Supervised and Unsupervised Neural Approaches to Text Readability

Matej Martinc
Jožef Stefan Institute, Ljubljana, Slovenia
Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
matej.martinc@ijs.si

Senja Pollak
Jožef Stefan Institute, Ljubljana, Slovenia
senja.pollak@ijs.si

Marko Robnik-Šikonja
University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia
marko.robnik@fri.uni-lj.si

We present a set of novel neural supervised and unsupervised approaches for determining the readability of documents. In the unsupervised setting, we leverage neural language models, whereas in the supervised setting, three different neural classification architectures are tested. We show that the proposed neural unsupervised approach is robust, transferable across languages, and allows adaptation to a specific readability task and data set. By systematic comparison of several neural architectures on a number of benchmark and new labeled readability data sets in two languages, this study also offers a comprehensive analysis of different neural approaches to readability classification. We expose their strengths and weaknesses, compare their performance to current state-of-the-art classification approaches to readability, which in most cases still rely on extensive feature engineering, and propose possibilities for improvements.

1. Introduction

Readability is concerned with the relation between a given text and the cognitive load of a reader to comprehend it. This complex relation is influenced by many factors, such as a degree of lexical and syntactic sophistication, discourse cohesion, and background knowledge (Crossley et al. 2017). In order to simplify the problem of measuring readability, traditional readability formulas focused only on lexical and syntactic features

Submission received: 26 July 2019; revised version received: 22 November 2020; accepted for publication: 18 December 2020.

https://doi.org/10.1162/COLI_a_00398

expressed with statistical measurements, such as word length, sentence length, and word difficulty (Davison and Kantor 1982). These approaches have been criticized because of their reductionism and weak statistical bases (Crossley et al. 2017). Another problem is their objectivity and cultural transferability, since children from different environments master different concepts at different ages. For example, a word *television* is quite long and contains many syllables but is well-known to most young children who live in families with a television.

With the development of novel natural language processing (NLP) techniques, several studies attempted to eliminate deficiencies of traditional readability formulas. These attempts include leveraging high-level textual features for readability modeling, such as semantic and discursive properties of texts. Among them, cohesion and coherence received the most attention, and several readability predictors based on these text features have been proposed (see Section 2). Nevertheless, none of them seems to predict the readability of the text as well as much simpler readability formulas mentioned above (Todirascu et al. 2016).

With the improvements in machine learning, the focus shifted once again, and most newer approaches consider readability as being a classification, regression, or a ranking task. Machine learning approaches build prediction models to predict human assigned readability scores based on several attributes and manually built features that cover as many text dimensions as possible (Schwarm and Ostendorf 2005; Petersen and Ostendorf 2009; Vajjala and Meurers 2012). They generally yield better results than the traditional readability formulas and text cohesion–based methods but require additional external resources, such as labeled readability data sets, which are scarce. Another problem is the transferability of these approaches between different corpora and languages, because the resulting feature sets do not generalize well to different types of texts (Xia, Kochmar, and Briscoe 2016; Filighera, Steuer, and Rensing 2019).

Recently, deep neural networks (Goodfellow, Bengio, and Courville 2016) have shown impressive performance on many language-related tasks. In fact, they have achieved state-of-the-art performance in all semantic tasks where sufficient amounts of data were available (Collobert et al. 2011; Zhang, Zhao, and LeCun 2015). Even though very recently some neural approaches toward readability prediction have been proposed (Nadeem and Ostendorf 2018; Filighera, Steuer, and Rensing 2019), these types of studies are still relatively scarce, and further research is required in order to establish what type of neural architectures are the most appropriate for distinct readability tasks and data sets. Furthermore, language model features designed to measure lexical and semantic properties of text, which can be found in many of the readability studies (Schwarm and Ostendorf 2005; Petersen and Ostendorf 2009; Xia, Kochmar, and Briscoe 2016), are generated with traditional n-gram language models, even though language modeling has been drastically improved with the introduction of neural language models (Mikolov et al. 2011).

The aim of the present study is two-fold. First, we propose a novel approach to readability measurement that takes into account neural language model statistics. This approach is unsupervised and requires no labeled training set but only a collection of texts from the given domain. We demonstrate that the proposed approach is capable of contextualizing the readability because of the trainable nature of neural networks and that it is transferable across different languages. In this scope, we propose a new measure of readability, RSRS (ranked sentence readability score), with good correlation with true readability scores.

Second, we experiment to find how different neural architectures with automatized feature generation can be used for readability classification and compare their

performance to state-of-the-art classification approaches. Three distinct branches of neural architectures—recurrent neural networks (RNN), hierarchical attention networks (HAN), and transfer learning techniques—are tested on four gold standard readability corpora with good results.

The article is structured as follows. Section 2 addresses the related work on readability prediction. Section 3 offers a thorough analysis of data sets used in our experiments, and in Section 4, we present the methodology and results for the proposed unsupervised approach to readability prediction. The methodology and experimental results for the supervised approach are presented in Section 5. We present conclusions and directions for further work in Section 6.

2. Related Work

Approaches to the automated measuring of readability try to find and assess factors that correlate well with human perception of readability. Several indicators, which measure different aspects of readability, have been proposed in the past and are presented in Section 2.1. These measures are used as features in newer approaches, which train machine learning models on texts with human-annotated readability levels so that they can predict readability levels on new unlabeled texts. Approaches that rely on an extensive set of manually engineered features are described in Section 2.2. Finally, Section 2.3 covers the approaches that tackle readability prediction with neural classifiers. Besides tackling the readability as a classification problem, several other supervised statistical approaches for readability prediction have been proposed in the past. They include regression (Sheehan et al. 2010), Support Vector Machine (SVM) ranking (Ma, Fosler-Lussier, and Lofthus 2012), and graph-based methods (Jiang, Xun, and Qi 2015), among many others. We do not cover these methods in the related work because they are not directly related to the proposed approach.

2.1 Readability Features

Classical readability indicators can be roughly divided into five distinct groups: traditional, discourse cohesion, lexico-semantic, syntactic, and language model features. We describe them below.

2.1.1 Traditional Features. Traditionally, readability in texts was measured by statistical readability formulas, which try to construct a simple human-comprehensible formula with a good correlation to what humans perceive as the degree of readability. The simplest of them is average sentence length (ASL), though they take into account various other statistical factors, such as word length and word difficulty. Most of these formulas were originally developed for the English language but are also applicable to other languages with some modifications (Škvorc et al. 2019).

The Gunning fog index (Gunning 1952) (GFI) estimates the years of formal education a person needs to understand the text on the first reading. It is calculated with the following expression:

$$\text{GFI} = 0.4 \left(\frac{totalWords}{totalSentences} + 100 \frac{longWords}{totalSentences} \right)$$

where *longWords* are words longer than 7 characters. Higher values of the index indicate lower readability.

Flesch reading ease (Kincaid et al. 1975) (FRE) assigns higher values to more readable texts. It is calculated in the following way:

$$\text{FRE} = 206.835 - 1.015\left(\frac{totalWords}{totalSentences}\right) - 84.6\left(\frac{totalSyllables}{totalWords}\right)$$

The values returned by the Flesch-Kincaid grade level (Kincaid et al. 1975) (FKGL) correspond to the number of years of education generally required to understand the text for which the formula was calculated. The formula is defined as follows:

$$\text{FKGL} = 0.39\left(\frac{totalWords}{totalSentences}\right) + 11.8\left(\frac{totalSyllables}{totalWords}\right) - 15.59$$

Another readability formula that returns values corresponding to the years of education required to understand the text is the Automated Readability Index (Smith and Senter 1967) (ARI):

$$\text{ARI} = 4.71\left(\frac{totalCharacters}{totalWords}\right) + 0.5\left(\frac{totalWords}{totalSentences}\right) - 21.43$$

The Dale-Chall readability formula (Dale and Chall 1948) (DCRF) requires a list of 3,000 words that fourth-grade US students could reliably understand. Words that do not appear in this list are considered difficult. If the list of words is not available, it is possible to use the GFI approach and consider all the words longer than 7 characters as difficult. The following expression is used in calculation:

$$\text{DCRF} = 0.1579\left(\frac{difficultWords}{totalWords} * 100\right) + 0.0496\left(\frac{totalWords}{totalSentences}\right)$$

The SMOG grade (Simple Measure of Gobbledygook) (McLaughlin 1969) is a readability formula originally used for checking health messages. Similar to FKGL and ARI, it roughly corresponds to the years of education needed to understand the text. It is calculated with the following expression:

$$\text{SMOG} = 1.0430\sqrt{numberOfPolysyllables\frac{30}{totalSentences}} + 3.1291$$

where the *numberOfPolysyllables* is the number of words with three or more syllables.

We are aware of one study that explored the transferability of these formulas across genres (Sheehan, Flor, and Napolitano 2013), and one study that explored transferability across languages (Madrazo Azpiazu and Pera 2020). The study by Sheehan, Flor, and Napolitano (2013) concludes that, mostly due to vocabulary specifics of different genres, traditional readability measures are not appropriate for cross-genre prediction, because they underestimate the complexity levels of literary texts and overestimate that of educational texts. The study by Madrazo Azpiazu and Pera (2020), on the other hand,

concludes that the readability level predictions for translations of the same text are rarely consistent when using these formulas.

All of the above-mentioned readability measures were designed for the specific use on English texts. There are some rare attempts to adapt these formulas to other languages (Kandel and Moles 1958) or to create new formulas that could be used on languages other than English (Anderson 1981).

To show a multilingual potential of our approach, we address two languages in this study, English and Slovenian, a Slavic language with rich morphology and orders of magnitude fewer resources compared to English. For Slovenian, readability studies are scarce. Škvorc et al. (2019) researched how well the above statistical readability formulas work on Slovenian text by trying to categorize text from three distinct sources: children's magazines, newspapers and magazines for adults, and transcriptions of sessions of the National Assembly of Slovenia. Results of this study indicate that formulas that consider the length of words and/or sentences work better than formulas that rely on word lists. They also noticed that simple indicators of readability, such as percentage of adjectives and average sentence length, work quite well for Slovenian. To our knowledge, the only other study that employed readability formulas on Slovenian texts was done by Zwitter Vitez (2014). In that study, the readability formulas were used as features in the author recognition task.

2.1.2 Discourse Cohesion Features. In the literature, we can find at least two distinct notions of discourse cohesion (Todirascu et al. 2016). First is the notion of **coherence**, defined as the "semantic property of discourse, based on the interpretation of each sentence relative to the interpretation of other sentences" (Van Dijk 1977). Previous research that investigates this notion tries to determine whether a text can be interpreted as a coherent message and not just as a collection of unrelated sentences. This can be done by measuring certain observable features of the text, such as the repetition of content words or by analysis of words that explicitly express connectives (*because, consequently, as a result*, etc.) (Sheehan et al. 2014). A somewhat more investigated notion, due to its easier operationalization, is the notion of **cohesion**, defined as "a property of text represented by explicit formal grammatical ties (discourse connectives) and lexical ties that signal how utterances or larger text parts are related to each other."

According to Todirascu et al. (2016), we can divide cohesion features into five distinct classes, outlined below: co-reference and anaphoric chain properties, entity density and entity cohesion features, lexical cohesion measures, and part of speech (POS) tag-based cohesion features. **Co-reference and anaphoric chain properties** were first proposed by Bormuth (1969), who measured various characteristics of anaphora. These features include statistics, such as the average length of reference chains or the proportion of various types of mention (noun phrases, proper names, etc.) in the chain. **Entity density** features include statistics such as the total number of all/unique entities per document, the average number of all/unique entities per sentence, and so forth. These features were first proposed in Feng, Elhadad, and Huenerfauth (2009) and Feng et al. (2010), who followed the theoretical line from Halliday and Hasan (1976) and Williams (2006). **Entity cohesion** features assess relative frequency of possible transitions between syntactic functions played by the same entity in adjacent sentences (Pitler and Nenkova 2008). **Lexical cohesion measures** include features such as the frequency of content word repetition across adjacent sentences (Sheehan et al. 2014), a Latent Semantic Analysis (LSA)-based feature for measuring the similarity of words and passages to each other proposed by Landauer (2011), or a measure called Lexical

Tightness (LT), suggested by Flor, Klebanov, and Sheehan (2013), defined as the mean value of the Positive Normalized Pointwise Mutual Information (PMI) for all pairs of content-word tokens in a text. The last category is **POS tag-based cohesion features**, which measure the ratio of pronoun and article parts-of-speech, two crucial elements of cohesion (Todirascu et al. 2016).

Todirascu et al. (2016), who analyzed 65 discourse features found in the readability literature, concluded that they generally do not contribute much to the predictive power of text readability classifiers when compared with the traditional readability formulas or simple statistics such as sentence length.

2.1.3 Lexico-semantic Features. According to Collins-Thompson (2014), vocabulary knowledge is an important aspect of reading comprehension, and lexico-semantic features measure the difficulty of vocabulary in the text. A common feature is **Type-token ratio (TTR)**, which measures the ratio between the number of unique words and the total number of words in a text. The length of the text influences TTR; therefore, several corrections, which produce a more unbiased representation, such as Root TTR and Corrected TTR, are also used for readability prediction.

