1.01 UDC: 070.18:329.052 (497.5)"1945/1990"

Tadej Škvorc,[*] Simon Krek,[**] Senja Pollak,[***]
Špela Arhar Holdt,[****] Marko Robnik-Šikonja[*****]

Predicting Slovene Text Complexity Using Readability Measures

IZVLEČEK

NAPOVEDOVANJE KOMPLEKSNOSTI SLOVENSKIH BESEDIL
Z UPORABO MER BERLJIVOSTI

Večina obstoječih formul za merjenje berljivosti je zasnovana za besedila v angleškem jeziku, na katerih je tudi ocenjena njihova kakovost. V našem članku predstavimo prilagoditev izbranih mer za slovenščino. Uspešnost desetih znanih formul ter osmih dodatnih kriterijev berljivosti ocenimo na petih skupinah besedil: otroških revijah, splošnih revijah, časopisih, tehničnih revijah in zapisnikih sej državnega zbora. Te skupine besedil imajo različne ciljne publike, zaradi česar predpostavimo, da uporabljajo različne stile pisanja, ki bi jih formule in kriteriji berljivosti morali zaznati. V analizi pokažemo, katere formule in kriteriji berljivosti delujejo dobro in s katerimi razlik med skupinami nismo mogli zaznati.

Ključne besede: berljivost, obdelava naravnega jezika, analiza besedil

[*] **University of Ljubljana, Faculty of Computer and Information Science, Večna Pot 113, SI-1000 Ljubljana, Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, tadej.skvorc@fri.uni-lj.si**

[**] **Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, University of Ljubljana, Faculty of Arts, Aškerčeva 2, SI-1000 Ljubljana, simon.krek@guest.arnes.si**

[***] **Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, senja.pollak@ijs.si**

[****] **University of Ljubljana, Faculty of Arts, Aškerčeva 2, SI-1000 Ljubljana, University of Ljubljana, Faculty of Computer and Information Science, Večna Pot 113, SI-1000 Ljubljana, spela.arharholdt@ff.uni-lj.si**

[*****] **University of Ljubljana, Faculty of Computer and Information Science, Večna Pot 113, SI-1000 Ljubljana, marko.robnik@fri.uni-lj.si**

ABSTRACT

The majority of existing readability measures are designed for English texts. We aim to adapt and test the readability measures on Slovene. We test ten well-known readability formulas and eight additional readability criteria on five types of texts: children's magazines, general magazines, daily newspapers, technical magazines, and transcriptions of national assembly sessions. As these groups of texts target different audiences, we assume that the differences in writing styles should be reflected in their readability scores. Our analysis shows which readability measures perform well on this task and which fail to distinguish between the groups.

Keywords: readability, natural language processing, text analysis

Introduction

In English, the problem of determining text readability (i.e. how easy a text is to understand) has long been a topic of research, with its origins in the 19th century (Sherman 1893). Since then, many different methods and readability measures have been developed, often with the goal of determining whether a text is too difficult for its target age group. Even though the question of readability is complex from a linguistic standpoint, a large majority of existing measures are based on simple heuristics. There has been little research on readability of languages other than English, therefore we aim to apply these measures to Slovene and evaluate how well they perform.

There are several factors that might cause these measures to perform poorly on non-English languages, such as:

- Many measures are fine-tuned to correspond to the grade levels of the United States education system. It is likely a different fine-tuning would be needed for other languages, as a.) their education system is different from the US system, and b.) the differences in readability between grade levels are likely to be different between languages, meaning that each language would require specifically tuned parameters.
- Some measures utilize a list of common English words and their results depend on the definition of this list. For Slovene, there currently does not exist a publicly available list of common words, so it is not known how such measures would perform.
- The existing readability measures do not use the morphological information to determine difficult words but rely on syllable and character counts, or a list of difficult words. As Slovene is morphologically much more complex than English, words with complex morphology are harder to understand than those with simple morphology, even if they have the same number of characters or syllables.

We analyze the commonly used readability measures (as well as some novel measures) on Slovene texts and propose a word list needed to implement the word-list-based measures. We calculate statistical distributions of scores for each readability measure across subcorpora and assess the ability of measures to distinguish between different subcorpora using a variety of statistical tests. We show that machine learning classification models, using a combination of readability measures, can predict the subcorpus a given text belongs to.

The paper extends the short version of the paper presented in Škvorc et al. (2018) and is structured as follows. We first present the related work on readability measures and describe the readability measures used in our analysis. The methodology of the analysis is presented next, followed by the results split into three sections. The last section concludes the paper and presents ideas for further work.

Related Work

For English, there exists a variety of works focused on determining readability by using readability formulas. Those formulas rely on different features of the text such as the average sentence length, percentage of difficult words, and the average number of characters per word. Examples of such measures include the Coleman-Liau index (Coleman and Liau 1975), LIX (Björnsson 1968), and the automated readability index (ARI) (Senter and Smith 1967). Some formulas, like the Flesch-Kincaid grade level (Kincaid et al. 1975) and SMOG (Mc Laughlin 1969) use the number of syllables per word to determine if a word is difficult. Additionally, some measures (e.g., the Spache readability formula (Spache 1953) and Dale-Chall readability formula (Dale and Chall 1948) rely on a pre-constructed list of difficult words.

Aside from the readability formulas, there exists a variety of other approaches that can be used to determine readability (Bailin and Grafstein 2016). For example, various machine-learning approaches can be used to obtain better results than readability formulas, such as the approach presented in Francois and Miltsakaki (2012), which outperforms readability formulas on French text.

There is little work attempting to apply these measures to Slovene texts. Most work dealing with the readability of Slovene text is focused on manual methods. For example, Justin (2009) analyzes Slovene textbooks from a variety of angles, including readability. On the other hand, works that focus on automatic readability measures are rare. Zwitter Vitez (2014) uses a variety of readability measures for author recognition in Slovene text, but we found no works that used them to determine readability.

In addition to Slovene, some related works evaluate readability measures on other languages. Debowski et al. (2015) evaluate readability formulas on Polish text and show that they obtain better results by using a more complex, machine-learning-based approach.

Readability Measures

In our analysis, we used two groups of readability measures:

- **Existing readability formulas for English:** we focused mainly on popular methods that have been shown to achieve good results on English texts. These measures mostly rely on easy-to-obtain features such as a number of difficult words, sentence length, and word length.
- **Natural-language-processing-based readability criteria:** we used additional criteria that are not present in the existing readability formulas but can be obtained from tools for automatic language processing, such as the percentage of verbs, number of unique words, and morphological difficulty of words. In the existing English formulas, such criteria are not used but they might contain useful information for determining the readability of Slovene texts.

