Jasnopis – A Program to Compute Readability of Texts in Polish Based on Psycholinguistic Research[1]

Łukasz Dębowski[1], Bartosz Broda[1], Bartłomiej Nitoń[1], and Edyta Charzyńska[2]

[1] Polish Academy of Sciences, Institute of Computer Science,
ul. Jana Kazimierza 5, 01-248 Warsaw, Poland
ldebowsk@ipipan.waw.pl
bartosz.broda@gmail.com
bartek.niton@gmail.com

[2] University of Silesia, Institute of Pedagogy
ul. Grażyńskiego 53, 40-126 Katowice, Poland
edyta.charzynska@us.edu.pl

Abstract. Readability of a text is a measure how difficult the text is to understand on average. The aim of the present paper is twofold. First, we have determined through a psychological experiment and statistical data analysis how readability of texts in Polish depends on syntactical and lexical statistics of the texts. Second, we have implemented a computer program, called Jasnopis, which computes readability of a given text according to the developed formula and suggests how to make the text easier to comprehend.

1 Introduction

Texts that circulate in public discourse, such as fairy tales, press articles, legal acts, or scientific articles, vary with respect to their ease of reading, called readability. Some degree of text difficulty stems from an intention of communicating complex meanings, but ideally we would expect that the intended ideas were put across as simply as possible. The reality is far from this ideal state. We are surrounded by more and more complex texts, such as legal decisions or medical leaflets, which we are supposed to understand. Unfortunately, these important texts are difficult to understand to nonspecialists, for they are written using a specific language register and contain very long sentences or highly specialized terms. We think that there is a need for a computer application that would help to measure how difficult a given text is to comprehend and would suggest possible ways of simplifying it. Consequently, in this paper we propose a computational method of predicting readability of texts in Polish.

A great deal of research concerning readability of texts has been already done for English. Since the end of 19th century, researchers in the United States have been

[1] The research reported in this paper has been funded by the NCN grant "Mierzenie stopnia zrozumiałości polskich tekstów użytkowych" no. 2011/03/BHS2/05799.

interested in variation of texts with respect to their ease of understanding (Sherman, 1893). In the 1920's the interest of readability researchers shifted towards practical application such as assessing difficulty of textbooks and adjusting them to progressing abilities of schoolchildren. Hence, various empirical formulas have been proposed that allowed to estimate readability of a text given its certain statistics (Lively and Pressey, 1923; Washburne and Vogel, 1928; Lewerenz, 1929; Patty and Painter, 1931). The next decade brought interest in measuring actual understanding by adults through psychological tests and using results of the experiments to scale and improve readability formulas (Dale and Tyler, 1934; Gray and Leary, 1935). In 1940's, Lorge (1944a,b) and Flesh (1948) observed that readability can be predicted surprisingly well using only two or three statistics related to syntactical and lexical text complexity, such as the average sentence length (ASL) and the average word length (AWL). Let us observe that restricting ourselves to the AWL, we would treat the text as a bag of words. In contrast, taking into account the ASL, we indirectly measure also the complexity of syntax, which is a desirable property.

An example of a simple readability predictor is the Flesh formula (Flesh, 1948):

$$\text{Text Readability} = 206.835 - 1.015 * \text{ASL} - 84{,}6 * \text{AWL}, \quad (1)$$

where ASL – average sentence length (in words), AWL – average word length (in syllables). The readability index given above ranges between 100 (a very simple text) and 0 (a very difficult text). Many more similar readability formulas have been advocated by various researchers since then (Dale and Chall, 1948; Gunning, 1952; McLauglin, 1969; Caylor et al., 1973; Kincaid et al., 1975). The FOG index by Gunning (1952) became particularly famous. It reads:

$$\text{Fog Index} = 0.4 * (\text{ASL} + \text{PHW}), \quad (2)$$

where ASL – average sentence length (in words), PHW – percentage of words longer than two syllables. Two other popular readability indices are ARI (Senter and Smith, 1967) and LIX (Bjornsson, 1968). ARI (Automated Readability Index) was proposed for English and it reads ARI = – 21.43 + 0.5 * ASL + 4.71 * AWL, where ASL – average sentence length (in words), and AWL – average word length (in characters). In contrast, LIX (Lasbarhetsindex) was designed for Swedish and it reads LIX = ASL + PHW, where ASL – average sentence length (in words) and PHW – percentage of words longer than 6 letters.