Other frequently used features in classification approaches to readability are ***n*-gram lexical features**, such as word and character *n*-grams (Vajjala and Meurers 2012; Xia, Kochmar, and Briscoe 2016). While **POS-based lexical features** measure lexical variation (i.e., TTR of lexical items such as nouns, adjectives, verbs, adverbs, and prepositions) and density (e.g., the percentage of content words and function words), **word list-based features** use external psycholinguistic and Second Language Acquisition (SLA) resources, which contain information about which words and phrases are acquired at the specific age or English learning class.

2.1.4 Syntactic Features. Syntactic features measure the grammatical complexity of the text and can be divided into several categories. **Parse tree features** include features such as an average parse tree height or an average number of noun- or verb-phrases per sentence. **Grammatical relations features** include measures of grammatical relations between constituents in a sentence, such as the longest/average distance in the grammatical relation sets generated by the parser. **Complexity of syntactic unit features** measure the length of a syntactic unit at the sentence, clause (any structure with a subject and a finite verb), and T-unit level (one main clause plus any subordinate clause). Finally, **coordination and subordination features** measure the amount of coordination and subordination in the sentence and include features such as a number of clauses per T-unit or number of coordinate phrases per clause, and so on.

2.1.5 Language Model Features. The standard task of language modeling can be formally defined as predicting a probability distribution of words from the fixed size vocabulary V, for word w_{t+1}, given the historical sequence $w_{1:t} = [w_1, \ldots, w_t]$. To measure its performance, traditionally a metric called perplexity is used. A language model m is evaluated according to how well it predicts a separate test sequence of words $w_{1:N} = [w_1, \ldots, w_N]$. For this case, the perplexity (PPL) of the language model $m()$ is defined as:

$$\text{PPL} = 2^{-\frac{1}{N}\sum_{i=1}^{N}\log_2 m(w_i)} \tag{1}$$

where $m(w_i)$ is the probability assigned to word w_i by the language model m, and N is the length of the sequence. The lower the perplexity score, the better the language model

predicts the words in a document—that is, the more predictable and aligned with the training set the text is.

All past approaches for readability detection that use language modeling leverage older n-gram language models rather than the newer neural language models. Schwarm and Ostendorf (2005) train one n-gram language model for each readability class c in the training data set. For each text document d, they calculate the likelihood ratio according to the following formula:

$$LR(d, c) = \frac{P(d|c)P(c)}{\sum_{\bar{c} \neq c} P(d|\bar{c})P(\bar{c})}$$

where $P(d|c)$ denotes the probability returned by the language model trained on texts labeled with class c, and $P(d|\bar{c})$ denotes probability of d returned by the language model trained on the class $\bar{c}$. Uniform prior probabilities of classes are assumed. The likelihood ratios are used as features in the classification model, along with perplexities achieved by all the models.

In Petersen and Ostendorf (2009), three statistical language models (unigram, bigram and trigram) are trained on four external data resources: Britannica (adult), Britannica Elementary, CNN (adult), and CNN abridged. The resulting 12 n-gram language models are used to calculate perplexities of each target document. It is assumed that low perplexity scores calculated by language models trained on the adult level texts and high perplexity scores of language models trained on the elementary/abridged levels would indicate a high reading level, and high perplexity scores of language models trained on the adult level texts and low perplexity scores of language models trained on the elementary/abridged levels would indicate a low reading level.

Xia, Kochmar, and Briscoe (2016) train 1- to 5-gram word-based language models on the British National Corpus, and 25 POS-based 1- to 5-gram models on the five classes of the WeeBit corpus. Language models' log-likelihood and perplexity scores are used as features for the classifier.

2.2 Classification Approaches Based on Feature Engineering

The above approaches measure readability in an unsupervised way, using the described features. Alternatively, we can predict the level of readability in a supervised way. These approaches usually require extensive feature engineering and also leverage many of the features described earlier.

One of the first classification approaches to readability was proposed by Schwarm and Ostendorf (2005). It relies on a SVM classifier trained on a WeeklyReader corpus,[1] containing articles grouped into four classes according to the age of the target audience. Traditional, syntactic, and language model features are used in the model. This approach was extended and improved upon in Petersen and Ostendorf (2009).

Altogether, 155 traditional, discourse cohesion, lexico-semantic, and syntactic features were used in an approach proposed by Vajjala and Lučić (2018), tested on a recently published OneStopEnglish corpus. Sequential Minimal Optimization (SMO) classifier with the linear kernel achieved the classification accuracy of 78.13% for three readability classes (elementary, intermediate, and advanced reading level).

1 http://www.weeklyreader.com.

A successful classification approach to readability was proposed by Vajjala and Meurers (2012). Their multilayer perceptron classifier is trained on the WeeBit corpus (Vajjala and Meurers 2012) (see Section 3 for more information on WeeBit and other mentioned corpora). The texts were classified into five classes according to the age group they are targeting. For classification, the authors use 46 manually crafted traditional, lexico-semantic, and syntactic features. For the evaluation, they trained the classifier on a train set consisting of 500 documents from each class and tested it on a balanced test set of 625 documents (containing 125 documents per each class). They report 93.3% accuracy on the test set.[2]

Another set of experiments on the WeeBit corpus was conducted by Xia, Kochmar, and Briscoe (2016), who conducted additional cleaning of the corpus because it contained some texts with broken sentences and additional meta-information about the source of the text, such as copyright declaration and links, strongly correlated with the target labels. They use similar lexical, syntactic, and traditional features as Vajjala and Meurers (2012) but add language modeling (see Section 2.1.5 for details) and discourse cohesion-based features. Their SVM classifier achieves 80.3% accuracy using the 5-fold crossvalidation. This is one of the studies where the transferability of the classification models is tested. The authors used an additional CEFR (Common European Framework of Reference for Languages) corpus. This small data set of CEFR-graded texts is tailored for learners of English (Council of Europe 2001) and also contains 5 readability classes. The SVM classifier trained on the WeeBit corpus and tested on the CEFR corpus achieved the classification accuracy of 23.3%, hardly beating the majority classifier baseline. This low result was attributed to the differences in readability classes in both corpora, since WeeBit classes are targeting children of different age groups, and CEFR corpus classes are targeting mostly adult foreigners with different levels of English comprehension. However, this result is a strong indication that transferability of readability classification models across different types of texts is questionable.

Two other studies that deal with the multi-genre prospects of readability prediction were conducted by Sheehan, Flor, and Napolitano (2013) and Napolitano, Sheehan, and Mundkowsky (2015). Both studies describe the problem in the context of the TextEvaluator Tool (Sheehan et al. 2010), an online system for text complexity analysis. The system supports multi-genre readability prediction with the help of a two-stage prediction workflow, in which first the genre of the text is determined (as being informational, literary, or mixed) and after that its readability level is predicted with an appropriate genre-specific readability prediction model. Similarly to the study above, this work also indicates that using classification models for cross-genre prediction is not feasible.

When it comes to multi- and crosslingual classification, Madrazo Azpiazu and Pera (2020) explore the possibility of a crosslingual readability assessment and show that their methodology called CRAS (Crosslingual Readability Assessment Strategy), which includes building a classifier that uses a set of traditional, lexico-semantic, syntactic, and discourse cohesion-based features works well in a multilingual setting. They also show that classification for some low resource languages can be improved by including documents from a different language into the train set for a specific language.

2 Later research by Xia, Kochmar, and Briscoe (2016) called the validity of the published experimental results into question; therefore, the reported 93.3% accuracy might not be the objective state-of-the-art result for readability classification.

2.3 Neural Classification Approaches

Recently, several neural approaches for readability prediction have been proposed. Nadeem and Ostendorf (2018) tested two different architectures on the WeeBit corpus regression task, namely, sequential Gated Recurrent Unit (GRU) (Cho et al. 2014) based RNN with the attention mechanism and hierarchical RNNs (Yang et al. 2016) with two distinct attention types: a more classic attention mechanism proposed by Bahdanau, Cho, and Bengio (2014), and multi-head attention proposed by Vaswani et al. (2017). The results of the study indicate that hierarchical RNNs generally perform better than sequential. Nadeem and Ostendorf (2018) also show that neural networks can be a good alternative to more traditional feature-based models for readability prediction on texts shorter than 100 words, but do not perform that competitively on longer texts.

Another version of a hierarchical RNN with the attention mechanism was proposed by Azpiazu and Pera (2019). Their system, named Vec2Read, is a multi-attentive RNN capable of leveraging hierarchical text structures with the help of word and sentence level attention mechanisms and a custom-built aggregation mechanism. They used the network in a multilingual setting (on corpora containing Basque, Catalan, Dutch, English, French, Italian, and Spanish texts). Their conclusion was that although the number of instances used for training has a strong effect on the overall performance of the system, no language-specific patterns emerged that would indicate that prediction of readability in some languages is harder than in others.

An even more recent neural approach for readability classification on the cleaned WeeBit corpus (Xia, Kochmar, and Briscoe 2016) was proposed by Filighera, Steuer, and Rensing (2019), who tested a set of different embedding models, word2vec (Mikolov et al. 2013), the uncased Common Crawl GloVe (Pennington, Socher, and Manning 2014), ELMo (Peters et al. 2018), and BERT (Devlin et al. 2019). The embeddings were fed to either a recurrent or a convolutional neural network. The BERT-based approach from their work is somewhat similar to the BERT-based supervised classification approach proposed in this work. However, one main distinction is that no fine-tuning is conducted on the BERT model in their experiments (i.e., the extraction of embeddings is conducted on the pretrained BERT language model). Their best ELMo-based model with a bidirectional LSTM achieved an accuracy of 79.2% on the development set, slightly lower than the accuracy of 80.3% achieved by Xia, Kochmar, and Briscoe (2016) in the 5-fold crossvalidation scenario. However, they did manage to improve on the state of the art by an ensemble of all their models, achieving the accuracy of 81.3%, and the macro averaged F_1-score of 80.6%.

A somewhat different neural approach to readability classification was proposed by Mohammadi and Khasteh (2019), who tackled the problem with deep reinforcement learning, or more specifically, with a deep convolutional recurrent double dueling Q network (Wang et al. 2016) using a limited window of 5 adjacent words. GloVe embeddings and statistical language models were used to represent the input text in order to eliminate the need for sophisticated NLP features. The model was used in a multilingual setting (on English and Persian data sets) and achieved performance comparable to the state of the art on all of the data sets, among them also on the Weebit corpus (accuracy of 91%).

Finally, a recent study by Deutsch, Jasbi, and Shieber (2020) used predictions of HAN and BERT models as additional features in their SVM model that also utilized a set of syntactic and lexico-semantic features. Although they did manage to improve the performance of their SVM classifiers with the additional neural features, they concluded

that additional syntactic and lexico-semantic features did not generally improve the predictions of the neural models.

3. Data Sets

In this section, we first present the data sets used in the experiments (Section 3.1) and then conduct their preliminary analysis (Section 3.2) in order to assess the feasibility of the proposed experiments. Data set statistics are presented in Table 1.

3.1 Data Set Presentation

All experiments are conducted on four corpora labeled with readability scores:

- **The WeeBit corpus**: The articles from WeeklyReader[3] and BBC-Bitesize[4] are classified into five classes according to the age group they are targeting. The classes correspond to age groups 7–8, 8–9, 9–10, 10–14, and 14–16 years. Three classes targeting younger audiences consist of articles from WeeklyReader, an educational newspaper that covers a wide range of nonfiction topics, from science to current affairs. Two classes targeting older audiences consist of material from the BBC-Bitesize Web site, containing educational material categorized into topics that roughly match school subjects in the UK. In the original corpus of Vajjala and Meurers (2012), the classes are balanced and the corpus contains altogether 3,125 documents, 625 per class. In our experiments, we followed recommendations of Xia, Kochmar, and Briscoe (2016) to fix broken sentences and remove additional meta information, such as copyright declaration and links, strongly correlated with the target labels. We reextracted the corpus from the HTML files according to the procedure described in Xia, Kochmar, and Briscoe (2016) and discarded some documents because of the lack of content after the extraction and cleaning process. The final corpus used in our experiments contains altogether 3,000 documents, 600 per class.
- **The OneStopEnglish corpus** (Vajjala and Lučić 2018) contains aligned texts of three distinct reading levels (beginner, intermediate, and advanced) that were written specifically for English as Second Language (ESL) learners. The corpus was compiled over the period 2013–2016 from the weekly news lessons section of the language learning resources `onestopenglish.com`. The section contains articles sourced from the Guardian newspaper, which were rewritten by English teachers to target three levels of adult ESL learners (elementary, intermediate, and advanced). Overall, the document-aligned parallel corpus consists of 189 texts, each written in three versions (567 in total). The corpus is freely available.[5]

3 `http://www.weeklyreader.com`.
4 `http://www.bbc.co.uk/bitesize`.
5 `https://zenodo.org/record/1219041`.

