Evaluation of Statistical Readability Measures on Slovene texts

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

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

[◦] University of Ljubljana, Faculty of Arts
Aškerčeva 2, SI-1000 Ljubljana
spela.arharholdt@ff.uni-lj.si

[†] Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana
simon.krek@guest.arnes.si senja.pollak@ijs.si

Abstract

The majority of existing readability measures are explicitly designed for and tested on English texts. The aim of our paper is to adapt and test the readability measures on Slovene. We test a set of 10 well-known readability formulas and 8 additional readability criteria on different types of texts: children's magazines, general magazines, daily newspapers, technical magazines, and transcriptions from the national assembly. As these groups of texts target different audiences, we assume that the differences in writing styles should also be reflected in different readability scores. Our analysis shows which readability measures perform well on this task and which fail to distinguish between the groups.

1. 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 it's 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. Nevertheless, it makes sense to apply these measures to Slovene and evaluate how well they perform, since there has been little work dedicated to this question.

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 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 a more complex morphology are likely harder to understand than those with a simple morphology, even if they have the same number of characters or syllables.

These are only a few of the reasons explaining why it is hard to evaluate the performance of the original measures on other languages. In this paper, we analyze the commonly used readability measures (as well as some novel measures) on Slovene texts and propose a word list needed for implementing 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.

The paper is structured as follows. In Section 2. we present the related work on readability measures. In Section 3. we describe the readability measures used in our analysis. The methodology of the analysis is presented in Section 4. The results are contained in Section 5. and Section 6. concludes the paper.

2. 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 average sentence length, percentage of difficult words, and the average number of characters per word. Examples of such measures are given in Section 3. and 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 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 François and Miltsakaki (2012) which outperforms readability formulas on French text.

To the best of our knowledge, there is little existing work that attempts to apply these measures to Slovene texts. Most work dealing with readability of Slovene text is focused on manual methods. For example, Justin (2009) analyzes Slovene textbooks from a variety of angles, including readability. 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 work evaluates readability measures on other languages. Debowski et al. (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.

3. 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 number of difficult words, sentence length and word length).

Additional readability criteria: we used additional criteria that are not present in the existing readability formulas, such as the percentage of verbs, number of unique words, and morphological difficulty of words. In English formulas, such criteria are not used, but they might contain useful information for readability of Slovene texts.

In this section, we present these two groups of readability measures. In Section 3.1. we present the established readability measures for grading English text and in Section 3.2. we present the additional criteria.

3.1. Existing Readability Formulas

There exists a variety of ways to measure 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:

$$\text{GFI} = 0.4\left(\frac{\text{words}}{\text{sentences}} + 100\frac{\text{complex words}}{\text{words}}\right),$$

where a word is considered complex if it contains three or more syllables [1]. The resulting score is calibrated to the grade level of the USA education system.

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

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

The score does not correspond to grade levels. Instead, the higher the value is 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:

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

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

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

The formula requires a predefined list of common (easy) words and the words which are not on the list are considered as difficult. The originality of the Dale-Chall Formula was that it did not use word-length counts but uses a count of 'hard' words, which are the words that 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 in Section 4.2..
2. By considering words with more than seven characters as difficult.

Spache readability formula (Spache, 1953) is calculated as:

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

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.

[1] As there exists no established automatic method for counting syllables of Slovene words, we used a rule-based approach designed for English.

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

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

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

SMOG (Simple Measure of Gobbledygook) (Mc Laughlin, 1969) can be calculated as:

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

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

LIX (Björnsson, 1968) is calculated as:

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

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:

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

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

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

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.

3.2. Additional readability criteria

As mentioned in Section 1. the readability formulas mentioned in Section 3.1. are simple and 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.

Most 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 words. The 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 was surrounded by other difficult words.

Average morphological difficulty. To calculate this, we use Sloleks (Arhar Holdt, 2009) to assign a morphological richness score to each word. Sloleks contains frequency information for morphological variants of over 100 000 lemmas, and we use the relative frequency of a variant compared to other variants of the same lemma as the morphological difficulty score.

