Measuring Readability of Polish Texts: Baseline Experiments*

Bartosz Broda[1], Maciej Ogrodniczuk[1], Bartłomiej Nitoń[1], Włodzimierz Gruszczyński[2]
[1]Institute of Computer Science
Polish Academy of Sciences
Jana Kazimierza 5, Warsaw, Poland

[2]Warsaw School of Social Sciences and Humanities
Chodakowska 19/31, Warsaw, Poland

Abstract

Measuring readability of a text is the first sensible step to its simplification. In this paper we present an overview of the most common approaches to automatic measuring of readability. Of the described ones, we implemented and evaluated: Gunning FOG index, Flesch-based Pisarek method. We also present two other approaches. The first one is based on measuring distributional lexical similarity of a target text and comparing it to reference texts. In the second one, we propose a novel method for automation of Taylor test – which, in its base form, requires performing a large amount of surveys. The automation of Taylor test is performed using a technique called statistical language modelling. We have developed a free on-line web-based system and constructed plugins for the most common text editors, namely Microsoft Word and OpenOffice.org. Inner workings of the system are described in detail. Finally, extensive evaluations are performed for Polish – a Slavic, highly inflected language. We show that Pisarek's method is highly correlated to Gunning FOG Index, even if different in form, and that both the similarity-based approach and automated Taylor test achieve high accuracy. Merits of using either of them are discussed.

Keywords: readability, distributional similarity, text understandability

1. Introduction

Text readability is the measure used to determine how easy (or difficult) a given text can be to read and understand. Various methods are applied for this purpose, originating in psycholinguistics, cognitive linguistics or statistical linguistics. There are many cues that can indicate the level of readability of a given text. The cues range from simple ones like lengths of the sentences to very complicated like the depth of the parse tree of the sentence. Usually, calculating the cues for a given text is a time-consuming task, whose difficulty rises with complexity of the cues used. As the cue calculation is a repetitive task, we hope that it can be automated by the computer. Thus, our aim is to construct a computer system focused on the analytic approach – computing a numerical indicator based on selected features of a text – and comparing readability assessments using different methods.

The task is highly language-dependent. Even the definition of one of the simplest of cues, i.e., length of the sentence, is a language dependent (e.g., if we measure the length of the sentences in words, then one has to precisely define what a word is). Most research in the field has obviously been conducted for English, but we want to focus on another language, namely Polish. The field is largely unexplored for Polish. Second, as a Slavic, highly inflected language, Polish represent a few interesting challenges, especially for automated methods.

Results of research on readability are of immediate practical significance: they create a rapid and objective way of rating intelligibility of various texts – manuals, medicine brochures, legal regulations, school textbooks and many others. Secondly, they can help formulate rules for writing texts accessible for people with a given educational background. Last but not least, they have theoretical consequences since they can provide grounds for verification or hypotheses concerning influence of certain features of a Polish text (e.g. negation, double negation, use of passive voice, impersonal forms, rare vocabulary or loan words etc.) on its comprehensibility. We hope that availability of a computer system for readability evaluation may contribute to improvement of language communication in Poland.

The most popular Polish formula for computing readability has been proposed in 1960's by Walery Pisarek based on research of Rudolf Flesch (Flesch, 1960) and Josef Mistrik (Mistrík, 1968). It takes into account only two features of the text: the mean length of a sentence and the percentage of "potentially difficult" words (longer than three syllables). The existing methods, including the Pisarek method, were established in times when neither frequency lists nor large text corpora were available. Thus, it wasn't possible to apply the methods on a large scale.

One of the main contributions of this paper is the development of an automated version of Taylor test. Traditionally, in the Taylor test, every fifth word from some text is removed. Then, a person is asked to fill in the gaps. With a large population of language users filling in the gaps one can draw some conclusions about the readability of a given text. Instead of an approach based on interviews with (preferably) a large sample of language users, we train an n-gram language model on reference corpora. For language models to give good n-gram estimates, the corpora should be large, authored by different people and represent different levels of readability. Then we apply the language mod-

*The work reported here was carried out within the *Measuring the degree of readability of nonliterary Polish texts* project financed by the Polish National Science Centre (contract number 2011/03/B/HS2/05799).

els to fill in the gaps in a text with removed words (or we can measure the perplexity of given language models). The last step of automated Taylor test involves comparing performances of different language models in the task. The readability level of the reference corpus used to train the best performing language model now roughly corresponds to the readability of input text.