- **The Newsela corpus** (Xu, Callison-Burch, and Napoles 2015): We use the version of the corpus from 29 January 2016 consisting of altogether 10,786 documents, out of which we only used 9,565 English documents. The corpus contains 1,911 original English news articles and up to four simplified versions for every original article, that is, each original news article has been manually rewritten up to 4 times by editors at Newsela, a company that produces reading materials for pre-college classroom use, in order to target children at different grade levels and help teachers prepare curricula that match the English language skills required at each grade level. The data set is a document-aligned parallel corpus of original and simplified versions corresponding to altogether eleven different imbalanced grade levels (from 2nd to 12th grade).

- **Corpus of Slovenian school books (Slovenian SB):** In order to test the transferability of the proposed approaches to other languages, a corpus of Slovenian school books was compiled. The corpus contains 3,639,665 words in 125 school books for nine grades of primary schools and four grades of secondary school. It was created with several aims, like studying different quality aspects of school books, extraction of terminology, and linguistic analysis. The corpus contains school books for 16 distinct subjects with very different topics ranging from literature, music, and history to math, biology, and chemistry, but not in equal proportions, with readers being the largest type of school books included.

  Whereas some texts were extracted from the Gigafida reference corpus of written Slovene (Logar et al. 2012), most of the texts were extracted from PDF files. After the extraction, we first conduct some light manual cleaning on the extracted texts (removal of indices, copyright statements, references, etc.). Next, in order to remove additional noise (tips, equations, etc.), we apply a filtering script that relies on manually written rules for sentence extraction (e.g., a text is a sentence if it starts with an uppercase and ends with an end-of-sentence punctuation) to obtain only passages containing sentences. Final extracted texts come without structural information (where does a specific chapter end or start, which sentences constitute a paragraph, where are questions, etc.), since labeling the document structure would require a large amount of manual effort; therefore we did not attempt it for this research.

  For supervised classification experiments, we split the school books into chunks 25 sentences long, in order to build a train and test set with a sufficient number of documents.[6] The length of 25 sentences was chosen due to size limitations of the BERT classifier, which can be fed documents that contain up to 512 byte-pair tokens (Kudo and Richardson 2018),[7] which on average translates to slightly less than 25 sentences.

6 Note that this chunking procedure might break the text cohesion and that topical similarities between chunks from the same chapter (or paragraphs) might have a positive effect on the performance of the classification. However, because the corpus does not contain any high-level structural information (e.g., the information about paragraph or chapter structure of a specific school book), no other more refined chunking method is possible.

7 Note that the BERT tokenizer uses byte-pair tokenization (Kudo and Richardson 2018), which in some cases generates tokens that correspond to sub-parts of words rather than entire words. In the case of Slovenian SB, 512 byte-pair tokens correspond to 306 word tokens on average.

Table 1
Readability classes, number of documents, tokens per specific readability class, and average tokens per document in each readability corpus.

Readability class	#documents	#tokens	#tokens per doc.
	Wikipedia		
simple	130,000	10,933,710	84.11
balanced	130,000	10,847,108	83.44
normal	130,000	10,719,878	82.46
	OneStopEnglish		
beginner	189	100,800	533.33
intermediate	189	127,934	676.90
advanced	189	155,253	820.49
All	567	383,987	677.23
	WeeBit		
age 7–8	600	77,613	129.35
age 8–9	600	100,491	167.49
age 9–10	600	159,719	266.20
age 10–14	600	89,548	149.25
age 14–16	600	152,402	254.00
All	3,000	579,773	193.26
	Newsela		
2nd grade	224	74,428	332.27
3rd grade	500	197,992	395.98
4th grade	1,569	923,828	588.80
5th grade	1,342	912,411	679.89
6th grade	1,058	802,057	758.09
7th grade	1,210	979,471	809.48
8th grade	1,037	890,358	858.59
9th grade	750	637,784	850.38
10th grade	20	19,012	950.60
11th grade	2	1,130	565.00
12th	1,853	1,833,781	989.63
All	9,565	7,272,252	760.30
	KRES-balanced		
balanced	/	2,402,263	/
	Slovenian SB		
1st-ps	69	12,921	187.26
2nd-ps	146	30,296	207.51
3rd-ps	268	62,241	232.24
4th-ps	1,007	265,242	263.40
5th-ps	1,186	330,039	278.28
6th-ps	959	279,461	291.41
7th-ps	1,470	462,551	314.66
8th-ps	1,844	540,944	293.35
9th-ps	2,154	688,149	319.47
1st-hs	1,663	578,694	347.98
2nd-hs	590	206,147	349.40
3rd-hs	529	165,845	313.51
4th-hs	45	14,313	318.07
All	11,930	3,636,843	304.85

Language models are trained on large corpora of texts. For this purpose, we used the following corpora.

- **Corpus of English Wikipedia** and **Corpus of Simple Wikipedia** articles: We created three corpora for the use in our unsupervised English experiments:[8]
  - **Wiki-normal** contains 130,000 randomly selected articles from the Wikipedia dump, which comprise 489,976 sentences and 10,719,878 tokens.
  - **Wiki-simple** contains 130,000 randomly selected articles from the Simple Wikipedia dump, which comprise 654,593 sentences and 10,933,710 tokens.
  - **Wiki-balanced** contains 65,000 randomly selected articles from the Wikipedia dump (dated 26 January 2018) and 65,000 randomly selected articles from the Simple Wikipedia dump. Altogether the corpus comprises 571,964 sentences and 10,847,108 tokens.
- **K**RES-balanced: The KRES corpus (Logar et al. 2012) is a 100 million word balanced reference corpus of Slovenian language: 35% of its content is books, 40% periodicals, and 20% Internet texts. From this corpus we took all the available documents from two children's magazines (Ciciban and Cicido), all documents from four teenager magazines (Cool, Frka, PIL plus, and Smrklja), and documents from three magazines targeting adult audiences (Življenje in tehnika, Radar, City magazine). With these texts, we built a corpus with approximately 2.4 million words. The corpus is balanced in a sense that about one-third of the sentences come from documents targeting children, one-third is targeting teenagers, and the last third is targeting adults.

3.2 Data Set Analysis

Overall, there are several differences between our data sets:

- **Language:** As already mentioned, we have three English (Newsela, OneStopEnglish and WeeBit), and one Slovenian (Slovenian SB) test data set.
- **Parallel corpora vs. unaligned corpora:** Newsela and OneStopEnglish data sets are parallel corpora, which means that articles from different readability classes are semantically similar to each other. On the other hand, WeeBit and Slovenian SB data sets contain completely different articles in each readability class. Although this might not affect traditional readability measures, which do not take semantic information into account, it might prove substantial for the performance of classifiers and the proposed language model-based readability measures.

8 English Wikipedia and Simple Wikipedia dumps from 26 January 2018 were used for the corpus construction.

- **Length of documents:** Another difference between Newsela and OneStopEnglish data sets on one side, and WeeBit and Slovenian SB data set on the other, is the length of data set documents. Newsela and OneStopEnglish data sets contain longer documents, on average about 760 and 677 words long, and documents in the WeeBit and Slovenian SB corpora are on average about 193 and 305 words long, respectively.
- **Genre:** OneStopEnglish and Newsela data sets contain news articles, WeeBit is made of educational articles, and the Slovenian SB data set is composed of school books. For training of the English language models, we use Wikipedia and Simple Wikipedia, which contain encyclopedia articles, and for Slovene language model training, we use the KRES-balanced corpus, which contains magazine articles.
- **Target audience:** OneStopEnglish is the only test data set that specifically targets adult ESL learners and not children, as do other test data sets. When it comes to data sets used for language model training, KRES-balanced corpus is made of articles that target both adults and children. The problem with Wikipedia and Simple Wikipedia is that no specific target audience is addressed because articles are written by volunteers. In fact, using Simple Wikipedia as a data set for the training of simplification algorithms has been criticized in the past because of its lack of specific simplification guidelines, which are based only on the declarative statement that Simple Wikipedia was created for "children and adults who are learning the English language" (Xu, Callison-Burch, and Napoles 2015). This lack of guidelines also contributes to the decrease in the quality of simplification according to Xu, Callison-Burch, and Napoles (2015), who found that the corpus can be noisy and that half of its sentences are not actual simplifications but rather copied from the original Wikipedia.

This diversity of the data sets limits ambitions of the study to offer general conclusions true across genres, languages, or data sets. On the other hand, it offers an opportunity to determine how the specifics of each data set affect each of the proposed readability predictors and also to determine the overall robustness of the applied methods.

Although many aspects differ from one data set to another, there are also some common characteristics across all the data sets, which allow using the same prediction methods on all of them. These are mostly connected to the common techniques used in the construction of the readability data sets, no matter the language, genre, or target audience of the specific data set. The creation of parallel simplification corpora (i.e., Newsela, OneStopEnglish, and Simple Wikipedia) generally involves three techniques, splitting (breaking a long sentence into shorter ones), deletion (removing unimportant parts of a sentence), and paraphrasing (rewriting a text into a simpler version via reordering, substitution, and occasionally expansion) (Feng 2008). Even though there might be some subtleties involved (because what constitutes simplification for one type of user may not be appropriate for another), how these techniques are applied is rather general. Also, although there is no simplification used in the non-parallel corpora (WeeBit, Slovenian SB), the contributing authors were nevertheless instructed to write the text for a specific target group and adapt the writing style accordingly. In most

cases, this leads to the same result (e.g., shorter, less complex sentences and simpler vocabulary used in texts intended for younger or less fluently speaking audiences).

The claim of commonality between data sets can be backed up by the fact that even traditional readability indicators correlate quite well with human assigned readability, no matter the specific genre, language, or purpose of each data set. Results in Table 2 demonstrate this point by showcasing readability scores of traditional readability formulas from Section 2.1.1. We can see that the general pattern of increased difficulty on all data sets and for all indicators—larger readability scores (or in the case of FRE, smaller) are assigned to those classes of the data set that contain texts written for older children or more advanced ESL learners. This suggests that multi-data set, multi-genre, and even multilingual readability prediction is feasible on the set of chosen data sets, even if only the shallow traditional readability indicators are used.

However, the results do indicate that cross-genre or even cross-data set readability prediction might be problematic because the data sets do not cover the same readability range according to the shallow prediction formulas (and also ground truth readability labels). For example, documents in the WeeBit 14–16 age group have scores very similar to the Newsela 6th grade documents, which means that a classifier trained on the WeeBit corpus might have a hard time classifying documents belonging to higher Newsela grades since the readability of these documents is lower than for the most complex documents in the WeeBit corpus according to all of the shallow readability indicators. For this reason, we opted not to perform any supervised cross-data set or cross-genre experiments. Nevertheless, the problem of cross-genre prediction is important in the context of the proposed unsupervised experiments, because the genre discrepancy between the data sets used for training the language models and the data sets on which the models are used might influence the performance of the proposed language model-based measures. A more detailed discussion on this topic is presented in Section 4.2.

The analysis in Table 2 also confirms the findings by Madrazo Azpiazu and Pera (2020), who have shown that crosslingual readability prediction with shallow readability indicators is problematic. For example, if we compare the Newsela corpus and Slovenian SB corpus, which both cover roughly the same age group, we can see that for some readability indicators (FRE, FKGL, DCRF, and ASL) the values are on entirely different scales.

4. Unsupervised Neural Approach

In this section, we explore how neural language models can be used for determining the readability of the text in an unsupervised way. In Section 4.1, we present the neural architectures used in our experiments; in Section 4.2, we describe the methodology of the proposed approach; and in Section 4.3, we present the conducted experiments.

4.1 Neural Language Model Architectures

Mikolov et al. (2011) have shown that neural language models outperform *n*-gram language models by a high margin on large and also relatively small (less than 1 million tokens) data sets. The achieved differences in perplexity (see Equation (1)) are attributed to a richer historical contextual information available to neural networks, which are not limited to a small contextual window (usually of up to 5 previous words) as is the case of *n*-gram language models. In Section 2.1.5, we mentioned some approaches that use

Table 2
Scores of traditional readability indicators from Section 2.1.1 for specific classes in the readability data sets.