As we can see, each of the proposed readability formulas uses a bit different text statistics, with different coefficients, and returns values in a different range. Thus, there is a problem of putting readability indices onto a common human-readable scale. To overcome this problem, Dale and Chall (1948) proposed to scale readability index according to the number of years of education that is needed by the intended reader of the text. Another way of putting the readability index onto a common scale is to use some standardized and universal psychological test of text understanding and to construct the best predictor of this test based on text statistics. In fact, Taylor (1953, 1956) developed a method, called the Cloze test, which seems to measure how well human subjects understand a given text. The Cloze test consists in asking a person to complete gaps in a version of the text in which every 5th word has been deleted. The Cloze score is the percentage of gaps that have been completed correctly. It has been

confirmed that the Cloze score correlates well with other psycholinguistic methods of assessing text readability (Rankin 1959; Bormuth 1966). Using the Cloze test, quality of various readability formulas was computed by DuBay (2006). For example the Pearson correlation between the Cloze score and the Flesh formula (1) is 0.91, the same result was obtained for the Fog index (2) whereas the best result, correlation 0.93, was observed for the formula by Dale and Chall (1948).

It is reasonable to expect that readability formulas should be language dependent to a certain extent. The typical length of a sentence or a word clearly depends on a language. For this reason, readability formulas, particularly those suggesting the required level of reader's education, should be tuned to a particular language, such as Polish. Until recently, there was not much interest in the readability research for the Polish language. This research area started to gain more interest in the last few years, e.g., Broda et al. (2010), but the most known readability formula was proposed by Pisarek in 1960's (as reported in Pisarek, 2007):

$$\text{Text Difficulty} = \frac{1}{2}\sqrt{ASL^2 + PHW^2}, \tag{3}$$

where ASL – average sentence length (in words), PHW – percentage of words longer than three syllables. In his 2007 paper, Pisarek also published a graphical scale for computing readability, which corresponds to a bit different formula, namely

$$\text{Text Difficulty} = \frac{ASL}{3} + \frac{PHW}{3} + 1. \tag{4}$$

Pisarek has not verified his formulas in a psycholinguistic experiment on human subjects. In contrast, we will discuss the results of a larger research project in which:

1. The Cloze test and an open-ended question test was applied to 35 texts in Polish, read by a sample of 1759 persons.
2. The results of the psycholinguistic experiment were analyzed statistically to provide a new readability formula, which is better than Pisarek's formula.
3. A computer application, called Jasnopis, was written to compute this readability formula for a given text in Polish. Besides estimating readability according to our new formula, Jasnopis returns many other text statistics for a given text and prompts how to adjust the text to make it more readable.

The organization of the paper is as follows. In Section 2 we discuss the psycholinguistic experiment. Section 3 is devoted to statistical analysis of the data and the development of a new readability formula. In Section 4 we describe the Jasnopis program. Conclusions are presented in Section 5.

2 The psycholinguistic experiment

The purpose of the psycholinguistic experiment was fourfold: a) to validate Pisarek's formula, b) to find text variables that influence text readability but are different than ASL and PHW, c) to identify psychological variables that influence text compre-

hension (such as reader's interest in the text), and d) to use the results of the experiment to develop a new readability formula.

Before conducting the psycholinguistic experiment, we have constructed an a priori scale of seven classes of growing text difficulty, measured in the number of required years of education to understand the text correctly. (Class 1 are texts that should be understandable by students of elementary schools, whereas class 7 are those whose comprehension requires the doctorate level of education.) Subsequently, we have compiled a corpus of texts that presumably belong to the respective text difficulty classes. The texts were chosen by the project members: psycholinguists, linguists and computational linguists. Using the FOG index (2), we have next chosen 5 most typical texts for each difficulty class. In this way we have obtained a sample of 35 texts, on which we performed the psycholinguistic experiment.

In the experiment, 1759 persons have participated: 63% female and 37% male. Participants were of diversified ages (average=35.6, standard deviation=14.65, min=15, max=87), education (from elementary to higher), occupation (including manual workers, white-collars, unemployed and pensioners), coming from villages (20%), smaller towns (28.3%), medium-size cities (28.7%) and big cities (21.7%). Each participant of the study received 2 texts to read. Each text was accompanied with a set of 5 open-ended questions or the Cloze test. The experiment was performed using the traditional pen and paper approach.