In the following two subsections we present the established readability measures for grading English text and our proposed additional criteria.

Existing Readability Formulas

There exists a variety of ways to measure the readability of texts written in English. For our analysis, we used 10 readability formulas given below. The entities used in the expressions correspond to the number of occurrences of a given entity, e.g., word corresponds to the number of words in a measured text.

- **Gunning fog index** (Gunning 1952) is calculated as:

$$\mathrm{GFI} = 0.4\frac{\text{words}}{\text{sentences}} + 100\frac{\text{complex words}}{\text{words}},$$

where a word is considered complex if it contains three or more syllables. As there exists no established automatic method for counting syllables of Slovene words, we used a rule-based approach designed for English. The resulting score is calibrated to the grade level of the USA education system.

- **Flesch reading ease** (Kincaid et al. 1975) is calculated as:

$$\mathrm{FRE} = 206.835 - 1.015\frac{\text{words}}{\text{sentences}} - 84.6\frac{\text{syllables}}{\text{words}}.$$

The score does not correspond to grade levels. Instead, the higher the value, the easier the text is considered to be. A text with a score of 100 should be easily understood by 11-year-old students, while a text with a score of 0 should be intended for university graduates.

– **Flesch–Kincaid grade level** (Kincaid et al. 1975) is similar to Flesch reading ease, but does correspond to grade levels. It is calculated as:

$$\mathrm{FKGL} = 0.39\frac{\text{words}}{\text{sentences}} + 11.8\frac{\text{syllables}}{\text{words}} - 15.59\,.$$

– **Dale–Chall readability formula** (Dale and Chall 1948) is calculated as:

$$\mathrm{DCRF} = 0.1579\frac{\text{difficult words}}{\text{words}} + 0.0496\frac{\text{words}}{\text{sentences}}\,.$$

The formula requires a predefined list of common (easy) words and the words which are not on the list are considered as difficult. The novelty of the Dale-Chall Formula was that it did not use word-length counts but a count of "hard" words which do not appear on a specially designed list of common words. This list was defined as the words familiar to most of the 4th-grade students: when 80 percent of the fourth--graders indicated that they knew a word, the word was added to the list.
Higher scores indicate that the text is harder, but the resulting score does not correspond to grade levels, nor is it appropriate for text aimed at children below 4th grade. In our analysis, we obtained the difficult words in two ways:

1. By constructing a list of "easy" words and considering every word not on the list as difficult. The list of easy words is described later in the paper.
2. By considering words with more than seven characters as difficult.

– **Spache readability formula** (Spache 1953) is calculated as:

$$\mathrm{SRF} = 0.141\frac{\text{words}}{\text{sentences}} + 8.6\frac{\text{unique difficult words}}{\text{unique words}} + 0.839.$$

Difficult words are defined as words that do not appear in the list of commonly used words, which is the same as the one used in the Dale–Chall readability formula. This method was specifically designed for texts targeting children up to the fourth grade and was not designed to perform well on harder text. The obtained score corresponds to grade levels.

– **Automated readability index** (Senter and Smith 1967) is calculated as:

$$\mathrm{ARI} = 4.71\frac{\text{characters}}{\text{words}} + 0.5\frac{\text{words}}{\text{sentences}} - 21.43.$$

The formula was designed so that it could be automatically captured in times when texts were written on typewriters and therefore it does not use information relating to syllables or difficult words. The obtained score corresponds to grade levels.

– **SMOG (Simple Measure of Gobbledygook)** (McLaughlin 1969) can be calculated as:

$$\mathrm{SMOG} = 1.043\sqrt{\text{difficult words}\frac{30}{\text{sentences}}} + 3.1291,$$

where difficult words are defined as words with three or more syllables. The score corresponds to grade levels.

– **LIX** (Bjornsson 1968) is calculated as:

$$\mathrm{LIX} = \frac{\text{words}}{\text{sentences}} + 100\frac{\text{long words}}{\text{words}},$$

where long words are defined as words consisting of more than six characters. LIX is the only measure we used that was not designed specifically for English but for a variety of languages. Because of this, it does not use syllables or a list of unique words. The score does not correspond to grade levels.

– **RIX** (Anderson 1983) is a simplification of LIX, and is calculated as:

$$\mathrm{RIX} = \frac{\text{long words}}{\text{sentences}}.$$

– **Coleman-Liau index** (Coleman and Liau 1975) is calculated as:

$$\mathrm{CLI} = 0.0588\mathrm{L} - 0.296\mathrm{S} - 15.8,$$

where L is the average number of letters per 100 words and S is the average number of sentences per 100 words. The obtained score corresponds to grade levels.

Language-Processing-Based Readability Criteria

The readability formulas described in the previous section use a low number of common criteria, such as the number of syllables in words or the number of words in a sentence. In our analysis, we also analyzed Slovene texts using the following additional statistics:

– percentage of long words,
– percentage of difficult words,
– percentage of verbs,
– percentage of adjectives,
– percentage of unique words,
– average sentence length.

Many of these (percentage of long words, difficult words, unique words, and average sentence length) are used as features in the readability measures described above. We evaluate them individually to determine how important each of them is for Slovene texts. The **percentage of verbs** is used because a higher number of verbs can indicate more complex sentences with multiple clauses. The **percentage of adjectives** was chosen because we assumed a higher percentage of adjectives could indicate longer, more descriptive sentences that are harder to understand.

To take into account richer morphology of Slovene and a less fixed word order compared to English, we computed two additional criteria:

– **Context of difficult words**, which is the average number of difficult words that appear in a context (i.e. the three words before or after the word) of a difficult word. Difficult words are defined as words that do not appear on the list of common words. The intuition behind this metric is that a difficult word that appears in the context of easy words is easier to understand than if it is surrounded by other difficult words since its meaning can be more easily inferred from the context.
– **Average morphological difficulty**, where we use the Slovene morphological lexicon Sloleks (Arhar Holdt 2009) to assign a "morphological difficulty" score to each word. Sloleks is a lexicon of word forms and contains frequency information for morphological variants of over 100,000 lemmas (base forms of words as defined in a dictionary). We use the relative frequency of a word variant compared to other variants of the same lemma as the morphological difficulty score.

In addition, we also calculated the number of words in each document, even if in our case, it cannot be interpreted as a criterion for determining readability since it is largely determined by the type of document. E.g., the documents belonging to the subcorpus of newspapers contain individual articles and are therefore short, while the subcorpus of computer magazines contains entire magazines which are considerably longer.