Class	GFI	FRE	FKGL	ARI	DCRF	SMOG	ASL
			Wikipedia				
simple	11.80	62.20	8.27	14.08	11.40	11.40	16.90
balanced	13.49	56.17	9.70	15.86	12.53	12.53	19.54
normal	15.53	49.16	11.47	18.06	13.89	13.89	23.10
			WeeBit				
age 7–8	6.91	83.41	3.82	8.83	7.83	7.83	10.23
age 8–9	8.45	76.68	5.34	10.33	8.87	8.87	12.89
age 9–10	10.30	69.88	6.93	12.29	10.01	10.01	15.69
age 10–14	9.94	75.35	6.34	11.20	9.67	9.67	16.64
age 14–16	11.76	66.61	8.09	13.56	10.81	10.81	18.86
			OneStopEnglish				
beginner	11.79	66.69	8.48	13.93	11.05	11.05	20.74
intermediate	13.83	59.68	10.19	15.98	12.30	12.30	23.98
advanced	15.35	54.84	11.54	17.65	13.22	13.22	26.90
			Newsela				
2nd grade	6.11	85.69	3.27	8.09	7.26	7.26	9.26
3rd grade	7.24	80.92	4.27	9.30	7.94	7.94	10.72
4th grade	8.58	76.05	5.40	10.50	8.88	8.88	12.72
5th grade	9.79	71.76	6.47	11.73	9.68	9.68	14.81
6th grade	11.00	67.46	7.53	12.99	10.47	10.47	16.92
7th grade	12.11	62.71	8.54	14.12	11.26	11.26	18.46
8th grade	13.05	60.37	9.38	15.19	11.83	11.83	20.81
9th grade	14.20	55.00	10.46	16.37	12.70	12.70	22.17
10th grade	14.15	55.70	10.60	16.50	12.83	12.83	23.33
11th grade	15.70	56.41	11.05	16.96	12.77	12.77	24.75
12th grade	14.52	55.58	10.71	16.70	12.79	12.79	23.69
			KRES-balanced				
balanced	12.72	29.20	12.43	14.88	14.08	14.08	15.81
			Slovenian SB				
1st-ps	9.54	31.70	10.38	11.72	11.12	11.12	7.63
2nd-ps	9.49	34.90	10.11	11.34	11.26	11.26	8.37
3rd-ps	10.02	32.89	10.61	11.78	11.80	11.80	9.31
4th-ps	10.96	30.29	11.18	12.84	12.39	12.39	10.40
5th-ps	11.49	28.13	11.62	13.33	12.79	12.79	11.02
6th-ps	13.20	20.10	12.84	14.57	13.61	13.61	11.45
7th-ps	12.94	22.97	12.61	14.52	13.64	13.64	12.24
8th-ps	13.48	18.12	13.09	14.78	13.71	13.71	11.32
9th-ps	13.69	19.26	13.13	15.07	13.94	13.94	12.27
1st-hs	15.12	12.66	14.33	16.22	14.96	14.96	13.62
2nd-hs	15.13	15.13	13.90	15.83	14.67	14.67	13.49
3rd-hs	14.76	13.09	14.00	15.62	14.44	14.44	12.57
4th-hs	14.66	14.39	13.64	15.54	14.03	14.03	11.62

n-gram language models for readability prediction. However, we are unaware of any approach that would use deep neural network language models for determining the readability of a text.

In this research, we utilize three neural architectures for language modeling. First are RNNs, which are suitable for modeling sequential data. At each time step t, the input vector x_t, and hidden state vector h_{t-1} are fed into the network, producing the next hidden vector state h_t with the following recursive equation:

$$h_t = f(Wx_t + Uh_{t-1} + b)$$

where f is a nonlinear activation function, W and U are matrices representing weights of the input layer and hidden layer, and b is the bias vector. Learning long-range input dependencies with plain RNNs is problematic because of vanishing gradients (Bengio, Simard, and Frasconi 1994); therefore, in practice, modified recurrent networks, such as Long Short-Term Memory networks (LSTMs) are used. In our experiments, we use the LSTM-based language model proposed by Kim et al. (2016). This architecture is adapted to language modeling of morphologically rich languages, such as Slovenian, by utilizing an additional character-level convolutional neural network (CNN). The convolutional level learns a character structure of words and is connected to the LSTM-based layer, which produces predictions at the word level.

Bai, Kolter, and Koltun (2018) introduced a new sequence modeling architecture based on convolution, called temporal convolutional network (TCN), which is also used in our experiments. TCN uses causal convolution operations, which make sure that there is no information leakage from future time steps to the past. This and the fact that TCN takes a sequence as an input and maps it into an output sequence of the same size makes this architecture appropriate for language modeling. TCNs are capable of leveraging long contexts by using a very deep network architecture and a hierarchy of dilated convolutions. A single dilated convolution operation F on element s of the 1-dimensional sequence x can be defined with the following equation:

$$F(s) = (x *_d f)(s) = \sum_{i=0}^{k-1} f(i) \cdot x_{s-d \cdot i}$$

where $f : 0, \ldots k-1$ is a filter of size k, d is a dilation factor, and $s - d \cdot i$ accounts for the direction of the past. In this way, the context taken into account during the prediction can be increased by using larger filter sizes and by increasing the dilation factor. The most common practice is to increase the dilation factor exponentially with the depth of the network.

Recently, Devlin et al. (2019) proposed a novel approach to language modeling. Their BERT uses both left and right context, which means that a word w_t in a sequence is not determined just from its left sequence $w_{1:t-1} = [w_1, \ldots, w_{t-1}]$ but also from its right word sequence $w_{t+1:n} = [w_{t+1}, \ldots, w_{t+n}]$. This approach introduces a new learning objective, a *masked language model*, where a predefined percentage of randomly chosen words from the input word sequence is masked, and the objective is to predict these masked words from the unmasked context. BERT uses a transformer neural network architecture (Vaswani et al. 2017), which relies on the self-attention mechanism. The distinguishing feature of this approach is the use of several parallel attention layers, the so-called attention heads, which reduce the computational cost and allow the system to attend to several dependencies at once.

All types of neural network language models, TCN, LSTM, and BERT, output softmax probability distribution calculated over the entire vocabulary, and present the probabilities for each word given its historical (and in the case of BERT also future) sequence. Training of these networks usually minimizes the negative log-likelihood (NLL) of the training corpus word sequence $w_{1:n} = [w_1, \ldots, w_n]$ by backpropagation through time:

$$\text{NLL} = -\sum_{i=1}^{n} \log P(w_i | w_{1:i-1}) \tag{2}$$

In the case of BERT, the formula for minimizing NLL also uses the right-hand word sequence:

$$\text{NLL} = -\sum_{i=1}^{n} \log P(w_i | w_{1:i-1}, w_{i+1:n})$$

where w_i are the *masked words*.

The following equation, which is used for measuring the perplexity of neural language models, defines the relationship between perplexity (PPL, see Equation (1)) and NLL (Equation (2)):

$$\text{PPL} = e^{(\frac{\text{NLL}}{N})}$$

4.2 Unsupervised Methodology

Two main questions we wish to investigate in the unsupervised approach are the following:

- Can standalone neural language models be used for unsupervised readability prediction?
- Can we develop a robust new readability formula that will outperform traditional readability formulas by relying not only on shallow lexical sophistication indicators but also on neural language model statistics?

4.2.1 Language Models for Unsupervised Readability Assessment. The findings of the related research suggest that a separate language model should be trained for each readability class in order to extract features for successful readability prediction (Petersen and Ostendorf 2009; Xia, Kochmar, and Briscoe 2016). On the other hand, we test the possibility of using a neural language model as a standalone unsupervised readability predictor.

Two points that support this kind of usage are based on the fact that neural language models tend to capture much more information compared to the traditional *n*-gram models. First, because *n*-gram language models used in the previous work on readability detection were in most cases limited to a small contextual window of up to five words, their learning potential was limited to lexico-semantic information (e.g., information about the difficulty of vocabulary and word *n*-gram structures in the text), and information about the text syntax. We argue that due to much larger contextual information of the neural models (e.g., BERT leverages sequences of up to 512 byte-pair tokens), which spans across sentences, the neural language models also learn high-level textual properties, such as long-distance dependencies (Jawahar, Sagot, and Seddah 2019), in order to minimize NLL during training. Second, *n*-gram models in

past readability research have only been trained on the corpora (or, more specifically, on parts of the corpora) on which they were later used. In contrast, by training the neural models on large general corpora, the model also learns semantic information, which can be transferred when the model is used on a smaller test corpus. The success of this knowledge transfer is, to some extent, dependent on the genre compatibility of the train and test corpora.

A third point favoring greater flexibility of neural language models relies on the fact that no corpus is a monolithic block of text made out of units (i.e., sentences, paragraphs, and articles) of exactly the same readability level. This means that a language model trained on a large corpus will be exposed to chunks of text with different levels of complexity. We hypothesize that, due to this fact, the model will to some extent be able to distinguish between these levels and return a lower perplexity for more standard, predictable (i.e., readable) text. Vice versa, complex and rare language structures and vocabulary of less readable texts would negatively affect the performance of the language model, expressed via larger perplexity score. If this hypothesis is correct, then ideally, the average readability of the training corpus should fit somewhere in the middle of the readability spectrum of the testing corpus.

To test these statements, we train language models on Wiki-normal, Wiki-simple, and Wiki-balanced corpora described in Section 3. All three Wiki corpora contain roughly the same amount of text, in order to make sure that the training set size does not influence the results of the experiments. We expect the following results:

- **Hypothesis 1:** Training the language models on a corpus with a readability that fits somewhere in the middle of the readability spectrum of the testing corpus will yield the best correlation between the language model's performance and readability. According to the preliminary analysis of our corpora conducted in Section 3.2 and results of the analysis in Table 2, this ideal scenario can be achieved in three cases: (i) if a language model trained on the Wiki-simple is used on the Newsela corpora, (ii) if a language model trained on the Wiki-balanced corpus is used on the OneStopEnglish corpus, and (iii) if the model trained on the KRES-balanced corpus is used on the Slovenian SB corpus, despite the mismatch of genres in these corpora.
- **Hypothesis 2:** The language models trained only on texts for adults (Wiki-normal) will show higher perplexity on texts for children (WeeBit and Newsela) because their training set did not contain such texts; this will negatively affect the correlation between the language model's performance and readability.
- **Hypothesis 3:** Training the language models only on texts for children (Wiki-simple corpus) will result in a higher perplexity score of the language model when applied to adult texts (OneStopEnglish). This will positively affect the correlation between the language model's performance and readability. However, this language model will not be able to reliably distinguish between texts for different levels of adult ESL learners, which will have a negative effect on the correlation.

To further test the viability of the unsupervised language models as readability predictors and to test the limits of using a single language model, we also explore the possibility of using a language model trained on a large general corpus. The English BERT

language model was trained on large corpora (Google Books Corpus [Goldberg and Orwant 2013] and Wikipedia) of about 3,300M words containing mostly texts for adult English speakers. According to hypothesis 2, this will have a negative effect on the correlation between the performance of the model and readability.

Because of the large size of the BERT model and its huge training corpus, the semantic information acquired during training is much larger than the information acquired by the models we train on our much smaller corpora, which means that there is a greater possibility that the BERT model was trained on some text semantically similar to the content in the test corpora and that this information can be successfully transferred. However, the question remains, exactly what type of semantic content does the BERT's training corpus contain? One hypothesis is that its training corpus contains more content specific for adult audiences and less content found in the corpora for children. This would have a negative effect on the correlation between the performance of the model and readability on the WeeBit corpus. Contrarily, because the two highest readability classes in the WeeBit corpus contain articles from different scientific fields used for the education of high school students, which can contain rather specific and technical content that is unlikely to be common in the general training corpus, this might influence a positive correlation between the performance of the model and readability. Newsela and OneStopEnglish, on the other hand, are parallel corpora, which means that the semantic content in all classes is very similar; therefore the success or failure of semantic transfer will most likely not affect these two corpora.

4.2.2 Ranked Sentence Readability Score. Based on the two considerations below, we propose a new Ranked Sentence Readability Score (RSRS) for measuring the readability with language models.

- The shallow lexical sophistication indicators, such as the length of a sentence, correlate well with the readability of a text. Using them besides statistics derived from language models could improve the unsupervised readability prediction.
- The perplexity score used for measuring the performance of a language model is an *unweighted* sum of perplexities of words in the predicted sequence. In reality, a small number of unreadable words might drastically reduce the readability of the entire text. Assigning larger weights to such words might improve the correlation of language model scores with the readability.

The proposed readability score is calculated with the following procedure. First, a given text is split into sentences with the default sentence tokenizer from the NLTK library (Bird and Loper 2004). In order to obtain a readability estimation for each word in a specific context, we compute, for each word in the sentence, the word negative log-likelihood (WNLL) according to the following formula:

$$\text{WNLL} = -(y_t \log y_p + (1 - y_t) \log (1 - y_p))$$

where y_p denotes the probability (from the softmax distribution) predicted by the language model according to the historical sequence, and y_t denotes the empirical distribution for a specific position in the sentence, that is, y_t has the value 1 for the word in the vocabulary that actually appears next in the sequence and the value 0 for all the other words in the vocabulary. Next, we sort all the words in the sentence in ascending

order according to their WNLL score, and the ranked sentence readability score (RSRS) is calculated with the following expression:

$$\text{RSRS} = \frac{\sum_{i=1}^{S} \sqrt{i} \cdot \text{WNLL}(i)}{S} \tag{3}$$

where S denotes the sentence length and i represents the rank of a word in a sentence according to its WNLL value. The square root of the word rank is used for proportionally weighting words according to their readability because initial experiments suggested that the use of a square root of a rank represents the best balance between allowing all words to contribute equally to the overall readability of the sentence and allowing only the least readable words to affect the overall readability of the sentence. For out-of-vocabulary words, square root rank weights are doubled, because these rare words are, in our opinion, good indicators of non-standard text. Finally, in order to obtain the readability score for the entire text, we calculate the average of all the RSRS scores in the text. An example of how RSRS is calculated for a specific sentence is shown in Figure 1.