Having collected the survey results, we performed statistical analysis of the data. We found out that Pisarek's formula was highly correlated with the results of the Cloze test ($r=-0.69$, $p=0.001$) and the open-ended questions ($r=0.8$ $p=0.001$). Moreover, we found out that the test results are highly correlated not only with the variables used by Pisarek (for ASL $r_{cloze}=-0.63$, $r_{questions}=-0.71$; for PHW $r_{cloze}=-0.67$, $r_{questions}=-0.83$; $p=0.001$) but also with some other text statistics. Among the top correlated variables were: the percentages of nouns, terminology, abstract nouns, foreign words, gerunds, verbs, the ratio of nouns to verbs, and the subjective probability of words (Imiołczyk, 1987). These results suggest that using these variables we may propose a readability formula that outperforms the Pisarek or Fog indices.

3 A new readability formula

Given the psycholinguistic survey described in Section 2, we were in position to analyze how readability of a text depends on particular text statistics. At our disposal we had 35 texts – 5 texts per each of 7 difficulty classes. For each text, we had the results of two psycholinguistic tests measuring the text comprehension – the Cloze test and the open question test. These were our response variables. Moreover, for each text, we have measured 33 lexical and syntactical text statistics, such as ASL, PHW, the percentages of nouns, terminology, abstract nouns, foreign words, gerunds, verbs, the ratio of nouns to verbs, and the subjective probability of words. These were our explanatory variables. The goal of the consecutive data analysis was to find out (i) how the difficulty class of a text could be predicted from the response variables and (ii) how the response variables could be predicted from the explanatory variables. As

a result, we obtained a new formula for text readability which was implemented in the Jasnopis tool, to be discussed in the next Section 4.

Since there were not so much data, we looked for a linear formula for readability:

$$Y_i = A_0 + \sum A_j X_{ij} + \varepsilon_i \quad . \tag{6}$$

where: Y_i – a chosen response variable for the i-th text, X_{ij} – j-th explanatory variable for the i-th text, ε_i – random noise, i=1,...,N, N=35 – the number of texts, K=33 – the number of explanatory variables.

The number of texts N=35 is close to the number of explanatory variables K=33. In this situation, choosing the coefficients A_j through least squares regression would lead to terrible overfitting, that is, formula (6) would not predict comprehension of texts different to the training sample. A possible solution to this problem is to use least squares regression with regularization, such as Lasso or Ridge regression (Tibshirani, 1996; Tikhonov, Arsenin, 1977). The least squares regression with regularization consists in choosing such coefficients A_j that minimize expression

$$\sum \left[Y_i - A_0 + \sum A_j X_{ij} \right]^2 + \beta \sum (A_j)^\alpha \quad , \tag{7}$$

where $\alpha = 1$ for the Lasso regression and $\alpha = 2$ for the Ridge regression, whereas β is chosen by cross validation. (For the least squares regression without regularization, we minimize expression (7) with $\beta = 0$.)

A priori it was not obvious that the Lasso or Ridge regression would give the best results. Therefore we compared these two methods with three other methods:

1. The baseline model: Text difficulty does not depend on text. That is, we minimized expression (7) with $\beta = 0$ and A_j being nonzero only for $j = 0$.
2. The least squares regression with two explanatory variables, ASL and PHW, as in Pisarek's formula (4). That is, we minimized expression (7) with $\beta = 0$ and A_j being nonzero only for $j = 0,1,2$.
3. The weighted average (committee) of least squares regressions with three explanatory variables: ASL, PHW, and one of the remaining 31 variables. That is, first, for each k in range {3,...,K}, we minimized expression (7) with $\beta = 0$ and A_j being nonzero only for $j = 0,1,2,k$, and second, we took an average over k of so obtained A_j.

The quality of each of these five methods of determining coefficients A_j was assessed by leave-one out cross validation. That is, we removed one text from the training sample, we fitted the coefficients A_j to the remaining texts, and we checked how well the model predicted the response variable for the removed text. The prediction error, defined as difference between the prediction and the response variable, was recorded for each text. We made a boxplot graph of the prediction error and we chose the method for which the prediction error is the smallest in general.