This paper is organized as follows. First, we summarize relevant approaches in Section 2. We then present an overview of our system's architecture, which is followed by the evaluation section. Finally, we conclude the paper with a short summary and directions for future work.

2. Related Work

Evaluations of readability fall into two broader categories: quantitative measures and cloze tests, which we may call psycholinguistic. Quantitative methods involve mathematical and statistical analysis of specific linguistic features of the text. Different metrics differ in their range of assessed features and complexity of formulas. Among these, a method based in the field of information theory (whose theoretical assumptions stand out among the rest) deserves a particular mention. The cloze (deletion) test, on the other hand, involves surveying speakers and can be considered a psychological or sociological test rather than a strictly linguistic one.

Measuring readability dates back to the early medieval times, when word counts were used to estimate ease of reading particular works (Taylor and Wahlstrom, 1986). Medieval Talmudists, for instance, counted occurrences of specific words and letters in order to distinguish their underlying meanings. 19th century investigations into the history of the English language led to an observation that sentence lengths are correlated with their difficulty. Around the same time, in Russia, Nikolai Rubakin published a list of 1,500 words which he estimated should be known by most Russians, based on his studies of everyday writings by people among the general public (Choldin, 1979).

Modern approaches to studying readability were pioneered by American scientists in the early 20th century, when frequency lists came into prominence. Within several years, dozens of readability formulas were proposed, mostly for American English. (Pisarek, 1969) has been the only one to attempt a creation of a readability metric for Polish, inspired by the American approach. His formula is seldom referenced, although it may be familiar to some journalists due to its inclusion in the author's other work, a popular handbook of journalism (Pisarek, 2002).

Among the various metrics for evaluating readability of English text, only a couple have achieved notability. These include: Flesch's reading ease formula, later modified to become Flesch-Kincaid readability test (Kincaid et al., 1975), Dale-Chall readability formula (Dale and Chall, 1948), Gunning FOG index (Gunning, 1971), Fry readability formula (Fry, 1968), Bormuth readability index (Bormuth, 1966). Their value and recognition in the United States is supported by the fact that some of them have been included in the American version of the popular word processor Microsoft Word, while the Flesch-Kincaid test is used by the United States Department of Defense for the purpose of evaluating readability of all documents produced for the Department.

Rudolf Flesch created the first widely recognised readability test in 1943 (Flesch, 1943). The formula measured an average number of syllables in a word and an average number of words in a sentence. Flesch reading ease formula looks as follows:

$$T = 206.835 - (1.015 \ \times T_w) - (84.6 \ \times \ T_s)$$

where: T – readability score; T_w – index of syntactic complexity (average words per sentence); T_s – index of lexical complexity (average syllables per word).

In addition, measuring the number of words and syllables requires following assumptions behind Flesch's methodology, i.e., only the main body of text is considered (ignoring headlines, titles, signatures, etc.), while acronyms, abbreviations, hyphenated words, numbers, special characters and their combinations are counted as single words.

Flesch-Kincaid readability test is a specific version of Flesch's formula, whose result is an estimated number of years of education in the American school system necessary for comprehending a given text. The same underlying idea is used in the Gunning FOG index (Miles, 1990) and a number of other measures.

In recent years, it has been noted that formulas such as Flesch reading ease are only concerned with the surface structure of the text, whereas its readability may depend on a number of other factors. Such criticisms have been raised by cognitive linguists and, later, by the so-called text linguists, who typically invoke the concepts of cohesion and coherence here. Both terms relate to the way words, structures and concepts involved in the text interact at the different levels of language, discourse, and real-world knowledge (Goldman et al., 1999). Cohesion refers to links between elements of text: words, phrases, and sentences. Coherence, meanwhile, is a result of the interaction between textual cohesion and the reader. A given level of cohesion may correspond to a different level of coherence across different readers. Coherence refers to relations created as a result of the influence of linguistic expressions on the reader's cognitive model. In other words, the authors of the theory see coherence and cohesion as crucial components of readability. Greater cohesion and coherence of a text translates into its readability.