The main idea behind the RSRS score is to avoid the reductionism of traditional readability formulas. We aim to achieve this by including high-level structural and semantic information through neural language model–based statistics. The first assumption is that complex grammatical and lexical structures harm the performance of the language model. Since WNLL score, which we compute for each word, depends on the context in which the word appears in, words appearing in more complex grammatical and lexical contexts will have a higher WNLL. The second assumption is that the semantic information is included in the readability calculation: Tested documents with semantics dissimilar to the documents in the language model training set will negatively affect the performance of the language model, resulting in the higher WNLL score for words with unknown semantics. The trainable nature of language models allows for customization and personalization of the RSRS for specific tasks,

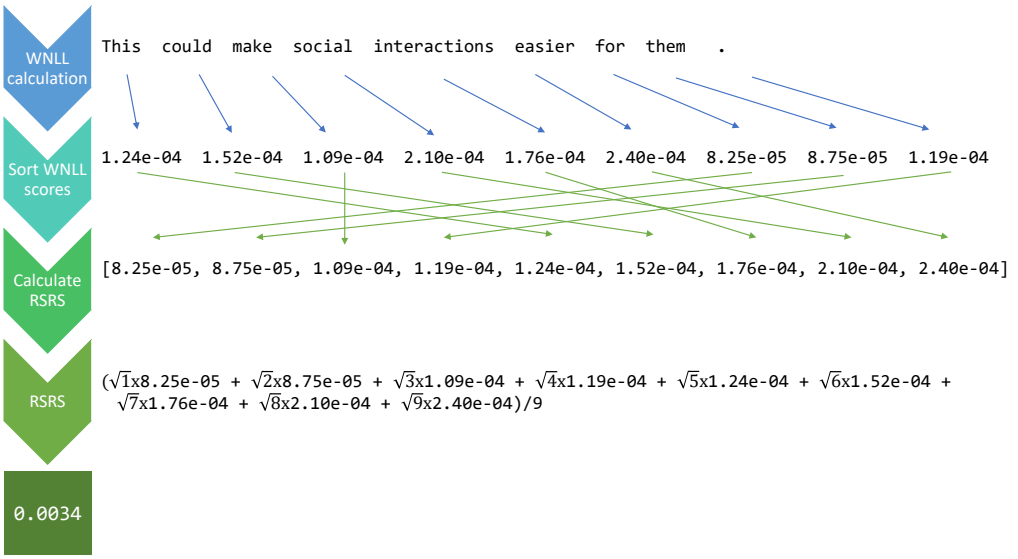


Figure 1
The RSRS calculation for the sentence *This could make social interactions easier for them.*

topics, and languages. This means that RSRS will alleviate the problem of cultural non-transferability of traditional readability formulas.

On the other hand, the RSRS also leverages shallow lexical sophistication indicators through the index weighting scheme, which ensures that less readable words contribute more to the overall readability score. This is somewhat similar to the counts of long and difficult words in the traditional readability formulas, such as GFI and DCRF. The value of RSRS also increases for texts containing longer sentences, since the square roots of the word rank weights become larger with increased sentence length. This is similar to the behavior of traditional formulas such as GFI, FRE, FKGL, ARI, and DCRF, where this effect is achieved by incorporating the ratio between the total number of words and the total number of sentences into the equation.

4.3 Unsupervised Experiments

For the presented unsupervised readability assessment methodology based on neural language models, we first present the experimental design followed by the results.

4.3.1 Experimental Design. Three different architectures of language models (described in Section 4.1) are used for experiments: a temporal convolutional network (TCN) proposed by Bai, Kolter, and Koltun (2018), a recurrent language model (RLM) using character-level CNN and LSTM proposed by Kim et al. (2016), and an attention-based language model, BERT (Devlin et al. 2019). For the experiments on the English language, we train TCN and RLM on three Wiki corpora.

To explore the possibility of using a language model trained on a general corpus for the unsupervised readability prediction, we use the BERT-base-uncased English language model, a pretrained uncased language model trained on BooksCorpus (0.8G words) (Zhu et al. 2015) and English Wikipedia (2.5G words). For the experiments on Slovenian, the corpus containing just school books is too small for efficient training of language models; therefore TCN and RLM were only trained on the KRES-balanced corpus described in Section 3. For exploring the possibility of using a general language model for the unsupervised readability prediction, a pretrained CroSloEngual BERT model trained on corpora from three languages, Slovenian (1.26G words), Croatian (1.95G words), and English (2.69G words) (Ulčar and Robnik-Šikonja 2020), is used. The corpora used in training the model are a mix of news articles and a general Web crawl.

The performance of language models is typically measured with the perplexity (see Equation (1)). To answer the research question of whether standalone language models can be used for unsupervised readability prediction, we investigate how the measured perplexity of language models correlates with the readability labels in the gold-standard WeeBit, OneStopEnglish, Newsela, and Slovenian SB corpora described in Section 3. The correlation to these ground truth readability labels is also used to evaluate the performance of the RSRS measure. For performance comparison, we calculate the traditional readability formula values (described in Section 2) for each document in the gold-standard corpora and measure the correlation between these values and manually assigned labels. As a baseline, we use the average sentence length (ASL) in each document.

The correlation is measured with the Pearson correlation coefficient (ρ). Given a pair of distributions X and Y, the covariance *cov*, and the standard deviation σ, the formula for ρ is:

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

A larger positive correlation signifies a better performance for all measures except the FRE readability measure. As this formula assigns higher scores to better-readable texts, a larger negative correlation suggests a better performance of the FRE measure.

4.3.2 Experimental Results. The results of the experiments are presented in Table 3. The ranking of measures on English and Slovenian data sets are presented in Table 4.

The correlation coefficients of all measures vary drastically between different corpora. The highest ρ values are obtained on the Newsela corpus, where the best performing measure (surprisingly this is our baseline—the average sentence length) achieves the ρ of 0.906. The highest ρ on the other two English corpora are much lower. On the WeeBit corpus, the best performance is achieved by GFI and FKGL measures (ρ of 0.544), and on the OneStopEnglish corpus, the best performance is achieved with the proposed TCN RSRS-simple (ρ of 0.615). On the Slovenian SB, the ρ values are higher, and the best performing measure is TCN RSRS score-balanced with ρ of 0.789.

The perplexity-based measures show a much lower correlation with the ground truth readability scores. Overall, they perform the worst of all the measures for both languages (see Table 4), but we can observe large differences in their performance across different corpora. Although there is either no correlation or low negative correlation between perplexities of all three language models and readability on the WeeBit corpus, there is some correlation between perplexities achieved by RLM and TCN on OneStopEnglish and Newsela corpora (the highest being the ρ of 0.566 achieved by TCN perplexity-simple on the Newsela corpus). The correlation between RLM and

Table 3
Pearson correlation coefficients between manually assigned readability labels and the readability scores assigned by different readability measures in the unsupervised setting. The highest correlation for each corpus is marked with bold typeface.

Measure/Data set	WeeBit	OneStopEnglish	Newsela	Slovenian SB
RLM perplexity-balanced	−0.082	0.405	0.512	0.303
RLM perplexity-simple	−0.115	0.420	0.470	/
RLM perplexity-normal	−0.127	0.283	0.341	/
TCN perplexity-balanced	0.034	0.476	0.537	0.173
TCN perplexity-simple	0.025	0.518	0.566	/
TCN perplexity-normal	−0.015	0.303	0.250	/
BERT perplexity	−0.123	−0.162	−0.673	−0.563
RLM RSRS-balanced	0.497	0.551	0.890	0.732
RLM RSRS-simple	0.506	0.569	0.893	/
RLM RSRS-normal	0.490	0.536	0.886	/
TCN RSRS-balanced	0.393	0.601	0.894	**0.789**
TCN RSRS-simple	0.385	**0.615**	0.894	/
TCN RSRS-normal	0.348	0.582	0.886	/
BERT RSRS	0.279	0.384	0.674	0.126
GFI	**0.544**	0.550	0.849	0.730
FRE	−0.433	−0.485	−0.775	−0.614
FKGL	**0.544**	0.533	0.865	0.697
ARI	0.488	0.520	0.875	0.658
DCRF	0.420	0.496	0.735	0.686
SMOG	0.456	0.498	0.813	0.770
ASL	0.508	0.498	**0.906**	0.683

Table 4
Ranking (lower is better) of measures on English and Slovenian data sets sorted by the average rank on all data sets for which the measure is available.

Measure	WeeBit	OneStopEnglish	Newsela	Slovenian SB
RLM RSRS-simple	4	4	4	/
TCN RSRS-balanced	11	2	2	**1**
RLM RSRS-balanced	5	5	5	3
GFI	**1**	6	10	4
TCN RSRS-simple	12	**1**	3	/
ASL	3	12	**1**	7
FKGL	2	8	9	5
RLM RSRS-normal	6	7	6	/
TCN RSRS-normal	13	3	7	/
ARI	7	9	8	8
SMOG	8	11	11	2
DCRF	10	13	13	6
FRE	9	14	12	9
TCN perplexity-simple	16	10	15	/
TCN perplexity-balanced	15	15	16	11
BERT RSRS	14	18	14	12
RLM perplexity-balanced	18	17	17	10
RLM perplexity-simple	19	16	18	/
TCN perplexity-normal	17	19	20	/
BERT perplexity	20	21	21	13
RLM perplexity-normal	21	20	19	/

TCN perplexity measures and readability classes on the Slovenian SB corpus is low, with RLM perplexity-balanced showing the ρ of 0.303 and TCN perplexity-balanced achieving ρ of 0.173.

BERT perplexities are negatively correlated with readability, and the negative correlation is relatively strong on Newsela and Slovenian school books corpora (ρ of -0.673 and -0.563, respectively), and weak on WeeBit and OneStopEnglish corpora. As BERT was trained on corpora that are mostly aimed at adults, the strong negative correlation on Newsela and Slovenian SB corpora seem to suggest that BERT language models might actually be less perplexed by the articles aimed at adults than the documents aimed at younger audiences. This is supported by the fact that the negative correlation is weaker on the OneStopEnglish corpus, which is meant for adult audiences, and for which our analysis (see Section 3.2) has shown that it contains more complex texts according to the shallow readability indicators.

Nevertheless, the weak negative correlation on the WeeBit corpus is difficult to explain as one would expect a stronger negative correlation because the same analysis showed that WeeBit contains the least complex texts out of all the tested corpora. If this result is connected with the successful transfer of the semantic knowledge, it supports the hypothesis that the two classes containing the most complex texts in the WeeBit corpus contain articles with rather technical content that perplex the BERT model. However, the role of the semantic transfer should also dampen the negative correlation on the Slovenian SB, which is a non-parallel corpus and also contains rather technical educational content meant for high-school children. Perhaps the transfer is less successful for Slovenian since the Slovenian corpus on which the CroSloEngual BERT was trained is smaller than the English corpora used for training of English BERT.

Although further experiments and data are needed to pinpoint the exact causes for the discrepancies in the results, we can still conclude that using a single language model trained on general corpora for unsupervised readability prediction of texts for younger audiences or English learners is, at least according to our results, not a viable option.

Regarding our expectations that performance of the language model trained on a corpus with average readability that fits somewhere in the middle of the readability spectrum of the testing corpus would yield the best correlation with manually labeled readability scores, it is interesting to look at the differences in performance between TCN and RLM perplexity measures trained on Wiki-normal, Wiki-simple, and Wiki-balanced corpora. As expected, the correlation scores are worse on the WeeBit corpus, since all classes in this corpus contain texts that are less complex than texts in any of the training corpora. On the OneStopEnglish corpus, both Wiki-simple perplexity measures perform the best, which is unexpected, since we would expect the balanced measure to perform better. On the Newsela corpus, RLM perplexity-balanced outperforms RLM perplexity-simple by 0.042 (which is unexpected), and TCN perplexity-simple outperforms TCN perplexity-balanced by 0.029, which is according to the expectations. Also, according to the expectation is the fact, that both Wiki-normal perplexity measures are outperformed by a large margin by Wiki-simple and Wiki-balanced perplexity measures on the OneStopEnglish and the Newsela corpora. Similar observations can be made with regard to RSRS, which also leverages language model statistics. On all corpora, the performance of Wiki-simple RSRS measures and Wiki-balanced RSRS measures is comparable, and these measures consistently outperform Wiki-normal RSRS measures.