Analysis of Slovene Texts

In this section, we describe the methodology used for our analysis. In the first subsection, we describe the data sets on which we conducted our analysis. In the second subsection, we describe how we constructed the list of easy words used in some of the readability measures.

Data Sets

We created a set of subcorpora from the Gigafida reference corpus of written Slovene (Logar et al. 2012). Gigafida contains 39,427 Slovene texts released from 1990 to 2011, for a total of 1,187,002,502 words. We focused on texts published in magazines, newspapers, and books while ignoring texts collected from the internet. The texts in the Gigafida corpus are segmented into paragraphs and sentences, tokenized, and part-of-speech tagged using the Obeliks tagger (Grčar et al. 2012). We grouped the texts based on the intended audience, resulting in the following subcorpora:
– **Children's magazines** include magazines aimed at younger children (to be read independently or by their parents), namely Cicido and Ciciban.
– **Pop magazines** contain magazines aimed at the general public, namely Lisa, Gloss, and Stop.

– **Newspapers** contain general adult population newspapers, namely Delo and Dolenjski list.
– **Computer magazines** include magazines focusing on technical topics relating to computers, namely Monitor, Računalniške novice, PC & Mediji, and Moj Mikro.
– **National Assembly** includes transcriptions of sessions from the National Assembly of Slovenia.

In Table 1 we show the number of documents in each subcorpus and the average number of words per document. The subcorpus of newspapers contains the largest number of documents, while the subcorpus of text sourced from the National Assembly of Slovenia contains the fewest.

Table 1: The number of documents and the average number of words per document for each subcorpus.

Subcorpus	#docs	Avg. #words / doc
Children's magazines	125	5,488
Pop magazines	247	33,967
Newspapers	14,011	12,881
Computer magazines	163	110,875
National Assembly	35	58,841

Our hypothesis is that the readability measures will be able to distinguish texts from different subcorpora. We assume that children's magazines will be easily distinguishable from other genres that are addressing an adult population. We also suppose that general magazines are less complex than specialized magazines. The National Assembly transcripts were included as they differ from other texts in two major ways: a.) they are transcripts of spoken language and b.) they relate to a highly technical subject matter. Because of this, we were interested in how readability measures would grade them. To test our hypothesis and to determine how well each readability measure works, we analyzed texts from each subcorpus to obtain a score distribution for each measure. The scores were calculated separately for each source text (e.g., one magazine article, a newspaper, or one assembly session).

List of Common Words

For designing the list of common words, we took a corpus-based approach. Note that the methodology to create a list of common words from language corpora was already tested for other languages, (see e.g., Kilgarriff et al. 2014). We used four corpora to create a list of common words: Kres, Janes, Gos, and Šolar:

– **Šolar** (Kosem et al. 2011) contains 2,703 texts written by pupils in Slovenia from grades 6 to 13 (grade 6 to 9 in primary school, and grade 1 to 4 in secondary school).

The texts include essays, summaries, and answers to examination questions.

- **Gos** (Verdonik et al. 2011) contains around 120 hours of recorded spoken Slovene (1,035,101 words), as well as transcriptions of the recordings. The recordings are collected from a variety of sources, including conversations, television, radio, and phone calls. Around 10% of the corpus consists of recorded lessons in primary and secondary schools.
- **Janes** (Fišer et al. 2014) contains Slovene texts from various internet sources, such as tweets, forum posts, blogs, comments, and Wikipedia talk pages.
- **Kres** (Logar Berginc and Šuster 2009) is a sub-corpus of Gigafida that is balanced with respect to the source (e.g. newspapers, magazines, or internet).

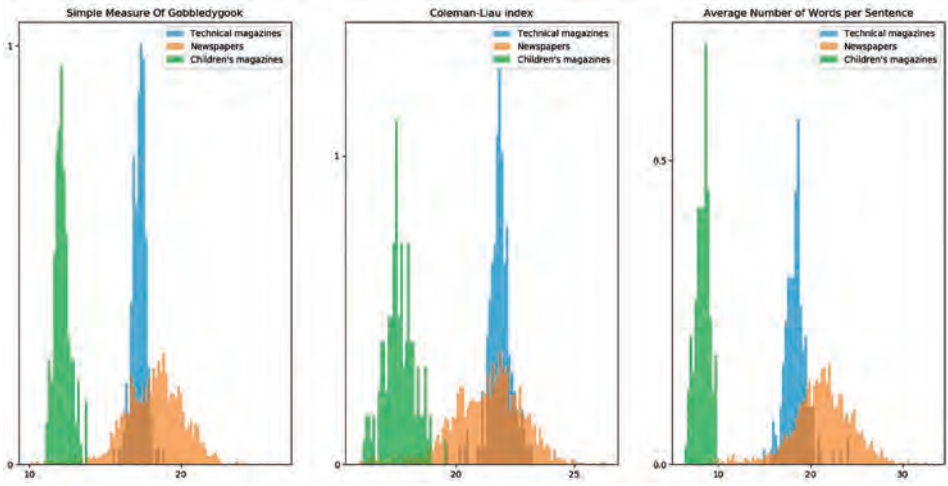
We extracted the most common words and defined the common words as the ones that appear frequently in all four corpora (and are therefore not specific to a certain text type). We use four corpora to include texts that primarily reflect language production by different language users (Gos, Janes, Šolar), as well as texts that primarily reflect standard language (Kres). We aimed at covering younger school-going population (Šolar) and adults. For some corpora, we could have assigned words to different age levels (e.g. using pupils' grade levels in Šolar or using the age groups available in Gos metadata), but these corpora are very specific and the resulting word groups would mainly reflect the genre instead of age levels. Because of this, we opted for the approach of crossing the word lists to obtain a single list. The overlap of the most common words in four corpora eliminates frequent words which are typical for only one of the corpora (e.g. administrative language in Kres, spoken language markers in Gos, Twitter-specific usage in Janes, and literary references from essays in Šolar).

From each corpus, we extracted the 10,000 most frequent word lemmas and part-of-speech tuples. In order to construct a list of common words representative of Slovene language, we selected the word lemmas that occurred in the most frequent word lists of all the four corpora. We obtained a list of 2,562 common words, which we used in readability measures.

Results

For each text in each subcorpus, we calculated readability scores using all readability measures described in the previous section. In Figure 1 we present a few examples of obtained score distributions. We show distributions for three text subcorpora (children's magazines, newspapers, and technical magazines) and three readability scores (Goobledygook, Coleman-Liau, and the average number of words in a sentence).