An advanced system for establishing the level of cohesion of coherence of English text (and thus assessing its readability) is being built in the United States. The Con-Metrix program is being designed as an interactive system consisting of computational, syntactic and lexical modules. The application is supposed to analyse the cohesion of a given text based on a set of established markers, then add different measures of readability for different levels of coherence. The authors intend to make it possible for the program to analyse syntactic structure to an extent which would allow for filling in gaps in its coherence, thus increasing its readability (Goldman et al., 1999).

Apart from measures based on analysing text, there are also empirical approaches referencing psycholinguistic experiments. This type of methodology was pioneered by an

American journalist Wilson Taylor (Taylor, 1953). His suggested "cloze procedure" method involves removing every nth word in a text (most commonly every 5th word) and replacing it with a gap, the size of which is always the same. A group of people (sharing demographic and psychographic features with the "target group" for a given text) is then asked to fill in the missing words. The percentage of properly reinserted words is treated as an index of readability of the text. The theory behind Taylor's measure, also known as the "cloze deletion test", references gestalt psychology, i.e., the empirical observation that every human being has an ability and a tendency to fill in elements missing from a (perceived) whole. The term "cloze" itself derives from the word "closure", referring to the tendency of the human mind to perceive complete images despite missing pieces.

The same principle may be applied to natural language. If we assume that a message expressed in natural language constitutes a certain whole (albeit a more complicated one than a simple geometric shape), an average language user will similarly attempt to fill in the defective message in accordance with their own knowledge and linguistic experience.

Taylor's test certainly takes into account more factors than any fixed formula used to evaluate readability of a text. Furthermore, it makes it possible to avoid the issue of predicting degrees to which particular linguistic features impact readability. Even if we know that sophisticated vocabulary and complicated syntax negatively impact readability, we cannot know which one has greater weight. In addition to this, relevance of individual factors may vary depending on age, education or individual skills of a particular reader. It is to be expected that certain core factors responsible for readability among child readers may be irrelevant for adults.

The "cloze deletion test", by its very nature, involves not only all of the factors known to be relevant to readability, but also potentially unknown and undiscovered ones. What is more, it takes into account extralinguistic features, such as familiarity with the topic or the reader's general interest in the subject. It also evaluates certain semantic features, e.g.: nonsensical word links, odd syntactic structures, or use of demonstrative pronouns as anaphora.

(Tanaka-Ishii et al., 2010) present an interesting approach based on machine learning. Instead of using traditional machine learning algorithms to train a complete model for different readability levels, they train a relation for sorting texts. Their method requires only easy and difficult texts, as opposed to standard machine learning approaches. Thus, the workload for developing training data is much smaller.

An approach to measuring the readability for Polish was previously presented by (Broda et al., 2012). The work done is somewhat similar to ours (using similarity-based approach, Taylor test — but performed manually — and FOG index). They also constructed a web-based system called Logios, but its functionality at the time of writing is limited to measuring FOG index of input text. On the other hand, they focused on manual and semi-automated analysis of target texts.

3. Measuring Readability in Jasnopis

As the first step of our work, we have implemented a web-based application for measuring the readability of a given text called *Jasnopis*[1]. At the moment, we focus on four methods of measuring readability:

1. FOG index (two variants: using words and base forms of words).

2. Pisarek's index (four variants: linear and non-linear versions using words and base forms of words).

3. Automated Taylor test (two variants: based on perplexity and hit count).

4. Measuring similarity (two variants: based on binary features and *tf.idf* weighting method).

Additionally, a few properties of a text are calculated: number of paragraphs, number of sentences, number of words, number of difficult words, average length of a sentence, average length of a paragraph, percentage of difficult words, percentage of nouns, percentage of difficult nouns, percentage of verbs and difficult verbs, percentage of adjectives and difficult adjectives and the ratio of nouns to verbs. Those properties can be used by the author as a source of additional analysis to help in meeting his specific readability goals of the text he or she is writing.

3.1. The FOG Index

We use a standard approach to calculation of a Gunning FOG index:

$$FOG = 0.4 \times \left(\frac{words}{sentences} + 100 \frac{complex\ words}{words} \right)$$

The first ratio can be interpreted as an average sentence length (ASL) and the second ratio as the percentage of complex words (PCW). A complex word is a word that has more than three syllables (of course, this is a simplification and approximation of a difficult problem of determining what a complex word is for a given speaker). The 0.4 factor was tuned experimentally so that the values of FOG index would roughly correspond to number of years of educations required to understand a given text.