These results are not entirely compatible with hypothesis 1 in Section 4.2.1 that Wiki-balanced measures would be most correlated with readability on the OneStopEnglish corpus and that Wiki-simple measures would be most correlated with readability on the Newsela corpus. Nevertheless, training the language models on the corpora with readability in the middle of the readability spectrum of the test corpus seems to be an effective strategy, because the differences in performance between Wiki-balanced and Wiki-simple measures are not large. On the other hand, the good performance of the Wiki-simple measures supports our hypothesis 3 in Section 4.2.1, that training the language models on texts with the readability closer to the bottom of the readability spectrum of the test corpus for children will result in a higher perplexity score of the language model when applied to adult texts, which will have a positive effect on the correlation with readability.

The fact that positive correlation between readability and both Wiki-simple and Wiki-balanced perplexity measures on the Newsela and OneStopEnglish corpora is quite strong supports the hypothesis that more complex language structures and vocabularies of less readable texts would result in a higher perplexity on these texts. Interestingly, strong correlations also indicate that the genre discrepancies between the language model train and test sets do not appear to have a strong influence on the performance. Whereas the choice of a neural architecture for language modeling does not appear to be that crucial, the readability of the language model training set is of utmost importance. If the training set on average contains more complex texts than the majority of texts in the test set, as in the case of language models trained just on the Wiki-normal corpus (and also BERTs), the correlation between readability and perplexity disappears or even gets reverted, since language models trained on more complex language structures learn how to handle these difficulties.

The low performance of perplexity measures suggests that neural language model statistics are not good indicators of readability and should therefore not be used alone for readability prediction. Nevertheless, the results of TCN RSRS and RLM RSRS

suggest that language models contain quite useful information when combined with other shallow lexical sophistication indicators, especially when readability analysis needs to be conducted on a variety of different data sets.

As seen in Table 4, shallow readability predictors can give inconsistent results on data sets from different genres and languages. For example, the simplest readability measure, the average sentence length, ranked first on Newsela and twelfth on OneStopEnglish. It also did not do well on the Slovenian SB corpus, where it ranked seventh. SMOG, on the other hand, ranked very well on the Slovenian SB corpus (rank 2) but ranked twice as eleventh and once as eighth on the English corpora. Among the traditional measures, GFI presents the best balance in performance and consistency, ranking first on WeeBit, sixth on OneStopEnglish, tenth on Newsela, and fourth on Slovenian SB.

On the other hand, RSRS-simple and RSRS-balanced measures offer more robust performance across data sets from different genres and languages according to ranks in Table 4. For example, the RLM RSRS-simple measure ranked fourth on all English corpora. The TCN RSRS-balanced measure, which was also used on Slovenian SB, ranked first on Slovenian SB and second on OneStopEnglish and Newsela. However, it did not do well on WeeBit, where the discrepancy in readability between the language model train and test sets was too large. RLM RSRS-balanced was more consistent, ranking fifth on all English corpora and third on Slovenian SB. These results suggest that language model statistics can improve the consistency of predictions on a variety of different data sets. The robustness of the measure is achieved by training the language model on a specific train set, with which one can optimize the RSRS measure for a specific task and language.

5. Supervised Neural Approach

As mentioned in Section 1, recent trends in text classification show the domination of deep learning approaches that internally use automatic feature construction. Existing neural approaches to readability prediction (see Section 2.3) tend to generalize better across data sets and genres (Filighera, Steuer, and Rensing 2019), and therefore solve the problem of classical machine learning approaches relying on an extensive feature engineering (Xia, Kochmar, and Briscoe 2016).

In this section, we analyze how different types of neural classifiers can predict text readability. In Section 5.1, we describe the methodology, and in Section 5.2 we present experimental scenarios and results of conducted experiments.

5.1 Supervised Methodology

We tested three distinct neural network approaches to text classification:

- Bidirectional long short-term memory network (BiLSTM). We use the RNN approach proposed by Conneau et al. (2017) for classification. The BiLSTM layer is a concatenation of forward and backward LSTM layers that read documents in two opposite directions. The max and mean pooling are applied to the LSTM output feature matrix in order to get the maximum and average values of the matrix. The resulting vectors are concatenated and fed to the final linear layer responsible for predictions.

- Hierarchical attention networks (HAN). We use the architecture of Yang et al. (2016) that takes hierarchical structure of text into account with the two-level attention mechanism (Bahdanau, Cho, and Bengio 2014; Xu et al. 2015) applied to word and sentence representations encoded by BiLSTMs.
- Transfer learning. We use the pretrained BERT transformer architecture with 12 layers of size 768 and 12 self-attention heads. A linear classification head was added on top of the pretrained language model, and the whole classification model was fine-tuned on every data set for three epochs. For English data sets, the BERT-base-uncased English language model is used, while for the Slovenian SB corpus, we use the CroSloEngual BERT model trained on Slovenian, Croatian, and English (Ulčar and Robnik-Šikonja 2020).[9]

We randomly shuffle all the corpora, and then Newsela and Slovenian SB corpora are split into a train (80% of the corpus), validation (10% of the corpus), and test (10% of the corpus) sets. Because of the small number of documents in OneStopEnglish and WeeBit corpora (see description in Section 3), we used five-fold stratified crossvalidation on these corpora to get more reliable results. For every fold, the corpora were split into the train (80% of the corpus), validation (10% of the corpus), and test (10% of the corpus) sets. We employ Scikit StratifiedKFold,[10] both for train-test splits and five-fold crossvalidation splits, in order to preserve the percentage of samples from each class.

BiLSTM and HAN classifiers were trained on the train set and tested on the validation set after every epoch (for a maximum of 100 epochs). The best performing model on the validation set was selected as the final model and produced predictions on the test sets. BERT models are fine-tuned on the train set for three epochs, and the resulting model is tested on the test set. The validation sets were used in a grid search to find the best hyperparameters of the models. For BiLSTM, all combinations of the following hyperparameter values were tested before choosing the best combination, which is written in bold in the list below:

- Batch size: **8**, 16, 32
- Learning rates: 0.00005, **0.0001**, 0.0002, 0.0004, 0.0008
- Word embedding size: 100, **200**, 400
- LSTM layer size: **128**, 256
- Number of LSTM layers: 1, **2**, 3, 4
- Dropout after every LSTM layer: 0.2, **0.3**, 0.4

For HAN, we tested all combinations of the following hyperparameter values (the best combination is written in bold):

- Batch size: 8, **16**, 32
- Learning rates: 0.00005, **0.0001**, 0.0002, 0.0004, 0.0008

9 Both models are available through the Transformers library `https://huggingface.co/transformers/`.
10 `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html.`,

- Word embedding size: 100, **200**, 400
- Sentence embedding size: **100**, 200, 400

For BERT fine-tuning, we use the default learning rate of 0.00002. The input sequence length is limited to 512 byte-pair tokens, which is the maximum supported input sequence length.

We used the same configuration for all the corpora and performed no corpus specific tweaking of classifier parameters. We measured the performance of all the classifiers in terms of accuracy (in order to compare their performance to the performance of the classifiers from the related work), weighted average precision, weighted average recall, and weighted average F_1-score.[11] Since readability classes are ordinal variables (in our case ranging from 0 to n = number of classes−1), not all mistakes of classifiers are equal; therefore we also utilize the Quadratic Weighted Kappa (QWK) measure, which allows for mispredictions to be weighted differently, according to the cost of a specific mistake. Calculation of the QWK involves three matrices containing observed scores, ground truth scores, and the weight matrix scores, which in our case correspond to the distance d between the classes c_i and c_j and is defined as $d = |c_i - c_j|$. QWK is therefore calculated as:

$$\text{QWK} = 1 - \frac{\sum_{i=1}^{c}\sum_{j=1}^{c} w_{ij}x_{ij}}{\sum_{i=1}^{c}\sum_{j=1}^{c} w_{ij}m_{ij}} \tag{4}$$

where c is the number of readability classes and w_{ij}, x_{ij}, and m_{ij} are elements in the weight, observed, and ground truth matrices, respectively.

5.2 Supervised Experimental Results

The results of supervised readability assessment using different architectures of deep neural networks are presented in Table 5, together with the state-of-the-art baseline results from the related work (Xia, Kochmar, and Briscoe 2016; Filighera, Steuer, and Rensing 2019; Deutsch, Jasbi, and Shieber 2020). We only present the best result reported by each of the baseline studies; the only exception is Deutsch, Jasbi, and Shieber (2020), for which we present two results, SVM-BF (SVM with BERT features) and SVM-HF (SVM with HAN features) that proved the best on the WeeBit and Newsela corpora, respectively.

On the WeeBit corpus, by far the best performance according to all measures was achieved by BERT. In terms of accuracy, BERT outperforms the second-best BiLSTM by about 8 percentage points, achieving the accuracy of 85.73%. HAN performs the worst on the WeeBit corpus according to all measures. BERT also outperforms the accuracy result reported by Xia, Kochmar, and Briscoe (2016), who used the five-fold crossvalidation setting and the accuracy result on the development set reported by Filighera, Steuer, and Rensing (2019).[12] In terms of weighted F_1-score, both strategies

11 We use the Scikit implementation of the metrics (`https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics`) and set the "average" parameter to "weighted."

12 For the study by Filighera, Steuer, and Rensing (2019), we report accuracy on the development set instead of accuracy on the test set, as the authors claim that this result is more comparable to the results achieved in the crossvalidation setting. On the test set, Filighera, Steuer, and Rensing (2019) report the best accuracy of 74.4%.

Table 5
The results of the supervised approach to readability in terms of accuracy, weighted precision, weighted recall, and weighted F_1-score for the three neural network classifiers and methods from the literature.

Measure/Data set	WeeBit	OneStopEnglish	Newsela	Slovenian SB
Filighera et al. (2019) accuracy	0.8130	–	–	–
Xia et al. (2016) accuracy	0.8030	–	–	–
SVM-BF (Deutsh et al., 2020) F_1	0.8381	–	0.7627	–
SVM-HF (Deutsh et al., 2020) F_1	–	–	0.8014	–
Vajjala et al. (2018) accuracy	–	0.7813	–	–
BERT accuracy	**0.8573**	0.6738	0.7573	0.4545
BERT precision	**0.8658**	0.7395	0.7510	0.4736
BERT recall	**0.8573**	0.6738	0.7573	0.4545
BERT F_1	**0.8581**	0.6772	0.7514	0.4157
BERT QWK	**0.9527**	0.7077	0.9789	**0.8855**
HAN accuracy	0.7520	**0.7872**	**0.8138**	0.4887
HAN precision	0.7534	**0.7977**	**0.8147**	0.4866
HAN recall	0.7520	**0.7872**	**0.8138**	0.4887
HAN F_1	0.7520	**0.7888**	**0.8101**	0.4847
HAN QWK	0.8860	**0.8245**	**0.9835**	0.8070
BiLSTM accuracy	0.7743	0.6875	0.7111	**0.5277**
BiLSTM precision	0.7802	0.7177	0.6910	**0.5239**
BiLSTM recall	0.7743	0.6875	0.7111	**0.5277**
BiLSTM F_1	0.7750	0.6920	0.6985	**0.5219**
BiLSTM QWK	0.9060	0.7230	0.9628	0.7980

that use BERT (utilizing the BERT classifier directly or feeding BERT features to the SVM classifier as in Deutsch, Jasbi, and Shieber [2020]) seem to return similar results. Finally, in terms of QWK, BERT achieves a very high score of 95.27% and the other two tested classifiers obtain a good QWK score close to 90%.

The best result on Newsela is achieved by HAN, achieving the F_1-score of 81.01% and accuracy of 81.38%. This is similar to the baseline SVM-HF result achieved by Deutsch, Jasbi, and Shieber (2020), who fed HAN features to the SVM classifier. BERT performs less competitively on the OneStopEnglish and Newsela corpora. On OneStopEnglish, it is outperformed by the best performing classifier (HAN) by about 10 percentage points, and on Newsela, it is outperformed by about 6 percentage points according to accuracy and F_1 criteria. The most likely reason for the bad performance of BERT on these two corpora is the length of documents in these two data sets. On average, documents in the OneStopEnglish and Newsela corpora are 677 and 760 words long. On the other hand, BERT only allows input documents of up to 512 byte-pair tokens, which means that documents longer than that need to be truncated. This results in the substantial loss of information on the OneStopEnglish and Newsela corpora but not on the WeeBit and Slovenian SB corpora, which contain shorter documents, 193 and 305 words long.

The results show that BiLSTM also has problems when dealing with longer texts, even though it does not require input truncation. This suggests that the loss of context is not the only reason for the non-competitive performance of BERT and BiLSTM, and that the key to the successful classification of long documents is the leveraging of

hierarchical information in the documents, for which HAN was built for. The assumption is that this is particularly important in parallel corpora, where the simplified versions of the original texts contain the same message as the original texts, which forces the classifiers not to rely as much on semantic differences but rather focus on structural differences.