Figure 1: The score distributions for three text subcorpora and three readability measures. The distributions show that technical magazines readability scores are the most consistent, while newspapers' scores are more diverse. Children's magazines' scores have a strong peak on the left-hand side (easier texts) that is well separated from the other sources.



To show a compact overview of all included readability measures we calculated the median, first and third quartiles of the distribution for each score and each text subcorpus. The box-and-whiskers plots showing these results are visualized in Figure 2 which shows that most readability measures are able to distinguish between different subcorpora. Additionally, some of the readability measures confirm our original hypothesis, i.e. they are able to distinguish children's magazines from other genres that are addressing adult population, and evaluate general magazines as less complex than computer magazines.

Figure 2: The scores of each readability measure for each subcorpus of texts, represented with box plots. The subcorpora depicted from left to right are: 1.) Children's magazines, 2.) General magazines, 3.) Newspapers, 4.) Computer magazines, and 5.) National assembly transcriptions. The boxes show the first, second, and third quartile of the distributions while the whiskers extend for 1.5 IQR past the first and third quartile.

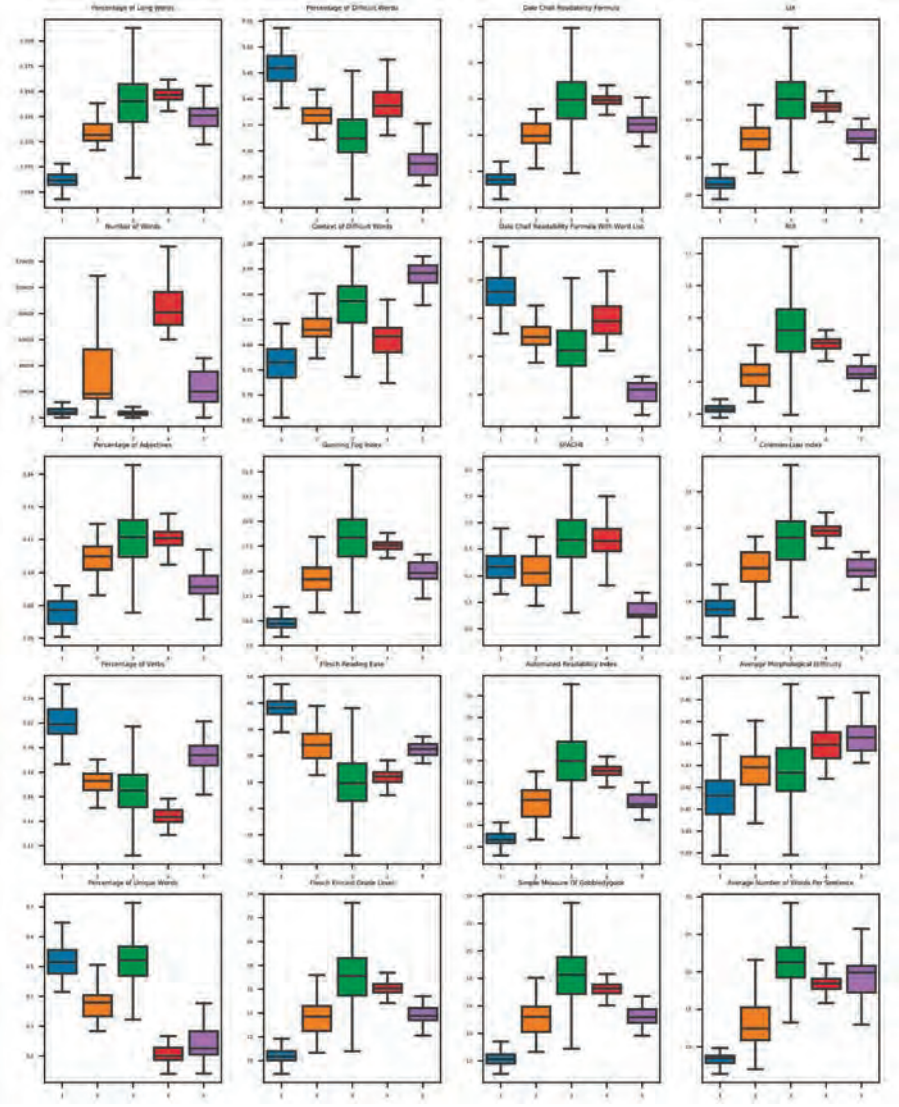


Figure 2 allows for an additional interpretation of readability measures. For example, children's magazines vs. general magazines vs. newspapers mean scores show increasing complexity in the following measures: Percentage of long words, Flesh Kincaid Grade Level, Gunning Fog Index, Dale-Chall Readability Formula (based on complexity defined by syllables), Context of Difficult Words, SMOG, LIX, RIX

and Automated Readability Index. All these measures consider the length of words and/or sentences. The percentage of adjectives also seems to correlate with the complexity of these three text types, although to a lesser extent. The same holds for Flesh Reading Ease, since higher scores indicate lower complexity. For the majority of these measures, the distinction between newspapers and specialized computer magazines is either less evident or not evident at all, but they do indicate that computer magazines are less readable than general magazines.

Scores using the list of common words do not lead to the same conclusions. Percentage of Difficult Words and Dale-Chall Readability Formula with word list do not reflect the complexity of genres, but to some extent, they do distinguish between general and specialized texts (i.e. newspapers and general magazines have lower scores than specialized computer magazines). One of the reasons for the relatively high scores for the complexity of children magazines might be in the large proportion of literary language, such as in poems for children with many words not in the list of common words. For example, "KRAH, KRAH, KRAH! MENE NIČ NI STRAH!" (Krah, krah, krah! I am not afraid!) has 7 words, out of which 4 are on the list of simple words, while the interjection KRAH is not on the simple words list. Therefore, the proportion of difficult words in this segment is 42.8% (3 occurrences of word KRAH out of 7 words in total). On the other hand, the words are short, therefore length-based measures consider them to be simple words.

The readability scores for the National Assembly subcorpus show high variability across the measures, which might be attributed to the fact that it is a different genre (spoken, but specialized). E.g., in several measures where the readability complexity rises from children's magazines to general magazines and newspapers, the National assembly scores are close to general magazines. Very long words are less likely used in spoken language, even in a political context. Average morphological difficulty and context of difficult words lead to the interpretation that this genre is more complex (less "readable"). The very high score for the context of difficult words might be attributed to enumeration of Assembly members (e.g., "Obveščen sem, da so zadržani in se današnje seje ne morejo udeležiti naslednje poslanke in poslanci: Ciril Pucko, Franc Kangler, Vincencij Demšar, Branko Kalalemina, …" (I was informed that the following deputies are occupied and cannot attend this session: …). The relatively high percentage of verbs can also be interpreted from this perspective, e.g., the National assembly text include many performatives, such as "Pričenjam nadaljevanje seje" (Starting the continuation of the session) and "Ugotavljamo prisotnost v dvorani" (Establishing the presence).