As Polish is an inflected language, in the calculation of the index one has to consider its rich morphology. Thus, we experimented with two variants of treating words. First, we use the original approach, that is, we use orthographic (orth) forms of words in calculation. In the second approach, we first run a morphological analyser called Morfeusz[2] on the text. For each word, Morfeusz assigns all potential analyses of the word. A single analysis consists of the word's base form and a morpho-syntactic description of a word. For example, for ambiguous word *lata* ('years', '(it) flies' or 'summers'), we can have the following analyses:

[1] The name is a neologism consisting of Polish words *jasno* (clear) and *pisać* (to write). See jasnopis.pl.

[2] http://sgjp.pl/morfeusz/index.html.en.

```
<tok>
  <orth>lata</orth>
  <lex>
    <base>rok</base>
    <ctag>subst:pl:nom:m3</ctag>
  </lex>
  <lex>
    <base>rok</base>
    <ctag>subst:pl:acc:m3</ctag>
  </lex>
  <lex>
    <base>lato</base>
    <ctag>subst:sg:gen:n</ctag>
  </lex>
  <lex>
   <base>lato</base>
   <ctag>subst:pl:nom:n</ctag>
  </lex>
  <lex>
   <base>lato</base>
   <ctag>subst:pl:acc:n</ctag>
  </lex>
  <lex>
   <base>lato</base>
   <ctag>subst:pl:voc:n</ctag>
  </lex>
  <lex>
   <base>latać</base>
   <ctag>fin:sg:ter:imperf</ctag>
  </lex>
</tok>
```

It is clear from the example above that we need to select the contextually appropriate analysis. For this purpose we use a WCRFT tagger (Radziszewski, 2013).

3.2. Pisarek's Index

There are two versions of Pisarek's index: a linear (P_L)[3] and non-linear (P_{NL}) one.

$$P_L = \frac{1}{3} \times ASL \times \frac{1}{3} \times PCW + 1$$

$$P_{NL} = \frac{1}{2}\sqrt{ASL^2 + PCW^2}$$

Where ASL is average sentence length and PCW is a percentage of complex words similarly as in FOG index.

We experimented with both base forms and orthographic form of words, as in the case of FOG index.

3.3. Automated Taylor Test

The automated Taylor test group, called "taylor" in our applications and in the Experiments section, represents analysing input text by automatic Taylor test. We train bigram[4] language models (Jurafsky and Martin, 2008) on reference corpora. A statistical bigram Language Model (LM) assigns a probability of word w_i based on the probability of previous word w_{i-1} in the following way:

$$p\left(w_i \mid w_{i-1}\right) = \frac{c\left(w_{i-1}w_i\right)}{\sum_{w_i} c\left(w_{i-1}w_i\right)}$$

Where $c(w_{i-1}w_i)$ denotes the number of times a bigram $w_{i-1}w_i$ occurred in a training corpus.

We train *n* LMs using *n* reference corpora. *n* is the number of readability levels that we want to capture (5 in our experiments). The corpora used should be large, authored by different people and represent different readability levels for a language model to give good bigram estimates. We then apply the language model to fill in the gaps in a text with removed words (or we can measure the perplexity of given language model).

The last step of automated Taylor test involves comparing performances of different LMs in the task. The readability level of the reference corpus used to train the best performing LM now roughly corresponds to the readability of input text.

Both variants of the algorithm use bigrams for selecting the optimal language model for a delivered file. They use bigram language models, where each word is represented by a concatenation of its part of speech tag and base form. Perplexity-based algorithm checks perplexity of test data with each of the existing language model. The proper language model is considered to be the one which acquired the lowest perplexity score:

$$P_{LM}\left(T\right) = 2^{H_{LM}(T)}$$

$$H_{LM}\left(T\right) = -\frac{1}{W_T} log_2 p(T)$$

Where $H_{LM}(T)$ is an entropy of a LM on text T.

The second variant of the algorithm cuts out every fifth word from the test file, then, using each bigram LM, fills those gaps with the most likely suitable words. After each gap is filled, it compares added words with the original file. The language model with the highest number of properly filled words is selected as the proper language model for test file. This algorithm is considered to be one-to-one implementation of automatic Taylor test.