We also collected the number of words in each document. In our case, this was not a useful criterion for determining readability since it was largely determined by the type of document (e.g., the documents belonging to the subcorpus of newspapers contained individual articles and were therefore short, while computer magazines contained the entire magazine and were longer).

4. Analysis of Slovene texts

In this section, we describe the methodology used for our analysis. In Section 4.1. we describe the datasets on which we conducted our analysis and in Section 4.2. we describe how we constructed the list of easy words used in some of the readability measures.

4.1. Datasets

For the analysis we have 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 tokenized, segmented into sentences and paragraphs, and part-of-speech tagged using the Obeliks tagger (Grcar et al., 2012). To determine the performance of readability

measures we grouped them based on the intended audience, obtaining the following subcorpora.

Children's magazines include magazines aimed at younger children (to be read from 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 of 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 contained the largest number of documents, while the subcorpus of text sourced from the National Assembly of Slovenia contained the fewest.

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

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

Our hypothesis is that the readability measures will be able to distinguish texts from different subcorpora. We assume children's magazines will be easily distinguishable from other genres that are addressing 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 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).

4.2. 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). From the corpora Kres, Janes, Gos and Šolar, we extracted the most common words and defined common words as the ones which appear in all four corpora (and are therefore not specific to a certain text type). With four corpora we aimed at an inclusion of corpus texts that primarily reflect language production by different language users (GOS, JANES, Šolar), as well as corpus texts that primarily reflect the language community's every-day language reception (Kres). We aimed at covering younger speakers (e.g. Šolar) and adult production. 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 reflecting 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).

More details on the four corpora used as a source of information for commonly used words, are provided below.

Šolar (Kosem et al., 2011) contains 2703 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 recordings of 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. newspaper, magazine or internet).

From each corpus, we extracted the top 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 2562 common words which we used in readability measures. In this paper, we are using the automatically assembled version of the list as described above. In future work, the list will be linguistically analyzed, refined, and made publicly available for further use.

5. Results

For each text in each subcorpus, we calculated readability scores using all readability measures described in Section 3. 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 average sentence length).

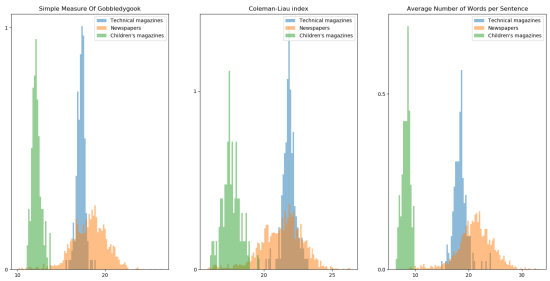

Figure 1: The score distributions for three text subcorpora and three readability measures. The distributions show that technical magazines readability scores are the most consistent, while newspapers' scores are more diverse. Children's magazines' scores have a strong peak on the left-hand side (easier 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 fit 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 also allows for 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 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!" has 7 words, out of which 4 are on the list of simple words, while the word KRAH is not on the simple words list. Therefore the proportion of difficult words in this segment is 42.8% (3 occurences 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 also 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 might be used in spoken language with lower probability, 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 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, ..."). 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" and "Ugotavljamo prisotnost v dvorani".

In summary, using a list of common words described in Section 4.2. did not improve the separation of the text subcorpora perceived as easy and difficult to read. Both measures that use them (Dale-Chall and Spache readability formulas) are poor separators. A number of simple readability measures worked well, such as the percentage of long words, 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.

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 the scores approximately follow a normal distribution, and calculated the distance using the following formula:

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

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

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

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.