We applied this procedure independently to three response variables: the Cloze score, the open question score, and a weighted average of these two scores, which we will refer to as the weighted readability score. The relative prediction errors for predicting the Cloze score and the open question score were similar whereas they were substantially smaller for the weighted readability score. Therefore we suppose that the weighted readability score is a better predictor of the actual text readability than the

Cloze score or the open question score considered individually. The boxplots of the prediction error for the weighted readability score and the five different methods of determining coefficients A_j are presented in Fig. 1.

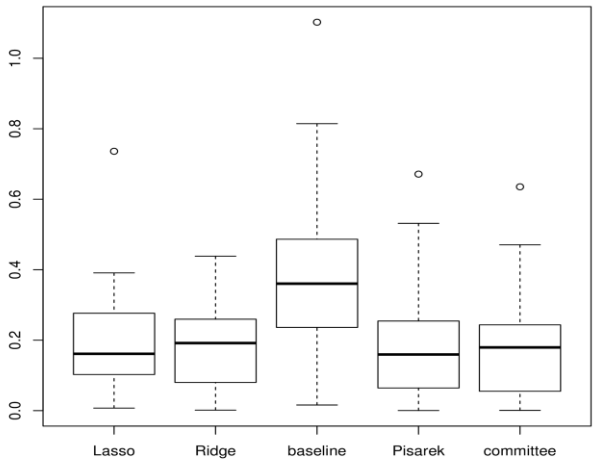

Fig 1. Boxplots of prediction error of the weighted readability score for the five methods of determining coefficients A_j described in Section 4.

In Fig. 1 we can see that the Ridge regression yields the smallest maximal prediction error. Therefore formula (6) with coefficients A_j given by the Ridge regression for the weighted readability score was adopted as a part of a new formula for readability of texts implemented in the Jasnopis tool. The second ingredient of the new Jasnopis formula for readability is a projection of the Rigde regression onto the scale of 7 difficulty classes, introduced in Section 2, so that the final readability score be more human readable. As we can see in Fig. 2, the dependence between the weighted readability score and the difficulty class is linear in a good approximation.

4 The Jasnopis program

The final aim of our project was to construct a computer application for measuring readability of Polish texts. The tool, called Jasnopis, implements the measure of text readability given by the Ridge regression, described in the previous section, and additionally computes a number of other text statistics, which may be of interest to end-users. The prototype web-based application is available at http://jasnopis.pl. Jasnopis accepts many different documents types on its input: from a plain text and a website URL address to document formats supported by OpenOffice. The first step of the document processing by Jasnopis is text extraction, which is a nontrivial problem on its own. Next, we perform morphological analysis using Morfeusz (Woliński, 2006, and part-of-speech tagging using WCRFT (Radziszewski, 2013). In the later stages of

document processing we use various tools and resources like the frequency lists from National Corpus of Polish, NCP (Przepiórkowski et al., 2012), the plWordnet (Piasecki et al., 2009) or the subjective probability lists of Imiołczyk (1987).

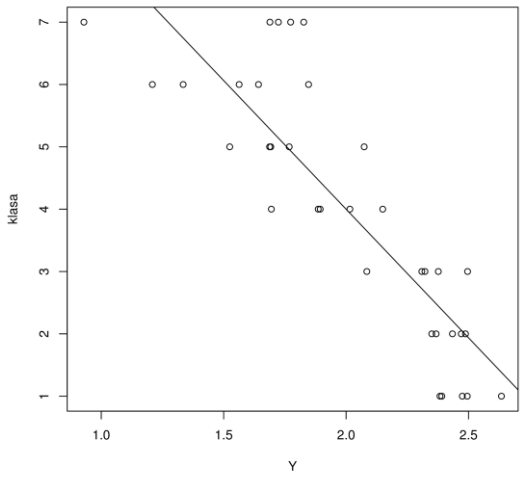

Fig. 2. The dependence between the weighted readability score (Y) and the difficulty class (klasa) for 35 texts.

The main statistic calculated by Jasnopis is the difficulty class on the 7-point scale, described in Section 2. To estimate the difficulty class we rescale the score of the Ridge regression via the linear function depicted in Fig. 2. Additionally, Jasnopis calculates the following indices: a few variants of the FOG index (2), Pisarek's indices (3) and (4), an automated Taylor test, similarity graphs and additional text statistics. Both the FOG and Pisarek indices depend on the variable PHW – the percentage of words longer than a certain number of syllables (for Polish, three). Since Polish is an inflective language, it is not a priori obvious whether when computing PHW one has to consider the orthographic forms or the base forms of words.

