated using ridge regression (Tikhonov and Arsenin, 1977), i.e. $\alpha = 2$ and λ is calculated in the process of cross-validation.

To select the best predictor we calculated:

1. mean absolute error:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |Y_i - F_i| \, .$$

where Y_i is the value of CLASS variable for i-th text and

$$F_i = A_0 + \sum_{j=1}^{K} A_j X_{ji}$$

is the value of selected formula for i-th text tested on all texts

2. mean absolute error of the cross-validated formula:

$$\text{MAE}_{CV} = \frac{1}{N}\sum_{i=1}^{N} |Y_i - F_i'| \, .$$

where

$$F_i' = A_0' + \sum_{j=1}^{K} A_j' X_{ji}$$

is the value of selected formula for i-th text tested on all texts except the i-th text

3. Pearson correlation coefficient between CLASS variable and the formula, $\rho = \text{corr}(Y, F)$,

Table 2: CLASS prediction evaluation

	MAE	MAE$_{CV}$	ρ	ρ_{CV}
BASELINE	1.710	1.760	NA	-1.00
PISAREK	0.580	0.641	0.93	0.91
NMK	0.235	0.521	0.99	0.94
LASSO	0.431	0.506	0.98	0.96
RIDGE	0.445	0.500	0.97	0.96

4. Pearson correlation coefficient between CLASS variable and the cross-validated formula, $\rho_{CV} = \text{corr}(Y, F')$.

The method with the lowest MAE$_{CV}$ turned out to be RIDGE, with cross-validation error amounting to half grade in 7-grade scale. However, this method resulted in an overly complicated readability formula:

$$\begin{aligned}\text{CLASS} \approx{}& -1.479 + 0.02708 \times \text{ASL} + 0.02909 \times \text{PHW}\\ &+ 0.0248 \times \text{PSUBST} + 0.04793 \times \text{PHSUBST}\\ &- 0.03267 \times \text{PVERB} + 0.04752 \times \text{PHVERB}\\ &+ 0.03114 \times \text{PADJ} + 0.06377 \times \text{PHADJ}\\ &+ 0.1585 \times \text{SUBSTVERB} + 0.002089 \times \text{APL}\\ &+ 0.9057 \times \text{AWL} + 0.09932 \times \text{AGEN}\\ &- 0.08596 \times \text{PIMPERS} + 0.0299 \times \text{PGER}\\ &+ 0.1938 \times \text{POSC} + 0.03792 \times \text{PHWIMIOL}\\ &+ 0.02781 \times \text{PGEN}.\end{aligned}$$

so if we are interested in a simpler variant, we can use the PISAREK method-based formula, just slightly more complex than original Pisarek's equation:

$$\text{CLASS} \approx -0.01413 + 0.0857 \times \text{ASL} + 0.2949 \times \text{PHW}.$$

4. Corpus-based readability assessment

Another method implemented in our study is based on reference corpora, each representing different readability level. Similarity between input text and a corpus measured directly or by using a corpus-based language model in a gap-filling task can be used to project the readability of a corpus to readability of the text being investigated.

In our evaluation we used 5 corpora representing documents of different expected readability levels, with their average FOG and CLASS values presented in Table 3:

1. corpus of children's literature consisting of 177 texts (186 149 words)
2. Wikipedia corpus consisting of 180 texts (183 093 words) selected from the Polish Wikipedia Corpus
3. press article corpus consisting of 180 articles (171 538 words) selected from Rzeczpospolita Corpus (Presspublica, 2013)
4. legal act corpus consisting of 175 legal acts of the Polish Sejm (172 627 words)
5. popular science corpus consisting of 183 texts (183 088 words) from *Wiedza i Życie* magazine.

Table 3: Average CLASS and amount of years of education required to understand texts from selected corpora

Corpus	FOG	CLASS
Children's literature	5	3
Press	10	4
Wikipedia	10	5
Popular science	11	4
Legal acts	10	5

Firstly, we implemented two automated Taylor test methods evaluating suitability of the bigram language models for investigated text. One based on calculating the lowest perplexity score and the other on counting the number of gaps properly filled in in the test file.

Secondly, we implemented two variants of counting similarity scores between a vector representing the document and reference corpora: using TF-IDF and cosine weighting schemas.