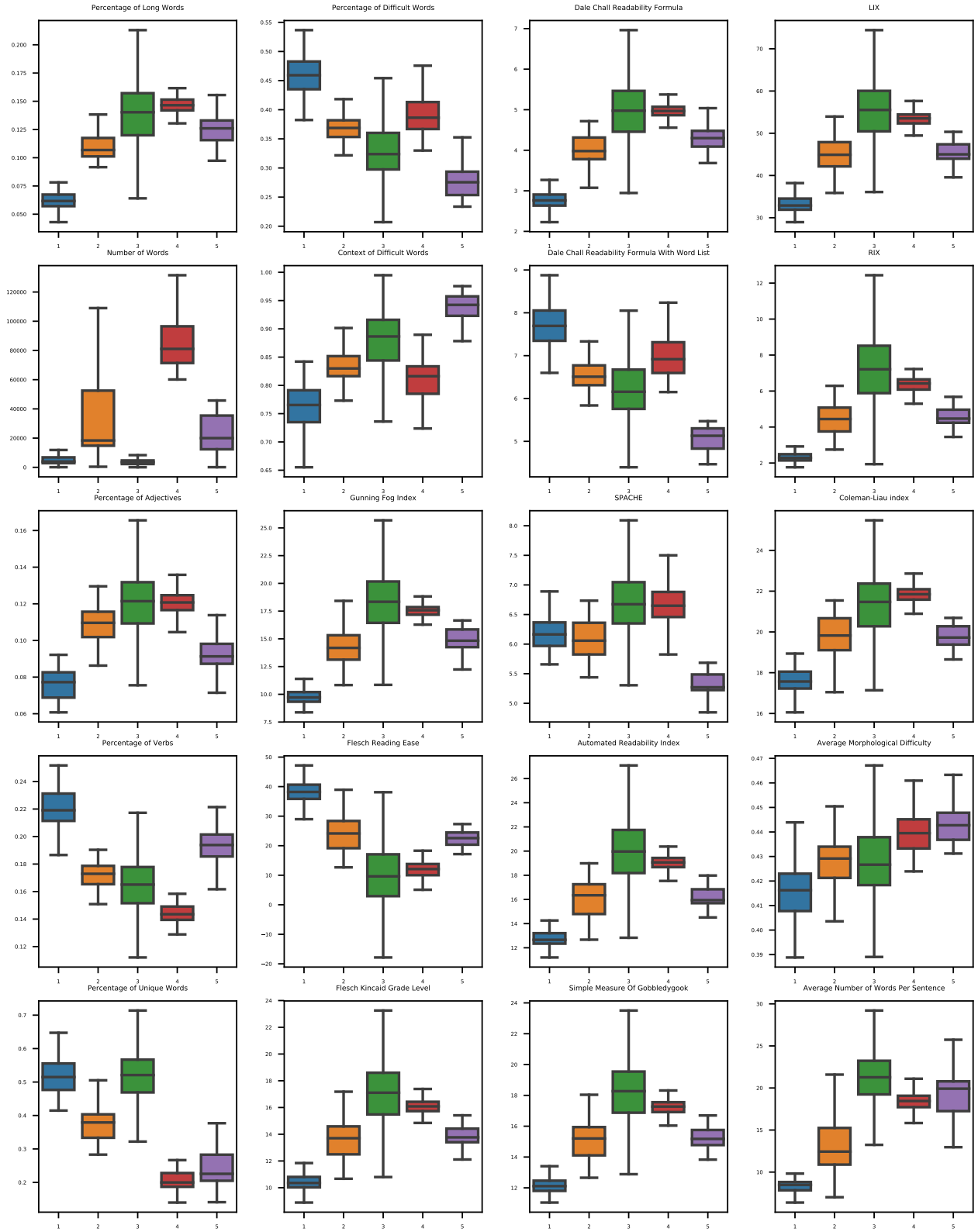

Figure 2: The scores of each readability measure for each subcorpus of texts, represented with box plots. The subcorpora are: 1.) Children's magazines, 2.) General magazines, 3.) Newspapers, 4.) Computer magazines, and 5.) National assembly text. The boxes show the first, second, and third quartile of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

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)

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

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

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.

6. Conclusion and Future work

We analyze statistical distributions of well-known readability measures designed for English on Slovene texts. We extract 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 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 focus on simple readability formulas along with some additional readability criteria. There exists a variety of more complex methods for evaluating the complexity of text, such as the one presented in (Lu, 2009) and (Wiersma et al., 2010). More advanced methods might be more suitable for Slovene texts than the simple methods used in this paper.