In summary, using a list of common words did not improve the partitioning of the text subcorpora perceived as easy and as difficult to read. Both measures that use it (Dale-Chall and Spache readability formulas) are poor separators. A number of simple readability measures worked well, such as the percentage of long words, the percentage of verbs/adjectives, and the average morphological difficulty.

We also calculated the sample mean and standard deviation of readability measures for each text subcorpus. The results are shown in Table 2.

Table 2: The mean and standard deviation for each subcorpus of texts and each readability score.

Measure	Children's mag.	Magazines	Newspapers	Technical mag.	National assembly
% long words	0.065 (0.015)	0.109 (0.011)	0.137 (0.029)	0.146 (0.010)	0.137 (0.046)
Number of words	5488 (6184)	33966 (34821)	12881 (84708)	110875 (151007)	58841 (106515)
% adjectives	0.078 (0.016)	0.111 (0.013)	0.120 (0.020)	0.120 (0.008)	0.096 (0.022)
% verbs	0.216 (0.026)	0.170 (0.015)	0.161 (0.034)	0.144 (0.013)	0.180 (0.044)
% unique words	0.517 (0.077)	0.375 (0.053)	0.513 (0.114)	0.244 (0.144)	0.277 (0.173)
Context of difficult words	0.756 (0.054)	0.834 (0.027)	0.849 (0.133)	0.808 (0.036)	0.929 (0.044)
% difficult words	0.464 (0.048)	0.369 (0.022)	0.356 (0.122)	0.389 (0.032)	0.280 (0.036)
Gunning Fog Index	9.950 (1.255)	14.272 (1.271)	18.662 (9.319)	17.470 (0.800)	15.901 (3.493)
Flesch reading ease	37.592 (4.989)	23.855 (5.217)	10.002 (24.128)	12.520 (4.340)	19.178 (13.098)
Flesch–Kincaid grade level	10.500 (0.894)	13.596 (1.193)	17.356 (8.959)	15.999 (0.741)	14.523 (2.761)
Dale–Chall	2.845 (0.425)	4.036 (0.306)	4.972 (1.270)	4.941 (0.258)	4.560 (0.971)
Dale–Chall with word list	7.781 (0.720)	6.534 (0.357)	6.643 (2.163)	6.955 (0.484)	5.208 (0.539)
Spache readability formula	6.217 (0.368)	6.079 (0.348)	6.977 (3.499)	6.685 (0.323)	5.482 (0.600)
Automated readability index	12.873 (1.086)	16.117 (1.428)	20.474 (11.456)	19.007 (0.885)	17.014 (3.371)
SMOG	12.206 (0.759)	15.095 (1.066)	18.200 (2.757)	17.194 (0.611)	15.849 (2.500)
LIX	33.676 (3.384)	44.999 (3.282)	56.016 (23.123)	53.260 (2.077)	47.909 (9.073)
RIX	2.381 (0.496)	4.481 (0.781)	7.370 (3.836)	6.354 (0.518)	5.250 (2.574)
Coleman-Liau index	17.785 (1.120)	19.823 (0.861)	21.220 (1.807)	21.762 (0.903)	20.318 (2.170)
Avg. morphological difficulty	0.419 (0.017)	0.428 (0.010)	0.436 (0.044)	0.441 (0.017)	0.445 (0.026)
Avg. sentence length	8.353 (0.820)	13.389 (2.843)	21.120 (4.043)	18.641 (1.960)	19.063 (3.826)

Using these results, we calculated the Bhattacharyya distance between the distributions of Children's magazines and newspapers for each score. The Bhattacharyya distance measures the similarity between two statistical distributions. We assumed the scores were distributed normally, as the results shown in Figure 1 show that the scores approximately follow a normal distribution, and calculated the distance using the following formula:

$$D_B(p,q) = \frac{1}{4}\ln\left[\frac{1}{4}\left(\frac{\sigma_p^2}{\sigma_q^2}+\frac{\sigma_q^2}{\sigma_p^2}+2\right)\right]+\frac{1}{4}\left(\frac{(\mu_p-\mu_q)^2}{\sigma_p^2+\sigma_q^2}\right)$$

We also show the Bhattacharyya coefficient, which measures the overlap between two statistical distributions and can be calculated as:

$$BC(p,q) = e^{(-D_B(p,q))}$$

The results are presented in Table 3. These results are similar to the ones shown in Figure 2, with the readability formulas using the list of difficult words showing less dichotomization power. The largest distance is obtained using average sentence lengths.

Table 3: The Bhattacharyya distances and coefficients between the distributions of scores for children's magazines and newspapers for each readability measure. The results are sorted by decreasing distance.

Measure	Distance	Coefficient
Average sentence length	2.866	0.057
SMOG	1.433	0.239
% long words	1.350	0.259
RIX	1.101	0.333
Flesch-Kincaid grade level	0.956	0.385
Automated readability index	0.945	0.389
Dale-Chall readability formula	0.885	0.413
Gunning fog index	0.880	0.415
LIX	0.853	0.426
Spache readability formula	0.797	0.451
Flesch reading ease	0.776	0.460
% adjectives	0.719	0.487
Coleman-Liau index	0.708	0.493
% verbs	0.432	0.649
% difficult words	0.365	0.694
Dale-Chall with word list	0.318	0.728
Context of difficult words	0.285	0.752
Avg. morphological difficulty	0.235	0.790
% unique words	0.039	0.961

Additional Statistical Tests

In addition to the initial analysis presented in the previous section, we performed additional, more thorough statistical tests to determine which of the evaluated measures are better at predicting the group a text belongs to. We used the following approaches:

– **Mutual information.** This measure reports the amount of information we get about a random variable *Y* by observing another random variable *X*. In our case, mutual information reports the amount of information we get about the group of texts by knowing a score of certain readability measure. Mutual information is defined as:

$$\sum_{y \in Y} \sum_{x \in X} p(x,y) log \left(\frac{p(x,y)}{p(x)p(y)} \right),$$

where p(x) and p(y) are the marginal probability distribution functions of *X* and *Y* and p(x, y) is the joint probability function of *X* and *Y*. In our case, X represents the distributions of readability measures and Y the distribution of groups. The higher the mutual information between the readability measure and the groups, the more useful the measure for determining the group membership.