3.4. Similarity

Having built some reference corpora we can also try another way of determining a given text's readability. One can measure similarity between input text and reference corpora. High similarity score denotes high similarity to texts in a given reference corpus which hopefully corresponds to the readability level of the input text.

One of the best-known methods of representing a document for similarity computation between documents is the Vector Space Model (Salton et al., 1975). In this approach, a document is represented as an *n*-dimensional vector D=[d_1,d_2,...,d_n], where d_i corresponds to words appearing in document D. To compare two documents one now has to compare two high-dimensional vectors. There are many ways to do it, but one of the most widely used measures is a cosine (Manning and Schütze, 1999). In practice, the values in the vector D do not just denote the number of occurrences of words in document D. There are many

[3] Reconstructed based on a graph in (Pisarek, 2007), p. 259.

[4] A *bigram* is a sequence of two words. The choice of using bigrams and not higher order n-grams in this work was arbitrary. In the future we will evaluate the performance of employing higher order n-grams.

weighting schemes that are employed to emphasize relative importance of words in documents. In this work, we use two fundamental approaches: using term frequency – inverse document frequency and binary occurrence (i.e., 1 if a given word occurs in document, 0 otherwise).

The obvious drawback of applying Vector Space Model to readability has to do with the bag–of-words model that the Vector Space Model is an instance of. In a bag of words, the order of the words is not preserved (syntactic information is also lost). Thus, what we do is basically compare documents on a purely lexical level. Nevertheless, lexicon is an important factor in measuring readability of a given text.

3.5. System Architecture

Our readability application called *Jasnopis* has been written using the Django framework[5] integrated with Celery task manager[6].

The main readability page has been designed using the Django framework. The system accepts three types of input sources: plain text, uploaded file and URL; the type of input can be selected in the application part of the readability web page. All entered data is saved on the readability server though it can be used later for building corpora and can be helpful in bug tracking. Plain text is saved as a text file, a page indicated by the URL is saved as HTML file. After saving the file, a process request is added to the database and then overtaken by the task manager. The task manager used for the purposes of the readability project is Celery with RabbitMQ task broker[7]. Owing to Celery and RabbitMQ asynchronous task mechanism, the readability page can handle more users at a time, though it is more expandable.

Requests are divided into three main groups: "index", "graph" and "taylor". The "index" group represents counting readability indices (two types of FOG index and four types of Pisarek's index). The "graph" group represents requests for drawing graphs of document similarity to easy and difficult texts. Difficult texts are represented by legal acts while easy ones by children's literature. Graph drawing uses word frequency lists for counting input document similarity with representative corpora. For each of those corpora, word frequency files are also built. At the moment, we use corpora representing children's literature, legal acts, press articles, popular science and Wikipedia.

For each type of request, different tasks are queued. For every type of request input data should be first of all prepared for tagging. To do this, the application invokes a conversion task. The conversion task uses a couple of standalone programs to convert input text data to a special premorph format accepted by the WCRFT tagger (Radziszewski, 2013). Depending on the input format, different conversion programs will be used. For URLs and HTML files the program uses justext python API[8] (Pomikálek, 2011) which strips HTML of boilerplate content, such as navigation links, headers and footers. So the created premorph file contains only main web page content. For Open Office file types (for example: `.odt`, `.doc`, `.rtf`) the converter uses unoconv program linked by UNO library[9] to headless Open Office application listening on a specified port. For other types of input files, the tool uses Tika[10] to change the file to text format and then pass it to the unoconv program. In both cases, unoconv creates Open Office syntax-based XML, which finally would be translated into premorph format. Presented approach allows uploading almost any kind of input file format to the application. The advantage of using Open Office-based formats is that during the conversion, the topics, bullets and enumeration will be recognized, in some cases the request result more accurate.

After the conversion task is done, the request with the premorph file will be passed on to the next task. For index measuring, it would be the calculation task, for graph drawing, the word frequency measure task and for "taylor" requests, creating a properly prepared model file and a file with unknown words. In each case, the premorph file will be tagged by a WCRFT tagger.

For the word frequency count task, the product of tagging would be used later for counting most frequently used words in input text. Word frequency is only counted for verbs, nouns and adjectives. The frequency list is saved later as a CSV file containing a list of words in descending order of frequency. Words are similarly when they have same base form and part of speech tag.