In addition to the above variants of the FOG and Pisarek indices we also calculate some discounted versions of them, because the original definition of "hard" words in PHW is too simplistic: not every long word is a difficult word. Many words that are long are so common, that an average person has no difficulty in understanding them. Thus, we exclude most frequent words in NCP from PHW calculation. There are many words that an average Polish native speaker knows, but which are rarely used in writing. Thus, from the PHW calculation, we also remove words that can be found on the subjective probability lists by Imiołczyk (1987).

Another text statistic implemented in Jasnopis, the automated Taylor test is somewhat inspired by the Cloze test by Taylor (1953, 1956). Instead of using human subjects, we train a few statistical language models and we check which one is the best at predicting the text. For simplicity, we use bigram language models (Jurafsky & Martin, 2008), trained on 5 reference corpora corresponding to different classes of text

difficulty. Each bigram model assigns a probability of a word w_i conditioned on a single previous word w_{i-1} as

$$p(w_i|w_{i-1}) = \frac{c(w_{i-1}w_i)}{\sum_{w_i} c(w_{i-1}w_i)}, \tag{5}$$

where $c(w_{i-1}w_i)$ denotes the number of times the bigram $w_{i-1}w_i$ occurred in a training corpus. In this way we obtain seven bigram models corresponding to the seven classes of text difficulty. Then, the difficulty class of a given new text is determined as the class of difficulty corresponding to the bigram model with the highest total probability (that is, the lowest perplexity).

Yet another automated score of text difficulty implemented in Jasnopis is also based on reference corpora. Namely, instead of building language models we use the Vector Space Model (Salton et al., 1975). In this approach the text is represented as the n-dimensional vector $D=[d_1,d_2,...,d_n]$, where d_i are frequencies of words appearing in the text. To compare two texts or corpora we use the cosine distance between the corresponding vectors. Subsequently, the difficulty class of a given new text is determined as the class of difficulty corresponding to the reference corpus with the highest smallest cosine distance. Let us observe that this procedure ignores syntactic difficulty of the text since we treat the text as a bag of words. Thus, we only compare texts on a lexical level. Nonetheless, as we have determined, the lexicon is an important factor in measuring readability of a given text.

We have experimentally verified performance of both the automated Taylor test and the similarity model. Using leave-one-out cross validation we achieved from 68.31% to 100% (depending on the reference corpus difficulty class) precision for automated Taylor test and from 71.74% to 100% for similarity-based approach. See Broda et al. (2014) for details.

Besides returning a number of text statistics, Jasnopis supports also computer-aided text simplification. In a given text, it marks difficult paragraphs, too long sentences, and hard words. For hard words, substitution suggestions are presented sometimes using synonyms, hyponyms and hyperonyms from the plWordNet. No word-sense disambiguation is implemented, so the user has to make the final decision. Since simplifying a document in a web-based environment might not be very convenient, we have developed also Jasnopis plugins for OpenOffice and MS Word that cover most important functionalities of the web application.

5 Concluding remarks

In this paper we have presented an approach for constructing a new readability formula for Polish, based on Ridge regression. We use 33 lexical and syntactic text variables for predicting the text difficulty class, which is an improvement over the received readability formulas, which only use two variables. The regression coefficients in the Ridge regression were fitted to the empirical text comprehension data – a psycholinguistic experiment with 35 texts and 1759 subjects. We have also presented a computer program for measuring text readability. The application, called Jasnopis, not only

implements the new formula but also provides other methods for measuring readability, both new and standard. By showing difficult sentences and words in a text, Jasnopis supports computer-aided text simplification, as well.

The proposed approach to measuring readability can be extended in several ways. One might search for additional explanatory text features. Especially, sophisticated syntactic features based on parse trees might provide additional benefits. Also, one could use other machine learning approaches to come up with even smaller error rates. Since Jasnopis already provides a few different methods for measuring readability, a straightforward approach would be to combine them using for example bagging. Last, but not least, having the ability to measure readability for Polish is a necessary step for (semi) automatic text simplification, which is an obvious direction of further research.

References

Bjornsson, C. H. (1968) *Lasbarhet.* Liber, Stockholm.

Bormuth, J. (1966) Readability: A new approach. *Reading Research Quarterly*, 1:79-132.