The evaluations of all methods are presented in Tables 4 and 5. We have achieved very high accuracy especially for legal acts and children's literature. Lower accuracy for Wikipedia can be interpreted as the result of many different kinds of texts in the corpus.

Table 4: Percentage of properly assigned documents to their corpora in leave-one-out cross-validation using similarity algorithm (experiments were done on the smaller versions of corpora, about 35 000 words each)

Corpus	TF-IDF	Binary
Children's literature	100.00%	100.00%
Wikipedia	85.37%	85.37%
Legal acts	100.00%	100.00%
Press	73.91%	71.74%
Popular science	100.00%	100.00%

Table 5: Percentage of properly assigned documents to their corpora in leave-one-out cross-validation using Taylor-based algorithm

Corpus	Perplexity	Hit count
Children's literature	97.18%	93.79%
Wikipedia	61.11%	80.56%
Legal acts	100.00%	86.29%
Press	66.11%	71.66%
Popular science	68.31%	73.77%

5. An application for measuring readability

Based on all presented formulae and readability assessment methods we have implemented a new Web-based application intended to facilitate measuring readability of Polish texts. It currently accepts three types of input sources: plain text, uploaded file and URL and presents different readability indices, starting from Gunning FOG and Flesch-based Pisarek index to methods comparing distributional lexical similarity of a target text with reference texts and using statistical language modelling for automation of a Taylor test.

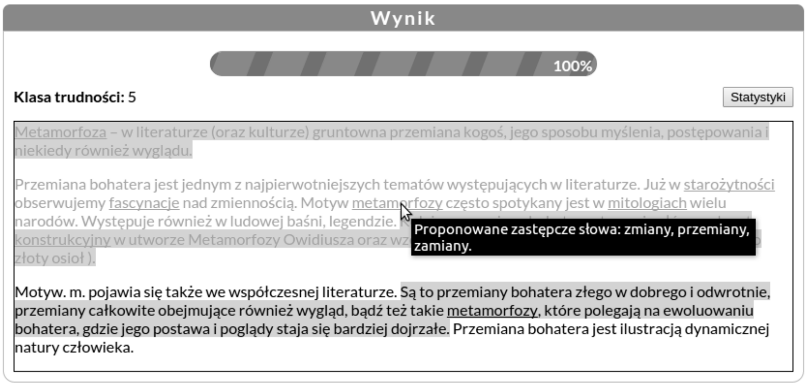
Apart from basic information about readability class for the whole text the interface also presents readability values for individual paragraphs and sentences, indicates difficult words and suggests thesaurus-based improvements (see Figure 1).

6. Conclusions and perspectives

To our best knowledge our work is the first empirical verification of Pisarek's equation and an attempt of verification of linguistic indicators of readability put forward in theoretical studies. The work described in the paper lays foundations of construction of readability indices of even higher predictive power than offered by other formula previously used for Polish.

Nevertheless, we need to mention certain limitations of our research, the first of which is representativeness of the surveys carried out, both with respect to the group of respondents and the set of texts being investigated. Future surveys should attempt at approaching much larger number of samples, allowing to measure linguistic variables of much higher diversity.

Taking into account the results of our psycholinguistic studies, much can still be done to investigate dependence of readability upon linguistic and cultural competences of a reader. Another useful field of research would be comparing readability of a text and its translations, e.g. in order to make a useful tool for European institutions, often (particularly in Poland) criticized for using too formal style, making the translation much more difficult to understand than the original. Last but not least, the application for measuring readability resulting from our work can be further improved.



Klasa trudności: 5

Metamorfoza – w literaturze (oraz kulturze) gruntowna przemiana kogoś, jego sposobu myślenia, postępowania i niekiedy również wyglądu.

Przemiana bohatera jest jednym z najpierwotniejszych tematów występujących w literaturze. Już w starożytności obserwujemy fascynacje nad zmiennością. Motyw metamorfozy często spotykany jest w mitologiach wielu narodów. Występuje również w ludowej baśni, legendzie.
konstrukcyjny w utworze Metamorfozy Owidiusza oraz wz
złoty osioł).

Proponowane zastępcze słowa: zmiany, przemiany, zamiany.

Motyw. m. pojawia się także we współczesnej literaturze. Są to przemiany bohatera złego w dobrego i odwrotnie, przemiany całkowite obejmujące również wygląd, bądź też takie metamorfozy, które polegają na ewoluowaniu bohatera, gdzie jego postawa i poglądy staja się bardziej dojrzałe. Przemiana bohatera jest ilustracją dynamicznej natury człowieka.