The calculation task uses text tagged by the WCRFT tagger for counting text statistics. Text statistics are used to calculate the type of readability index selected in the interface. This is the last task for "index" type requests, after it successfully ends, the readability index with text statistics will be presented in the application interface.

After word frequency is calculated, the file containing frequency statistics will be passed to the measure similarity task. It uses the word frequency list to estimate how similar the input text is to difficult texts and to easy ones. Similarity measuring can be done using one of two algorithms: tf.idf (term frequency – inversed document frequency) and binary (see previous section for details). After the task is done, a proper graph will be draw in the application interface using the jqPlot library[11].

For "taylor"-type requests, the next task is called the similarity search task. The similarity search task uses ngram language models created for each of the reference corpora to find the most suitable one for input text. For creating language models measuring similarity, the readability web page uses SRI Language Modelling Toolkit (Stolcke, 2002). Its standard functions support the creation of language models, perplexity and probability counting, as well as a gap filling mechanism. At the moment, "taylor"-type request are still in the beta version, nevertheless, finding a proper language model by perplexity and hit count mechanisms gives satisfying results. "Taylor"-type request use

[5] `https://www.djangoproject.com/`

[6] `http://celeryproject.org/`

[7] `http://www.rabbitmq.com/`

[8] `http://code.google.com/p/justext/`

[9] `http://dag.wieers.com/home-made/unoconv/` – unoconv file converter web page; `http://www.openoffice.org/udk/` – UNO library for Open Office web page.

[10] `http://tika.apache.org/`

[11] `http://www.jqplot.com/`

the same corpora for building language models as are used for drawing graphs.

When all tasks are done, the request status will be marked as "successful", if something goes wrong, it will be marked as "failure". Each task has its different status, every half a second the client side asks the database for actual request status (current task), and if one has been changed, it will also change it in the readability web page interface.

Each request covers information about actual request status, type of processing (which include information about request types), request start time, request end time, it also covers document statistics for index-type request and name of the most similar language model for "taylor"-type requests.

3.6. Experiments

In the first experiment, we wanted to check whether different readability indices (and their variants) are correlated with each other. The purpose of this experiment is twofold. First, we wanted to see if there is a real impact of morphosyntactic processing of a Slavic language — Polish – on the value of a readability index. Second, if the correlation is significant, we can simplify the user interface of Jasnopis and display only one index.

Correlation was counted as Spearman's rank correlation (Spearman, 1904). Texts for readability calculations were taken from the manually annotated 1-million-word subcorpus of NKJP (Przepiórkowski et al., 2012).

Every implemented readability index shows high correlation with all others. Correlation values can be seen in the Table 1. On one hand, the results support our initial intuition – that both Pisarek and FOG indices should be highly correlated as they employ similar features (percentage of difficult words, average sentence length), but use different formulas. On the other hand, we are surprised that inflection does not have as much impact as envisaged.

	Spearman's Correlation
$P(n,o)$ to $P(l,o)$	0.99
$P(n,b)$ to $FOG(b)$	0.96
$P(n,o)$ to $FOG(o)$	0.96
$P(l,b)$ to $FOG(b)$	0.97
$P(l,o)$ to $FOG(o)$	0.97
$FOG(o)$ to $FOG(b)$	0.94

Table 1: Spearman's rank correlation for different types of readability indices where: $P(n,o)$ – nonlinear Pisarek's index based on original lemma form; $P(l,o)$ – linear Pisarek's index based on original lemma form; $P(n,b)$ – nonlinear Pisarek's index based on base lemma form; $P(l,b)$ – linear Pisarek's index based on base lemma form; $FOG(o)$ – FOG index based on original lemma form; $FOG(b)$ – FOG index based on base lemma form.

For the next experiment with similarity based calculation and automated Taylor test, we first measured the FOG index for different sections of our reference corpora. The average amount of years of education was counted as a sum of FOG indices for each document in a selected corpus divided by the number of documents in it.

	FOG-Orth average value
Children's literature	7 years
Law acts	12 years
Press	12 years
Wikipedia	13 years
Popular-science (KPWr)	14 years
Law acts (KPWr)	14 years

Table 2: Average amount of years of education required to understand texts from selected corpora.