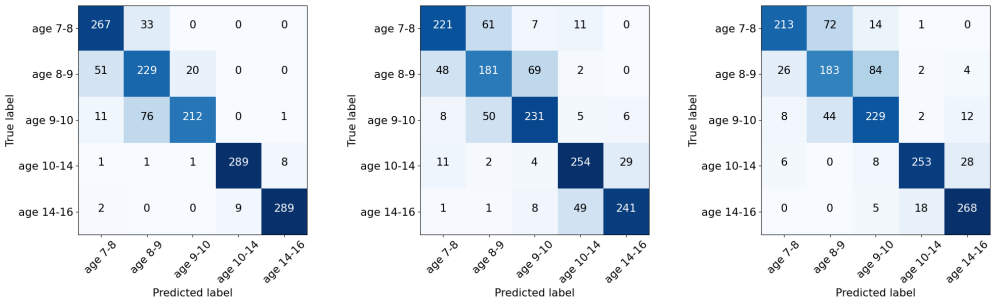
While F_1-scores and accuracies suggest large discrepancies in performance between HAN and two other classifiers on the OneStopEnglish and Newsela corpora, QWK scores draw a different picture. Although the discrepancy is still large on OneStopEnglish, all classifiers achieve almost perfect QWK scores on the Newsela data set. This suggests that even though BERT and BiLSTM make more classification mistakes than HAN, these mistakes are seldom costly on the Newsela corpus (i.e., documents are classified into neighboring classes of the correct readability class). QWK scores achieved on the Newsela corpus by all classifiers are also much higher than the scores achieved on other corpora (except for the QWK score achieved by BERT on the WeeBit corpus). This is in line with the results in the unsupervised setting, where the ρ values on the Newsela corpus were substantially larger than on other corpora.

The HAN classifier achieves the best performance on the OneStopEnglish corpus with an accuracy of 78.72% in the five-fold crossvalidation setting. This is comparable to the state-of-the-art accuracy of 78.13% achieved by Vajjala and Lučić (2018) with their SMO classifier using 155 hand-crafted features. BiLSTM and BERT classifiers perform similarly on this corpus, by about 10 percentage points worse than HAN, according to accuracy, F_1-score, and QWK.

The results on the Slovenian SB corpus are also interesting. In general, the performance of classifiers is the worst on this corpus, with the F_1-score of 52.19% achieved by BiLSTM being the best result. BiLSTM performs by about 4 percentage points better than HAN according to F_1-score and accuracy, while both classifiers achieve roughly the same QWK score of about 80%. On the other hand, BERT achieves lower F_1-score (about 45.45%) and accuracy (41.57%), but performs much better than the other two classifiers according to QWK, achieving QWK of almost 90%.

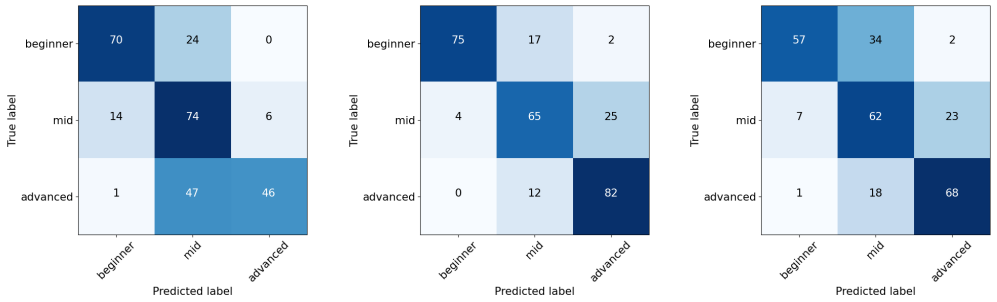
Confusion matrices for classifiers give us a better insight into what kind of mistakes are specific for different classifiers. For the WeeBit corpus, confusion matrices show (Figure 2) that all the tested classifiers have the most problems distinguishing between texts for children 8–9 years old and 9–10 years old. The mistakes where the text is falsely classified into an age group that is not neighboring the correct age group are rare. For example, the best performing BERT classifier misclassified only 16 documents into non-neighboring classes. When it comes to distinguishing between neighboring classes, the easiest distinction for the classifiers was the distinction between texts for children 9–10 years old and 10–14 years old. Besides fitting into two distinct age groups, the documents in these two classes also belong to two different sources (texts for children 9–10 years old consist of articles from WeeklyReader and texts for children 10–14 years old consist of articles from BBC-Bitesize), which suggests that the semantic and writing style dissimilarities between these two neighboring classes might be larger than for other neighboring classes, and that might have a positive effect on the performance of the classifiers.

On the OneStopEnglish corpus (Figure 3), the BERT classifier, which performs the worst on this corpus according to all criteria but precision, had the most problems correctly classifying documents from the advanced class, misclassifying about half of the documents. HAN had the most problems with distinguishing documents from the advanced and intermediate class, while the BiLSTM classifier classified a disproportionate amount of intermediate documents into the beginner class.



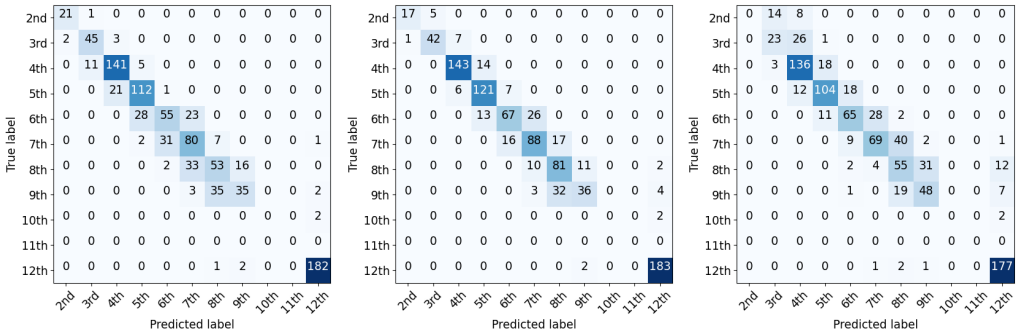
a) BERT b) HAN c) BiLSTM

Figure 2
Confusion matrices for BERT, HAN, and BiLSTM on the WeeBit corpus.



a) BERT b) HAN c) BiLSTM

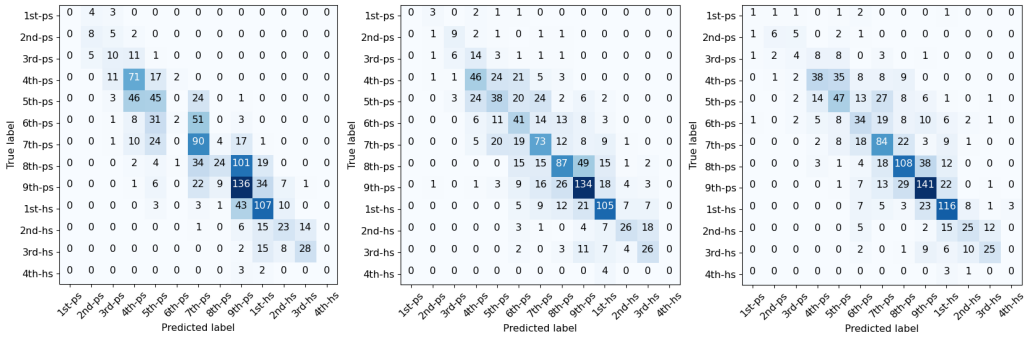
Figure 3
Confusion matrices for BERT, HAN, and BiLSTM on the OneStopEnglish corpus.



a) BERT b) HAN c) BiLSTM

Figure 4
Confusion matrices for BERT, HAN, and BiLSTM on the Newsela corpus.

Confusion matrices of all classifiers for the Newsela corpus (Figure 4) follow a similar pattern. Unsurprisingly, no classifier predicted any documents to be in two minority classes (10th and 11th grade) with minimal training examples. As the QWK score has already shown, all classifiers classified a large majority of misclassified instances into



a) BERT b) HAN c) BiLSTM

Figure 5
Confusion matrices for BERT, HAN, and BiLSTM on the Slovenian school books corpus.

neighboring classes, and costlier mistakes are rare. For example, the best performing HAN classifier altogether misclassified only 13 examples into non-neighboring classes.

Confusion matrices for the Slovenian SB corpus (Figure 5) are similar for all classifiers. The biggest spread of misclassified documents is visible for the classes in the middle of the readability range (from the 4th-grade of primary school to the 1st-grade of high school). The mistakes, which cause BERT to have lower F_1-score and accuracy scores than the other two classifiers, are most likely connected to the misclassification of all but two documents belonging to the school books for the 6th class of the primary school. Nevertheless, a large majority of these documents were misclassified into two neighboring classes, which explains the high QWK score achieved by the classifier. What negatively affected the QWK scores for HAN and BiLSTM is that the frequency of making costlier mistakes of classifying documents several grades above or below the correct grade is slightly higher for them than for BERT. Nevertheless, even though F_1-score results are relatively low on this data set for all classifiers (BiLSTM achieved the best F_1-score of 52.19%), the QWK scores around or above 80% and the confusion matrices clearly show that a large majority of misclassified examples were put into classes close to the correct one, suggesting that classification approaches to readability prediction can also be reliably used for Slovenian.

Overall, the classification results suggest that neural networks are a viable option for the supervised readability prediction. Some of the proposed neural approaches managed to outperform state-of-the-art machine learning classifiers that leverage feature engineering (Xia, Kochmar, and Briscoe 2016; Vajjala and Lučić 2018; Deutsch, Jasbi, and Shieber 2020) on all corpora where comparisons are available. However, the gains are not substantial, and the choice of an appropriate architecture depends on the properties of the specific data set.

6. Conclusion

We presented a set of novel unsupervised and supervised approaches for determining the readability of documents using deep neural networks. We tested them on several manually labeled English and Slovenian corpora. We argue that deep neural networks are a viable option both for supervised and unsupervised readability prediction and show that the suitability of a specific architecture for the readability task depends on the data set specifics.

We demonstrate that neural language models can be successfully used in the unsupervised setting, since they, in contrast to *n*-gram models, capture high-level textual properties and can successfully leverage rich semantic information obtained from the training data set. However, the results of this study suggest that unsupervised approaches to readability prediction that only take these properties of text into account cannot compete with the shallow lexical sophistication indicators. This is somewhat in line with the findings of the study by Todirascu et al. (2016), who also acknowledged the supremacy of shallow lexical indicators when compared with higher-level discourse features. Nevertheless, combining the components of both neural and traditional readability indicators into the new RSRS (ranked sentence readability score) measure does improve the correlation with human readability scores.

We argue that the RSRS measure is adaptable, robust, and transferable across languages. The results of the unsupervised experiments show the influence of the language model training set on the performance of the measure. While the results indicate that an exact match between the genres of the train and test sets is not necessary, the text complexity of a train set (i.e., its readability), should be in the lower or middle part of the readability spectrum of the test set for the optimal performance of the measure. This indicates that out of the two high-level text properties that the RSRS measure uses for determining readability, semantic information and long-distance structural information, the latter seems to have more effect on the performance. This is further confirmed by the results of using the general BERT language model for the readability prediction, which show a negative correlation between the language model perplexity and readability, even though the semantic information the model possesses is extensive due to the large training set.

The functioning of the proposed RSRS measure can be customized and influenced by choice of the training set. This is the desired property because it enables personalization and localization of the readability measure according to the educational needs, language, and topic. The usability of this feature might be limited for under-resourced languages because a sufficient amount of documents needed to train a language model that can be used for the task of readability prediction in a specific customized setting might not be available. On the other hand, our experiments on the Slovenian language show that a relatively small 2.4 million word training corpus for language models is sufficient to outperform traditional readability measures.

The results of the unsupervised approach to readability prediction on the corpus of Slovenian school books are not entirely consistent with the results reported by the previous Slovenian readability study (Škvorc et al. 2019), where the authors reported that simple indicators of readability, such as average sentence length, performed quite well. Our results show that the average sentence length performs very competitively on English but ranks badly on Slovenian. This inconsistency in results might be explained by the difference in corpora used for the evaluation of our approaches. Whereas Škvorc et al. (2019) conducted experiments on a corpus of magazines for different age groups (which we used for language model training), our experiments were conducted on a corpus of school books, which contains items for sixteen distinct school subjects with very different topics ranging from literature, music, and history to math, biology, and chemistry. As was already shown in Sheehan, Flor, and Napolitano (2013), the variance in genres and covered topics has an important effect on the ranking and performance of different readability measures. Further experiments on other Slovenian data sets are required to confirm this hypothesis.

In the supervised approach to determining readability, we show that the proposed neural classifiers can either outperform or at least compare with state-of-the-art

approaches leveraging extensive feature engineering as well as previously used neural models on all corpora where comparison data is available. While the improved performance and elimination of work required for manual feature engineering are desirable, on the downside, neural approaches tend to decrease the interpretability and explainability of the readability prediction. Interpretability and explainability are especially important for educational applications (Sheehan et al. 2014; Madnani and Cahill 2018), where the users of such technology (educators, teachers, researchers, etc.) often need to understand what causes one text to be judged as more readable than the other and according to which dimensions. Therefore in the future, we will explore the possibilities of explaining the readability predictions of the proposed neural classifier with the help of general explanation techniques such as SHAP (Lundberg and Lee 2017), or the attention mechanism (Vaswani et al. 2017), which can be analyzed and visualized and can offer valuable insights into inner workings of the system.