Most of the English readability formulas were designed to correlate with school grades and were tested on that domain. For Slovene, there currently does not exist a publicly available dataset where texts are tagged according to the grade level they are appropriate for. This makes analyzing the readability measures from this perspective difficult. In the future work, we plan prepare such a corpus and design several readability scores fit for different purposes. This will also allow us to frame determining readability as a classification problem with the goal of predicting the grade level of a text. A similar approach that is also worth considering would be to have experts annotate texts with readability scores. This would allow us to fit a regression model using the readablilty 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 rely on machine learning approaches that can rely on language specific features and would therefore need to be evaluated on Slovene text. The same applies to readability measures that rely on machine learning (François and Miltsakaki, 2012), which we also plan to analyze in the future.

Acknowledgements

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. P2-0209 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).

7. References

Jonathan Anderson. 1983. Lix and Rix: Variations on a little-known readability index. *Journal of Reading*, 26(6):490–496.

Špela Arhar Holdt. 2009. Učni korpus SSJ in leksikon besednih oblik za slovenščino. *Jezik in slovstvo*, 54(3–4):43–56.

Alan Bailin and Ann Grafstein. 2016. *Readability: Text and context*. Springer.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Carl Hugo Björnsson. 1968. *Läsbarhet*. Liber.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Scott A Crossley, 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(4):1227–1237.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Łukasz Debowski, Bartosz Broda, Bartłomiej Nitoń, and Edyta Charzyńska. 2015. Jasnopis–a program to compute readability of texts in polish based on psycholinguistic research. *Natural Language Processing and Cognitive Science*, page 51.

Darja Fišer, Tomaž Erjavec, Ana Zwitter Vitez, and Nikola Ljubešić. 2014. Janes se predstavi: metode, orodja in viri za nestandardno pisno spletno slovenščino. *V: Jezikovne tehnologije: zbornik*, 17:56–61.

Thomas François 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*, pages 49–57. Association for Computational Linguistics.

Miha Grcar, Simon Krek, and Kaja Dobrovoljc. 2012. Obeliks: statisticni oblikoskladenjski oznacevalnik in lematizator za slovenski jezik. In *Zbornik Osme konference Jezikovne tehnologije, Ljubljana, Slovenia*.

Robert Gunning. 1952. *The technique of clear writing*. McGraw-Hill, New York.

J Justin. 2009. *Učbenik kot dejavnik uspešnosti kurikularne prenove: poročilo o rezultatih evalvacijske študije*. Ljubljana: Pedagoški inštitut.

Adam Kilgarriff, 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(1):121–163.

J Peter Kincaid, 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. Technical report, Naval Technical Training Command Millington TN Research Branch.

Iztok Kosem, Tadeja Rozman, and M Stritar Kučuk. 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*, pages 20–22.

Nataša Logar Berginc and Simon Šuster. 2009. Gradnja novega korpusa slovenščine. *Jezik in slovstvo*, 54(3–4):57–68.

Nataša Logar, 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*. Trojina, zavod za uporabno slovenistiko.

Xiaofei Lu. 2009. Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics*, 14(1):3–28.

G Harry Mc Laughlin. 1969. SMOG grading-a new readability formula. *Journal of reading*, 12(8):639–646.

RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report, University of Cincinnati, Ohio.

Lucius Adelno Sherman. 1893. *Analytics of literature: A manual for the objective study of English prose and poetry*. Ginn, Boston.

George Spache. 1953. A new readability formula for primary-grade reading materials. *The Elementary School Journal*, 53(7):410–413.

Darinka Verdonik, Ana Zwitter Vitez, and Hotimir Tivadar. 2011. *Slovenski govorni korpus Gos*. Trojina, zavod za uporabno slovenistiko.

Wybo Wiersma, John Nerbonne, and Timo Lauttamus. 2010. Automatically extracting typical syntactic differences from corpora. *Literary and Linguistic Computing*, 26(1):107–124.

Ana Zwitter Vitez. 2014. Ugotavljanje avtorstva besedil: primer "trenirkarjev". In *Language technologies: Proceedings of the 17th International Multiconference Information Society – IS 2014*, pages 131–134, Ljubljana, Slovenia, October. Institut Jožef Stefan.