To evaluate both the similarity method and the automated Taylor method, we used a widely known protocol for evaluating machine learning algorithms called cross validation. We applied a leave-one-out cross validation, i.e., we took out one document from the corpora, and trained[12] the algorithms on the remaining documents. The results for the similarity method are presented in Table 3 and for automated Taylor test in Table 4.

	Binary	**tf.idf**
Children's literature	100.00%	100.00%
Wikipedia	85.37%	85.37%
Law acts	100.00%	100.00%
Press	71.74%	73.91%
Popular science	100.00%	100.00%

Table 3: Percentage of properly assigned documents to their corpora in Leave-one-out cross-validation using similarity algorithm

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

Table 4: Percentage of properly assigned documents to their corpora in Leave-one-out cross-validation using Taylor-based algorithm

Data in Tables 2 and 3 was collected using the following corpora:

1. small corpus of children's literature consisting of 33 texts (38 255 words);

2. small Wikipedia corpus consisting of 41 texts (34 280 words) selected from the Polish Wikipedia Corpus[13];

[12] In similarity-based method training can be seen as processing all the documents and converting them to the vector space model.

[13] http://clip.ipipan.waw.pl/

3. small press article corpus consisting of 46 articles (34 650 words) selected from Korpus "Rzeczpospolitej"[14];

4. small legal act corpus consisting of 24 legal acts (33 963 words).

For the purpose of counting average FOG index (Table 2), we used the Polish Corpus of Wrocław University of Technology (Broda et al., 2012): popular science texts for "popular-science" and law texts for "legal acts (KPWr)" (Table 2). For Table 3, as "popular science", we used a sample of 39 articles (33 476 words) from "Wiedza i życie" archives[15].

Data for Table 4 was collected using the following corpora:

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 Korpus "Rzeczpospolitej";

4. legal act corpus consisting of 175 legal acts (172 627 words);

5. popular science corpus consisting of 183 texts (183 088 words) from "Wiedza i Życie" archives.

The results of both the similarity method and automated Taylor test are interesting. We have achieved very high accuracy for both the most difficult and the easiest texts (legal acts and literature for children). Lower accuracy for Wikipedia can be interpreted as the result of many different kinds of texts in the corpus. They contain both more readable, better edited texts and less readable and more difficult articles (like math-related articles). Low scores for press articles are quite surprising, as the texts are of similar genre and style.

4. Conclusions and Future Work

The current version of the system provides the first computer implementation of Pisarek's and Taylor's method for evaluating text comprehensibility for Polish and shows considerable improvements compared to competing systems. Consequent steps of its development will concentrate on improving Taylor-based language models by collecting larger and more representative reference corpora. Furthermore, they will be cross-validated to eliminate non-representative texts and balanced in size (at the moment, we have collected a satisfying number of texts from the legal domain and the Polish Wikipedia; press registers are planned to be included in the next version of the system). Another branch of development will concentrate on enhancing presentation, e.g., pointing out features responsible for text incomprehensibility ("difficult" words, overly complex sentences).

We also believe that the readability scale needs elaboration; psycholinguistic methods are being included into this process. Moreover, Taylor method is being confronted with empirical data. Our current research also concentrates on adding other stylistic-statistic indicators, such as vocabulary richness, text subjectivity or egotism to the list of readability features.

It seems that Pisarek's formula (like Flesch's original test) relies on approximating and simplifying features of the text. In the age of fast computers and large text corpora, we may (and, in fact, should) determine how often specific lexical items appear in Polish texts, and how often specific features of the evaluated text (e.g., types of syntactic constructions, such as the passive voice or participle clauses) appear in comparison with the reference corpus (or a corpus of specifically selected works). We shall also attempt to involve Bayesian probability — see e.g.: (Imiołczyk, 1987), rank-ordered lists of technical vocabulary (Cygal-Krupa, 1986), rank-ordered lists of common vocabulary (Markowski, 1990) and other frequency lists.

In cooperation with psycholinguists, we shall test the extent to which particular features of a text translate into its readability. We shall rely on various methods of evaluating text comprehension, in particular cloze deletion tests and questions regarding content of the text (as in foreign language comprehension tests), among others.

We also intend to determine whether readability of a text may depend upon readers, and in particular, their linguistic and cultural competences. We shall attempt to create a tool for measuring readability of works aimed at teenagers, based on the contents of reading lists and textbooks for students of different grade levels, and for measuring readability of technical (e.g., scientific) documents, based on impressions of readers with expertise in the field (making use of corpora representing specific areas of knowledge).