Broda, B., Maziarz, M., Piekot, T., Radziszewski, A. (2010) Trudność tekstów o Funduszach Europejskich w świetle miar statystycznych. *Rozprawy Komisji Językowej Wrocławskiego Towarzystwa Naukowego*, 37.

Broda, B., Ogrodniczuk, M., Nitoń, B., Gruszczynski, W. (2014) Measuring Readability of Polish Texts: Baseline Experiments. In: *Proc. of the 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland*.

Caylor, J. S., Stitch, T. G., Fox, L. C., Ford, J. P. (1973) Methodologies for determining reading requirements of military occupational specialties. Technical Report 73-5, Human Resources Research Organization, Alexander, Virginia.

Dale, E., Chall, J. S. (1948) A formula for predicting readability. *Educational Research Bulletin*, 27:1-20, 37-54.

Dale, E., Tyler, R. (1934) A study of the factors influencing the difficulty of reading materials for adults of limited reading ability. *Library Quarterly*, 4;384-412.

DuBay, W. H. (2006) *Smart Language: Readers, Readability, and the Grading of Text.* Impact Information, Costa Mesa.

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

Gray, W., Leary, B. (1935) *What Makes a Book Readable.* University of Chicago Press, Chicago.

Gunning, R. (1952) *The Technique of Clear Writing.* McGraw-Hill, New York.

Imiołczyk, J. (1987), *Prawdopodobieństwo subiektywne wyrazów: podstawowy słownik frekwencyjny języka polskiego.* Warszawa.

Jurafsky, D. and Martin J. H. (2008). Speech and Language Processing (2nd Edition). Pearson Prentice Hall.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., Chissom, B. S. (1975) Derivation of new readability formulas (Automated Readability Index, Fog Count, and Flesch Read-

ing Ease Formula) for Navy enlisted personnel. CNTECHTRA Research Branch Report, pp. 8-75.

Lewerenz, A. S. (1929) Measurement of the diffi-culty of reading materials. *Los Angeles Educational Research Bulletin*, 8:11-16.

Lively, B. A., Pressey, S. L. (1923) A method for measuring the 'vocabulary burden' of textbooks. *Educational Administration and Supervision*, 9:389-398.

Lorge, I. (1944a) Predicting readability. *Teachers College Record*, 45:543-552.

Lorge, I. (1944b) Word lists as background for communication, *Teachers College Record*, 45:543-552.

McLaughlin, G. H. (1969) SMOG grading—a new readability formula, *Journal of Reading*, 22:639-646.

Patty, W. W., Painter, W. I. (1931) A technique for measuring the vocabulary burden of textbooks , *Journal of Educational Research*, 24:127-134.

Piasecki, M., Szpakowicz, S., Broda. B. (2009) A wordnet from the ground up. Oficyna wydawnicza Politechniki Wroclawskiej.

Pisarek, W. (2007) O mediach i języku. In: *Jak mierzyć zrozumiałość tekstu?*, pp.245-262. Universitas, Kraków.

Przepiórkowski, A., Bańko, M., Górski, R. L., and Lewandowska-Tomaszczyk, B. (2012). Narodowy Korpus Języka Polskiego. Wydawnictwo Naukowe PWN, Warsaw

Radziszewski, A (2013). A tiered CRF tagger for Polish. In: *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*. Springer Verlag.

Rankin, E. F. (1959) The cloze procedure – Its validity and utility. In: Causey O., Eller W. (eds.) *Eighth Yearbook of the National Reading Conference*, pp. 131–142.

Sherman L. (1893) *Analytics of literature: A manual for the objective study of English prose and poetry*. Ginn and Co, Boston.

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

Taylor, W. L. (1956) Recent developments in the use of 'cloze procedure'. *Journalism Quarterly*, 33:42–48.

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

Tikhonov, A. N., Arsenin, V. Y. (1977) *Solution of Ill-posed Problems*. Winston & Sons, Washington.

Salton, G., Wong, A. and C. S. Yang (1975) A vector space model for automatic indexing. Communications of ACM, 18(11):613–620.

Senter, R. J., Smith, E. A., (1967) Automated Readability Index. Wright-Patterson Air Force Base, p. iii. AMRL-TR-6620.

Washburne, C., Vogel, M. (1928) An objective method of determining grade placement of children's reading material, *Elementary School Journal*, 28:373-381.

Woliński M., 2006. Morfeusz – a Practical Tool for the Morphological Analysis of Polish. Intelligent Information Processing and Web Mining, Springer.