Figure 1: Application interface

Acknowledgements

The research financed from the research grant no. 2011/03/BHS2/05799, titled: *Mierzenie stopnia zrozumiałości polskich tekstów użytkowych (pozaliterackich)* (Pol. *Measuring the degree of readability of nonliterary Polish texts*) awarded by the National Science Centre of the Republic of Poland.

References

Banach, M., 2011. Tekst trudny, czyli... jaki? O czynnikach wpływających na trudność tekstu. *Polonica*, 31:27–35.

Bernhardt, E. B., 1991. *Reading development in a second-language*. Norwood, NJ: Ablex.

Copeland, L., T. Gedeon, and S. Mendis, 2014. Predicting reading comprehension scores from eye movements using artificial neural networks and fuzzy output error. *Artificial Intelligence Research*, 3(3):35–48.

Cygal-Krupa, Zofia, 1986. *Słownictwo tematyczne języka polskiego – zbiór wyrazów w układzie rangowym, alfabetycznym i tematycznym*. Kraków: Uniwersytet Jagielloński, Instytut Badań Polonijnych. Skrypty uczelniane nr 514.

Flesch, R., 1948. A new readability yardstick. *Journal of Applied Psychology*, 32:221–233.

Gajda, Stanisław, 1990. *Współczesna polszczyzna naukowa: język czy żargon?*. Instytut Śląski w Opolu.

Gizbert-Studnicki, Tomasz, 1986. *Język prawny z perspektywy socjolingwistycznej*. Państwowe Wydawnictwo Naukowe.

Gunning, Robert, 1952. *Technique of Clear Writing*. McGraw-Hill.

Imiołczyk, Janusz, 1987. *Prawdopodobieństwo subiektywne wyrazów: podstawowy słownik frekwencyjny języka polskiego*. Warszawa: Państwowe Wydawnictwo Naukowe.

Malinowska, Ewa, 1999. Język w urzędach. In Walery Pisarek (ed.), *Polszczyzna 2000. Orędzie o stanie języka na przełomie tysiącleci*. Uniwersytet Jagielloński. Ośrodek Badań Prawoznawczych, pages 75–96.

Markowski, Andrzej, 1990. *Leksyka wspólna różnym odmianom polszczyzny*. Warszawa: Wydawnictwo Wiedza o Kulturze.

Pisarek, W., 1972. *Frekwencja wyrazów w prasie: wiadomości, komentarze, reportaże*, volume 16 of *Biblioteka wiedzy o prasie, seria B*. Ośrodek Badań Prasoznawczych RSW „Prasa” w Krakowie.

Pisarek, Walery, 1966. Recepty na zrozumiałość wypowiedzi. *Zeszyty Prasoznawcze*, 2/3(28/29):44–53.

Presspublica, 2013. Rzeczpospolita corpus. [on-line] `http://www.cs.put.poznan.pl/dweiss/rzeczpospolita`.

Rada Języka Polskiego, 2012. Sprawozdanie o stanie ochrony języka polskiego za lata 2010–2011.

Royer, J. M., C. N. Hastings, and C. Hook, 1979. A sentence verification technique for measuring reading comprehension. *Journal of Reading Behavior*, 11:355–363.

Ruszkowski, M., 2004. *Statystyka w badaniach stylistyczno-składniowych*. Kielce: Wydawnictwo Akademii Świętokrzyskiej.

Seretny, Anna, 2006. Wskaźnik czytelności tekstu jako pomoc w określaniu stopnia jego trudności. *LingVaria*, (2):87–98.

Taylor, W. L., 1953. Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 30:415–433.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, series B*, 58:267–288.

Tikhonov, A. N. and V. Y. Arsenin, 1977. *Solution of Ill-Posed Problems*. Washington: Winston & Sons.

Wojtak, Maria, 1993. Styl urzędowy. In Jerzy Bartmiński (ed.), *Encyklopedia kultury polskiej XX w. T. II: Współczesny język polski*. pages 147–172.

Woronczak, J., 1976. O statystycznym określeniu spójności tekstu. In M. R. Mayenowa (ed.), *Semantyka tekstu i języka*. Wrocław: Ossolineum, page 165–173.