– **Analysis of variance (ANOVA).** This measure first splits samples of a statistical distribution into several groups (in our case, based on the group the texts belong to) and then calculates if the groups are significantly different from one another. We use this measure to determine if the distributions obtained by calculating a single measure on each group of texts are significantly different. If they are, they can be useful for determining the group membership of a given text.

– **Feature selection using a chi-squared test.** Similarly to mutual information, we use the chi-squared test to determine whether the readability measures and the group memberships are mutually dependent. If they are, this indicates that knowing the value of the readability measure is useful when determining which group a text belongs to.

In addition to the four statistical tests used above, we also ranked each feature using a random forest classifier (Breiman 2001). The classifier is capable of automatically combining different readability measures in order to predict which subcorpus a given text belongs to and is also capable of calculating how important each readability measure was when making the prediction. The classifier is described in more detail in the next section. Using each of these tests, we obtained scores that tell us how useful each readability measure is when trying to predict the subcorpus it came from. The results are presented in Table 4, with higher scores indicating better (more informative) readability measures.

Table 4: The ranks of readability measures obtained by the statistical tests, which report the usefulness of readability measures for predicting group membership. The measures are ordered from the most useful to the least useful.

Random Forest	ANOVA	Mutual information	Chi2
Average sentence length	Average sentence length	Average sentence length	% new words
% new words	% difficult words SPG	RIX	Number of words
Number of words	% long words	SMOG	% unique words
% unique words	SMOG	Percentage of new words	Flesch reading ease
% difficult words SPG	Dale-Chall	Automated readability index	LIX
Gunning fog index	Percentage of adjectives	Gunning fog index	Average sentence length
Percentage of verbs	Coleman-Liau index	LIX	% difficult words
RIX	Percentage of unique words	Number of words	Gunning fog index
Dale-Chall (word list)	RIX	Flesch-Kincaid grade level	Automated readability index
SMOG	% verbs	Flesch reading ease	% difficult words SPG
LIX	Flesch reading ease	Dale-Chall	Flesch-Kincaid grade level
Flesch-Kincaid grade level	Context of difficult words	% unique words	SMOG
Context of difficult words	LIX	% long words	RIX
Dale-Chall	Gunning fog index	% difficult words	Coleman-Liau index
% long words	Flesch-Kincaid grade level	% difficult words SPG	Dale-Chall
% difficult words	% difficult words	Spache readability formula	Spache readability formula
Avg morphological difficulty	Automated readability index	Context of difficult words	Dale-Chall (word list)
Automated readability index	% new words	Coleman-Liau index	% long words
% adjectives	Number of words	% verbs	Context of difficult words
Flesch reading ease	Dale-Chall (word list)	% adjectives	% verbs
Spache readability formula	Spache readability formula	Dale-Chall (word list)	% adjectives
Coleman-Liau index	Avg morphological difficulty	Avg morphological difficulty	Avg morphological difficulty

The results of the statistical tests show that the features commonly used by the readability formulas (i.e. an average sentence length and number of long words) are

useful when it comes to determining group membership. In particular, the average sentence length stands out since it is ranked as the most important measure in three out of the four tests. At least one of either LIX or RIX is also highly ranked (in the top 50% of all measures) by all the tests. Those measures are the only ones from the tested measures that were not designed specifically for English, which could be one of the reasons why they perform better on Slovene texts. The results also show that a number of proposed simpler readability criteria, such as the percentage of verbs, percentage of adjectives, and the average morphological difficulty are less useful than the established statistical formulas. The results are inconclusive about the most useful readability criterion for Slovene. Several formulas and statistics are useful, but the rankings are different by different tests. When using our list of common words Dale-Chall and Spache readability formulas are again shown to perform worse than the formulas that consider long words as difficult.

Classification Results

In addition to statistical evaluation, we also performed a test with machine learning classifiers (Kononenko and Kukar 2007) to see whether we could use our readability measures to predict which subcorpus a text belongs to. With classification models, we can automatically learn how to split the texts into different subcorpora based on readability formulas and other readability criteria. We used the following classification models.

- **Decision trees** construct a binary decision tree where each node splits the training set based on one readability measure. The trained tree can predict the subcorpus of a given text.
- **Random forests (Breiman 2001)** create multiple decision trees in a random manner. This reduces the variance of a model and often gives better prediction accuracy than using a single decision tree.
- **Naive Bayes** is a probabilistic model based on the Bayes' theorem. The model assumes that the readability measures are independent.
- **Extreme gradient boosting (Chen and Carlos 2016)** constructs a large number of simple classifiers and combines them to achieve state-of-the-art results on many classification problems.

In order to use classification models, we first train them on a training subset of our data set. We used randomly selected 75% of our data set for the training. To evaluate the models, we calculated the classification accuracy (i.e. the percentage of texts each model predicted correctly) on the remaining 25% of the data set. The obtained results are presented in Table 5. The results obtained by the majority classifier (i.e. classifying everything as the most frequent group) are presented as a baseline score.

Table 5: The classification accuracies for each of the models. The numbers show the percentage of texts for which the group membership was correctly predicted.

Model	Classification Accuracy
Random Forest	0.984
Extreme Gradient Boosting	0.979
Decision Tree	0.960
Majority Classifier	0.791
Naive Bayes	0.553

Table 5 shows that we are able to predict the correct group of a text with high accuracy, over 98% with the best-performing model (Random forest). This shows that a combination of readability measures that we evaluated in this paper can be used to accurately distinguish between different groups of text.

Conclusion and Future Work

We analyzed statistical distributions of well-known readability measures on Slovene texts. We extracted five subcorpora of texts from the Gigafida corpus with commonly perceived different readability levels: children magazines, popular magazines, newspapers, technical magazines, and national assembly texts. We find that the readability formulas are able to distinguish between these subcorpora reasonably well, with the exception of national assembly texts, which are of a different, spoken, genre and the used measures were not originally designed to handle it. A number of simple readability statistics, such as the context of difficult words and average sentence length, also dichotomize the different subcorpora of text.

In this work, we only focused on simple readability formulas along with some additional readability criteria. There exist several more complex methods for evaluating the complexity of texts, such as the one presented in Lu (2009) and Wiersma et al. (2010). Such advanced methods might be more suitable for Slovene texts than the simple methods used in this paper, and we plan to test them in future work.