We shall also attempt to design our tools in a fashion which would make our measurements compatible with analogical tools used for measuring readability of English text. This would allow for comparing the readability of source text and its translations (e.g., when translating EU documents).

From the technical point of view, providing an online system for measuring readability of a given text and suggesting improvements is already a big help. It might be convenient when working on an already finished text. On the other hand, copying and pasting a text to a website is not the most convenient option when writing the text, especially from scratch. Thus, we have implemented plugins for the most common text editors, namely Microsoft Word and `OpenOffice.org`[16], to help authors during all the stages of writing a text.

5. References

Bormuth, J. R. (1966). Readability: A new approach. *Reading research quarterly*, pages 79–132.

Broda, B., Marcińczuk, M., Maziarz, M., Radziszewski, A., and Wardyński, A. (2012). KPWr: Towards a Free Corpus of Polish. In Calzolari, N., Choukri, K., Declerck, T.,

`PolishWikipediaCorpus`

[14] `http://www.cs.put.poznan.pl/dweiss/research/rzeczpospolita/`

[15] `http://archiwum.wiz.pl/`

[16] Plugin for `OpenOffice.org` is still under development during writing the paper.

Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of LREC'12*, Istanbul, Turkey. ELRA.

Choldin, M. T. (1979). Rubakin, nikolai aleksandrovič. In Kent, A., Lancour, H., and Nasri, W. Z., editors, *Encyclopedia of Library and Information Science: Volume 26*, pages 178–79. CRC Press.

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

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

Flesch, R. F. (1943). *Marks of Readable Style: A Study in Adult Education*. Number 897 in Contributions to education. Teachers College, Columbia University.

Flesch, R. F. (1960). *How to write, speak, and think more effectively*. Harper – Row Publishers, New York.

Fry, E. (1968). A readability formula that saves time. *Journal of reading*, 11(7):513–578.

Goldman, S. R., Graesser, A. C., and van den Broek, P., editors. (1999). *Narrative Comprehension, Causality, and Coherence*. Erlbaum, Mahwah, NJ.

Gunning, R. (1971). *Technique of Clear Writing*. McGraw-Hill.

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

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

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, CNTECHTRA Research Branch Report 8-75.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

Markowski, A. (1990). *Leksyka wspólna różnym odmianom polszczyzny*. Wydawnictwo Wiedza o Kulturze, Warszawa.

Miles, T. H. (1990). *Critical thinking and writing for science and technology*. Harcourt Brace Jovanovich.

Mistrík, J. (1968). Meranie zrozumitel'nosti prehocoru. *Slovenská reč*, 33:171–178.

Pisarek, W. (1969). Jak mierzyć zrozumiałość tekstu. *Zeszyty Prasoznawcze*, 4(42):35–48.

Pisarek, W. (2002). *Nowa retoryka dziennikarska*. MIT Press, Cambridge, MA, USA.

Pisarek, W. (2007). Jak mierzyć zrozumiałość tekstu. In *O mediach i języku*, pages 245–262. Universitas, Kraków.

Pomikálek, J. (2011). *Removing Boilerplate and Duplicate Content from Web Corpora*. Ph.D. thesis, Univesity of Brno, Czech Republic.

Przepiórkowski, A., Bańko, M., Górski, R. L., and Lewandowska-Tomaszczyk, B., editors. (2012). *Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish]*. Wydawnictwo Naukowe PWN, Warsaw.

Radziszewski, A. (2013). A tiered CRF tagger for Polish. In Bembenik, R., Skonieczny, Ł., Rybiński, H., Kryszkiewicz, M., and Niezgódka, M., editors, *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*. Springer Verlag.

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

Spearman, C. (1904). The proof and measurement of association between two things. *The American journal of psychology*, 15:72–101.

Stolcke, A. (2002). Srilm – an extensible language modeling toolkit. In *Interspeech*, pages 901–904.

Tanaka-Ishii, K., Tezuka, S., and Terada, H. (2010). Sorting texts by readability. *Computational Linguistics*, 36(2):203–227.

Taylor, M. C. and Wahlstrom, M. W. (1986). Readability as applied to an abe assessment instrument. *International Journal for Basic Education*, 10(3)(42):155–170.

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