Another issue worth discussing is the trade-off between performance gains we can achieve by employing computationally demanding neural networks on the one side and the elimination of work on the other. For example, on the OneStopEnglish corpus, we report the accuracy of 78.72% when HAN is used, while Vajjala and Lučić (2018) report an accuracy of 78.13% with their classifier employing 155 hand-crafted features. While it might be worth opting for a neural network in order to avoid extensive manual feature engineering, on the other hand, the same study by Vajjala and Lučić (2018) also reports that just by employing generic text classification features, 2–5 character *n*-grams, they obtained the accuracy of 77.25%. Considering this, one might argue that, depending on the use case, it might not be worth dedicating significantly more time, work, or computational resources for an improvement of slightly more than 1%, especially if this also decreases the overall interpretability of the prediction.

The performance of different classifiers varies across different corpora. The major factor proved to be the length of documents in the data sets. The HAN architecture, which tends to be well equipped to handle long-distance hierarchical text structures, performs the best on these data sets. On the other hand, in terms of QWK measure, BERT offers significantly better performance on data sets that contain shorter documents, such as WeeBit and Slovenian SB. As was already explained in Section 5.2, a large majority of OneStopEnglish and Newsela documents need to be truncated in order to satisfy the BERT's limitation of 512 byte-pair tokens. Although it is reasonable to assume that the truncation and the consequential loss of information do have a detrimental effect on the performance of the classifier, the extent of this effect is still unclear. The problem of truncation also raises the question of what is the minimum required length of a text for a reliable assessment of readability and if there exists a length threshold, above which having more text does not influence the performance of a classifier in a significant manner. We plan to assess this in future work thoroughly. Another related line of research we plan to pursue in the future is the use of novel algorithms, such as Longformer (Beltagy, Peters, and Cohan 2020) and Linformer (Wang et al. 2020), in which the attention mechanism scales linearly with the sequence length, making it feasible to process documents of thousands of tokens. We will check if applying these two algorithms on the readability data sets with longer documents can further improve the state of the art.

The other main difference between WeeBit and Slovenian SB data sets on the one hand, and Newsela and OneStopEnglish data sets on the other, is that they are not parallel corpora, which means that there can be substantial semantic differences between the readability classes in these two corpora. It seems that pretraining BERT as a language model allows for better exploitation of these differences, which leads to better

performance. However, this reliance on semantic information might badly affect the performance of transfer learning based models on parallel corpora, since the semantic differences between classes in these corpora are much more subtle. We plan to assess the influence of available semantic information on the performance of different classification models in the future.

The differences in performance between classifiers on different corpora suggest that tested classifiers take different types of information into account. Provided that this hypothesis is correct, some gains in performance might be achieved if the classifiers are combined. We plan to test a neural ensemble approach for the task of predicting readability in the future.

While this study mostly focused on multilingual and multi-genre readability prediction, in the future, we also plan to test the cross-corpus, cross-genre, and cross-language transferability of the proposed supervised and unsupervised approaches. This requires new readability data sets for different languages and genres that are currently rare or not publicly available. On the other hand, this type of research will be capable of further determining the role of genre in the readability prediction and might open an opportunity to improve the proposed unsupervised readability score further.

Acknowledgments

The research was financially supported by the European Social Fund and Republic of Slovenia, Ministry of Education, Science, and Sport through project Quality of Slovene Textbooks (KaUČ). The work was also supported by the Slovenian Research Agency (ARRS) through core research programs P6-0411 and P2-0103, and the projects Terminology and Knowledge Frames Across Languages (J6-9372) and Quantitative and Qualitative Analysis of the Unregulated Corporate Financial Reporting (J5-2554). This work has also received funding from the European Union's Horizon 2020 Research and Innovation program under grant agreement no. 825153 (EMBEDDIA). The results of this publication reflect only the authors' views, and the EC is not responsible for any use that may be made of the information it contains.

References

Anderson, Jonathan. 1981. Analysing the readability of English and non-English texts in the classroom with LIX. In *Seventh Australian Reading Association Conference*, pages 1–12, Darwin.

Azpiazu, Ion Madrazo and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436. DOI: https://doi.org/10.1162/tacl_a_00278

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bai, Shaojie, J. Zico Kolter, and Vladlen Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Beltagy, Iz, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Bengio, Yoshua, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166. DOI: https://doi.org/10.1109/72.279181, PMID: 18267787

Bird, Steven and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*, page 31, Barcelona. DOI: https://doi.org/10.3115/1219044.1219075

Bormuth, John R. 1969. *Development of Readability Analysis*. ERIC Clearinghouse.

Cho, Kyunghyun, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha. DOI: https://doi.org/10.3115/v1/D14-1179

Collins-Thompson, Kevyn. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135. DOI: https://doi.org/10.1075/itl.165.2.01col

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Conneau, Alexis, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen. DOI: https://doi.org/10.18653/v1/D17-1070

Council of Europe, Council for Cultural Co-operation. Education Committee. Modern Languages Division. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.

Crossley, Scott A., Stephen Skalicky, Mihai Dascalu, Danielle S. McNamara, and Kristopher Kyle. 2017. Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, 54(5-6):340–359. DOI: https://doi.org/10.1080/0163853X.2017.1296264

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

Davison, Alice and Robert N. Kantor. 1982. On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly*, pages 187–209. DOI: https://doi.org/10.2307/747483

Deutsch, Tovly, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. *arXiv preprint arXiv:2006.00377*. DOI: https://doi.org/10.18653/v1/2020.bea-1.1

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, MN.

Feng, Lijun. 2008. Text simplification: A survey, The City University of New York.

Feng, Lijun, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 229–237, Athens. DOI: https://doi.org/10.3115/1609067.1609092

Feng, Lijun, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *COLING 2010: Posters*, pages 276–284, Beijing.

Filighera, Anna, Tim Steuer, and Christoph Rensing. 2019. Automatic text difficulty estimation using embeddings and neural networks. In *European Conference on Technology Enhanced Learning*, pages 335–348, Delft. DOI: https://doi.org/10.1007/978-3-030-29736-7_25

Flor, Michael, Beata Beigman Klebanov, and Kathleen M. Sheehan. 2013. Lexical tightness and text complexity. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, pages 29–38, Atlanta, GA.

Goldberg, Yoav and Jon Orwant. 2013. A data set of syntactic-ngrams over time from a very large corpus of English books. In *Second Joint Conference on Lexical and Computational Semantics*, pages 241–247, Atlanta, GA.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.

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

Halliday, Michael Alexander Kirkwood and Ruqaiya Hasan. 1976. *Cohesion in English*. Routledge.

Jawahar, Ganesh, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. DOI: https://doi.org/10.18653/v1/P19-1356

Jiang, Birong, Endong Xun, and Jianzhong Qi. 2015. A domain independent approach for extracting terms from research papers. In *Australasian Database Conference*, pages 155–166, Melbourne. DOI: https://doi.org/10.1007/978-3-319-19548-3_13

Kandel, Lilian and Abraham Moles. 1958. Application de l'indice de flesch à la langue française. *Cahiers Etudes de Radio-Télévision*, 19(1958):253–274.

Kim, Yoon, Yacine Jernite, David Sontag, and Alexander M. Rush. 2016. Character-aware neural language models. In *AAAI*, pages 2741–2749, Phoenix, AZ.

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*. Institute for Simulation and Training, University of Central Florida.

Kudo, Taku and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*. DOI: https://doi.org/10.18653/v1/D18-2012, PMID: 29382465

Landauer, Thomas K. 2011. Pearson's text complexity measure, Pearson.

Logar, 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*. Trojina, zavod za uporabno slovenistiko.

Lundberg, Scott M. and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4768–4777, Long Beach, CA.

Ma, Yi, Eric Fosler-Lussier, and Robert Lofthus. 2012. Ranking-based readability assessment for early primary children's literature. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 548–552, Montreal.

Madnani, Nitin and Aoife Cahill. 2018. Automated scoring: Beyond natural language processing. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1099–1109, Santa Fe, NM.

Madrazo Azpiazu, Ion and Maria Soledad Pera. 2020. Is crosslingual readability assessment possible? *Journal of the Association for Information Science and Technology*, 71(6):644–656. DOI: https://doi.org/10.1002/asi.24293

McLaughlin, G. Harry. 1969. SMOG grading—a new readability formula. *Journal of Reading*, 12(8):639–646.

Mikolov, Tomáš, Anoop Deoras, Stefan Kombrink, Lukáš Burget, and Jan Černockỳ. 2011. Empirical evaluation and combination of advanced language modeling techniques. In *Twelfth Annual Conference of the International Speech Communication Association*, pages 605–608.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, Florence.

Mohammadi, Hamid and Seyed Hossein Khasteh. 2019. Text as environment: A deep reinforcement learning text readability assessment model. *arXiv preprint arXiv:1912.05957*.

Nadeem, Farah and Mari Ostendorf. 2018. Estimating linguistic complexity for science texts. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 45–55, New Orleans, LA. DOI: https://doi.org/10.18653/v1/W18-0505

Napolitano, Diane, Kathleen M. Sheehan, and Robert Mundkowsky. 2015. Online readability and text complexity analysis with TextEvaluator. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 96–100, Denver, CO. DOI: https://doi.org/10.3115/v1/N15-3020

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, Doha. DOI: https://doi.org/10.3115/v1/D14-1162

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237, New Orleans, LA. DOI: https://doi.org/10.18653/v1/N18-1202

Petersen, Sarah E. and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer Speech & Language*, 23(1):89–106. DOI: https://doi.org/10.1016/j.csl.2008.04.003

Pitler, Emily and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, HI. DOI: https://doi.org/10.3115/1613715.1613742

Schwarm, Sarah E. and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 523–530, Ann Arbor, MI. DOI: https://doi.org/10.3115/1219840.1219905

Sheehan, Kathleen M., Michael Flor, and Diane Napolitano. 2013. A two-stage approach for generating unbiased estimates of text complexity. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, pages 49–58, Atlanta, GA.

Sheehan, Kathleen M., Irene Kostin, Yoko Futagi, and Michael Flor. 2010. Generating automated text complexity classifications that are aligned with targeted text complexity standards. *ETS Research Report Series*, 2010(2):i–44. DOI: https://doi.org/10.1002/j.2333-8504.2010.tb02235.x

Sheehan, Kathleen M., Irene Kostin, Diane Napolitano, and Michael Flor. 2014. The TextEvaluator tool: Helping teachers and test developers select texts for use in instruction and assessment. *The Elementary School Journal*, 115(2):184–209. DOI: https://doi.org/10.1086/678294

Škvorc, Tadej, Simon Krek, Senja Pollak, Špela Arhar Holdt, and Marko Robnik-Šikonja. 2019. Predicting Slovene text complexity using readability measures. *Contributions to Contemporary History (Spec. Issue on Digital Humanities and Language Technologies*, 59(1):198–220. DOI: https://doi.org/10.51663/pnz.59.1.10

Smith, Edgar A. and R. J. Senter. 1967. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories (US)*, pages 1–14.

Todirascu, Amalia, Thomas François, Delphine Bernhard, Núria Gala, and Anne-Laure Ligozat. 2016. Are cohesive features relevant for text readability evaluation? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 987–997, Osaka.

Ulčar, Matej and Marko Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT. In *International Conference on Text, Speech, and Dialogue*, pages 104–111, Brno.

Vajjala, Sowmya and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, LA.

Vajjala, Sowmya and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173, Montreal.

Van Dijk, Teun Adrianus. 1977. *Text and Context: Explorations in the Semantics and Pragmatics of Discourse*. Longman London.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, Long Beach, CA.

Wang, Sinong, Belinda Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*.

Wang, Ziyu, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. 2016. Dueling network architectures for deep reinforcement learning. In *International Conference on Machine Learning*, pages 1995–2003, New York.

Williams, Geoffrey. 2006. Michael Hoey. Lexical priming: A new theory of words and language. *International Journal of Lexicography*, 19(3):327–335. DOI: https://doi.org/10.1093/ijl/ecl017

Xia, Menglin, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. DOI: https://doi.org/10.18653/v1/W16-0502, PMCID: PMC4879617

Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, Lille.

Xu, Wei, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association of Computational Linguistics*, 3(1):283–297. DOI: https://doi.org/10.1162/tacl_a_00139

Yang, Zichao, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, CA. DOI: https://doi.org/10.18653/v1/N16-1174

Zhang, Xiang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, pages 649–657, Montreal.

Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 19–27, Santiago. DOI: https://doi.org/10.1109/ICCV.2015.11

Zwitter Vitez, Ana. 2014. Ugotavljanje avtorstva besedil: primer "trenirkarjev." In *Language Technologies: Proceedings of the 17th International Multiconference Information Society - IS 2014*, pages 131–134, Ljubljana.