Most of the used English readability formulas were designed to correlate with school grades and were initially tuned on that domain. For Slovene, there currently is no publicly available data set with texts tagged according to the appropriate grade level. This disallows analysis of the readability measures from this perspective. In future work, we plan to prepare such a corpus and design several readability scores fit for different purposes. This will allow us to frame text complexity as a classification problem with the goal of predicting the grade level of a text instead of predicting its group membership. In a similar approach, experts would annotate texts with readability scores. This would allow us to fit a regression model using the readability measures analyzed in this paper.

Another area that we plan to explore is the use of coherence and cohesion measures (Barzilay and Lapata 2008; Crossley et al. 2016), which are used to determine if words, sentences, and paragraphs are logically connected. Coherence and cohesion methods usually use machine learning approaches that mostly rely on language-specific features and shall be therefore evaluated on Slovene texts. The same applies to readability measures based on machine learning (Francois and Miltsakaki 2012) which we also plan to analyze in the future.

Acknowledgments

The research was financially supported by the Slovenian Research Agency through project J6-8256 (New grammar of contemporary standard Slovene: sources and methods), project J5-7387 (Influence of formal and informal corporate communications on capital markets), a young researcher grant, research core fundings no. P6-0411 and P2-0103; Republic of Slovenia, Ministry of Education, Science and Sport/European social fund/European fund for regional development/European cohesion fund (project Quality of Slovene textbooks, KaUč). This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153 (EMBEDDIA).

Sources and Literature

Literature:

- Anderson, Jonathan. 1983. "LIX and RIX: Variations on a Little-known Readability Index." *Journal of Reading* 26, No. 6: 490–96.
- Arhar Holdt, Špela. 2009. "Učni korpus SSJ in leksikon besednih oblik za slovenščino." *Jezik in slovstvo* 54, No. 3–4: 43–56.
- Bailin, Alan, and Ann Grafstein. 2016. *Readability: Text and context*. Springer.
- Barzilay, Regina, and Mirella Lapata. 2008. "Modeling Local Coherence: An Entity-based Approach." *Computational Linguistics* 34, No. 1: 1–34.
- Björnsson, Carl Hugo. 1968. *Läsbarhet*. Liber.
- Breiman, Leo. 2001. "Random forests." *Machine learning* 45, No. 1: 5–32.
- Chen, Tianqi, and Carlos Guestrin. 2016. "Xgboost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. ACM.
- Coleman, Meri, and Ta Lin Liau. 1975. "A Computer Readability Formula Designed for Machine Scoring." *Journal of Applied Psychology* 60, No. 2: 283.
- Crossley, Scott A., Kristopher Kyle, and Danielle S. McNamara. 2016. "The tool for the automatic analysis of text cohesion (TAACO): Automatic Assessment of Local, Global, and Text Cohesion." *Behavior Research Methods* 48, No. 4: 1227–37.
- Dale, Edgar, and Jeanne S. Chall. 1948. "A Formula for Predicting Readability: Instructions." *Educational Research Bulletin*: 37–54.

- Dębowski, Łukasz, Bartosz Broda, Bartłomiej Nitoń, and Edyta Charzyńska. 2015. "Jasnopis–A Program to Compute Readability of Texts in Polish Based on Psycholinguistic Research." In *Natural Language Processing and Cognitive Science*, edited by B. Sharp, W Lubaszewski and R. Delmonte, 51–61. Liberia Editrice Cafoscarina.
- Fišer, Darja, Tomaž Erjavec, Ana Zwitter Vitez, and Nikola Ljubešić. 2014. "JANES se predstavi: metode, orodja in viri za nestandardno pisno spletno slovenščino." In *Language technologies : proceedings of the 17th International Multiconference Information Society - IS 2014*, edited by Tomaž Erjavec and Jerneja Žganec Gros, 56–61. Ljubljana: Jožef Stefan Institute.
- François, Thomas, and Eleni Miltsakaki. 2012. "Do NLP and Machine Learning Improve Traditional Readability Formulas?" In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, edited by Sandra Williams, Advaith Siddharthan and Ani Nenkova, 49–57. Association for Computational Linguistics.
- Grčar, Miha, Simon Krek, and Kaja Dobrovoljc. 2012. "Obeliks: statisticni oblikoskladenjski oznacevalnik in lematizator za slovenski jezik." In *Proceedings of the Eighth Language Technologies Conference,* edited by Tomaž Erjavec and Jerneja Žganec Gros, 89–94. Ljubljana: Jožef Stefan Institute.
- Gunning, Robert. 1952. *The technique of clear writing*. McGraw-Hill.
- Justin, J. 2003. *Učbenik kot dejavnik uspešnosti kurikularne prenove: poročilo o rezultatih evalvacijske študije.*
- Kilgarriff, Adam, Frieda Charalabopoulou, Maria Gavrilidou, Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis, Robert Lew, Serge Sharoff, Ravikiran Vadlapudi, and Elena Volodina. 2014. "Corpus-based Vocabulary Lists for Language Learners for Nine Languages." *Language Resources and Evaluation* 48, No. 1: 121–63.
- Kincaid, J. Peter, Robert P. Fishburne Jr, Richard L. Rogers, and Brad S. Chissom. 1975. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease formula) for navy enlisted personnel*. Report No. 8–75.
- Kononenko, Igor, and Matjaž Kukar. 2007. *Machine Learning and Data Mining*. Chichester, Horwood Publishing.
- Kosem, Iztok, Tadeja Rozman, and Mojca Stritar. 2011. "How do Slovenian Primary and Secondary School Students Write and What Their Teachers Correct: A Corpus of Student Writing." In *Proceedings of Corpus Linguistics Conference 2011, ICC Birmingham*, 20–22.
- Logar Berginc, Nataša, and Simon Šuster. 2009. "Gradnja novega korpusa slovenščine." *Jezik in slovstvo* 54: 57–68.
- Logar Berginc, Nataša, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt, Simon Krek, and Iztok Kosem. 2012. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko and Faculty of Social Sciences.
- Lu, Xiaofei. 2009. "Automatic Measurement of Syntactic Complexity in Child Language Acquisition." *International Journal of Corpus Linguistics* 14, No. 1: 3–28.
- Mc Laughlin, G. Harry. 1969. "SMOG Grading - a New Readability Formula." *Journal of Reading* 12, No. 8: 639–46.
- Senter, R. J., and Edgar A. Smith. 1967. *Automated Readability Index*. Ohio; University of Cincinnati.
- Sherman, Lucius Adelno. 1893. *Analytics of Literature: A Manual for the Objective Study of English Prose and Poetry*. Boston: Ginn.
- Škvorc, Tadej, Simon Krek, Senja Pollak, Špela Arhar Holdt, and Marko Robnik-Šikonja. 2018. "Evaluation of Statistical Readability Measures on Slovene Texts." In *Proceedings of the conference on Language Technologies & Digital Humanities 2018,* edited by Darja Fišer and Andrej Pančur, 240–47. Ljubljana: Ljubljana University Press, Faculty of Arts.
- Spache, George. 1953. "A New Readability Formula for Primary-grade Reading Materials." *The Elementary School Journal* 53, No. 7: 410–13.

- Verdonik, Darinka, Ana Zwitter Vitez, and Hotimir Tivadar. 2011. *Slovenski govorni korpus Gos*. Trojina, zavod za uporabno slovenistiko.
- Wiersma, Wybo, John Nerbonne, and Timo Lauttamus. 2010. "Automatically Extracting Typical Syntactic Differences from Corpora." *Literary and Linguistic Computing* 26, No. 1: 107–24.
- Zwitter Vitez, Ana. 2014. "Ugotavljanje avtorstva besedil: primer »Trenirkarjev«." In *zbornik Devete konference Jezikovne Tehnologije Informacijska družba – IS*, edited by Tomaž Erjavec and Jerneja Žganec Gros, 131–34. Ljubljana: Jožef Stefan Institute.

Tadej Škvorc, Simon Krek, Senja Pollak, Špela Arhar Holdt, Marko Robnik-Šikonja

PREDICTING SLOVENE TEXT COMPLEXITY USING READABILITY MEASURES

SUMMARY

In English, the problem of determining text readability (i.e. how easy a text is to understand) has long been a topic of research, with its origins in the 19th century. Since then, many different methods and readability measures have been developed, often with the goal of determining whether a text is too difficult for its target age group. Even though the question of readability is complex from a linguistic standpoint, a large majority of existing measures are based on simple heuristics. Since most of these measures were developed for English texts, it is hard to say how well they would perform on Slovene texts. Measures designed for English are designed to correspond with the American school system, are sometimes based on pre-constructed lists of easy words which do not exist for Slovene and do not take into account morphological information when determining whether a word is difficult or not.

In our work, we analyze some common readability measures on Slovene text. We also introduce and analyze two additional readability criteria that do not appear in any of the analyzed readability measures: **morphological difficulty**, where we assume word forms that appear rarely are harder to understand than the ones that appear commonly and the **context of difficult words**, where we assume difficult words are easier to understand in a context of simple words, as their meaning can be inferred from that context. We performed the analysis on 14,581 text documents from the Gigafida corpus, which were split into five groups based on their target audience (childrens' magazines, pop magazines, newspaper articles, computer magazines, and transcriptions of sessions of the National Assembly). We assumed that the groups should have different readability scores due to their differing target audiences and writing styles.

For each analyzed readability measure we checked how well it separates texts from different groups. We did this by first obtaining the statistical distribution of readability

scores for texts in each group and checking how much the distributions differ. We show that a number of common readability measures designed for English work well on Slovene texts. To determine which of the measures perform the best we used several statistical tests.

We also show that machine-learning methods can be used to accurately (over 98% chance of a correct prediction) predict which group a text belongs to based on its readability scores. We trained four different machine-learning models (decision trees, random forests, naïve Bayes classifier, and extreme gradient boosting) and evaluated them on our dataset. We obtained the best result (98.4% classification accuracy) by using random forests.

Tadej Škvorc, Simon Krek, Senja Pollak, Špela Arhar Holdt, Marko Robnik-Šikonja

NAPOVEDOVANJE KOMPLEKSNOSTI SLOVENSKIH BESEDIL Z UPORABO MER BERLJIVOSTI

POVZETEK

Problem berljivosti (t.j. kako enostavno je besedilo za branje) je v angleščini dobro raziskan. Obstaja veliko različnih metod in formul, s katerimi lahko analiziramo angleška besedila z vidika berljivosti. Kljub temu, da je vprašanje berljivosti z lingvističnega vidika zapleteno večina metod za ugotavljanje berljivosti temelji na preprostih značilnostih besedil. Ker je bila večina mer berljivosti zasnovanih za angleška besedila, ne moremo biti prepričani da bodo enako dobro delovala na slovenskih besedilih. Angleške mere berljivosti so namreč usklajene z ameriškim šolskim sistemom, včasih temeljijo na vnaprej sestavljenih seznamih lahkih besed in ne upoštevajo težavnosti besed z morfološkega vidika.

V našem delu analiziramo pogoste mere berljivosti na slovenskih besedilih. Poleg tega uvedemo in analiziramo dva dodatna kazalnika berljivosti ki ne nastopata v pogostih merah berljivosti: **morfološka zahtevnost besed**, s katero želimo zajeti predpostavko da so redkejše morfološke oblike besed težko berljive, in **kontekst težkih besed**, s katero želimo zajeti predpostavko, da so neznane besede, ki se pojavijo v kontekstu znanih besed lažje berljive, saj lahko njihov pomen razberemo iz konteksta. Analizo smo izvedli na 14,581 besedilih iz korpusa Gigafida, ki smo jih razdelili v pet skupin glede na njihovo ciljno publiko (Otroške revije, splošne revije, časopisni članki, računalniške revije in transkripcije sej Državnega zbora). Predpostavili smo, da imajo revije zaradi različnih ciljnih publik in tematik različne sloge pisanja in posledično različne stopnje berljivosti.

Za vsako izmed mer berljivosti smo preverili, kako dobro med seboj loči besedila iz različnih skupin. Za vsako izmed njih smo pridobili statistično distribucijo vrednosti berljivosti vsake skupine in preverili, ali so distribucije ustrezno ločene. V analizi pokažemo, da se številne uveljavljene mere, ki so bile zasnovane za angleščino, dobro obnesejo tudi na slovenskih besedilih. Da bi ugotovili, katere mere najbolje razlikujejo med skupinami smo uporabili statistične teste.

Poleg tega pokažemo, da lahko z modeli strojnega učenja in kombinacijo analiziranih metod berljivosti z visoko točnostjo (nad 98 %) napovemo, v katero skupino spada določeno besedilo. Za to analizo smo uporabili štiri različne metode strojnega učenja (odločitvena drevesa, naključne gozdove, naivni Bayesov klasifikator, in extreme gradient boosting). Najboljši rezultat (98,4 %) smo dobili z metodo naključnih gozdov.