An Open Multilingual System for Scoring Readability of Wikipedia

Mykola Trokhymovych
Pompeu Fabra University
mykola.trokhymovych@upf.edu

Indira Sen
University of Konstanz
indira.sen@uni-konstanz.de

Martin Gerlach
Wikimedia Foundation
mgerlach@wikimedia.org

Abstract

With over 60M articles, Wikipedia has become the largest platform for open and freely accessible knowledge. While it has more than 15B monthly visits, its content is believed to be inaccessible to many readers due to the lack of readability of its text. However, previous investigations of the readability of Wikipedia have been restricted to English only, and there are currently no systems supporting the automatic readability assessment of the 300+ languages in Wikipedia. To bridge this gap, we develop a multilingual model to score the readability of Wikipedia articles. To train and evaluate this model, we create a novel multilingual dataset spanning 14 languages, by matching articles from Wikipedia to simplified Wikipedia and online children encyclopedias. We show that our model performs well in a zero-shot scenario, yielding a ranking accuracy of more than 80% across 14 languages and improving upon previous benchmarks. These results demonstrate the applicability of the model at scale for languages in which there is no ground-truth data available for model fine-tuning. Furthermore, we provide the first overview on the state of readability in Wikipedia beyond English.

1 Introduction

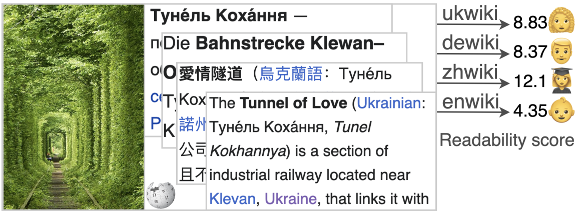
Туне́ль Коха́ння —
Die **Bahnstrecke Klewan–**
愛情隧道（烏克蘭語：Туне́ль
The **Tunnel of Love** (Ukrainian: Туне́ль Коха́ння, *Tunel Kokhannya*) is a section of industrial railway located near Klevan, Ukraine, that links it with
ukwiki
8.83
dewiki
8.37
zhwiki
12.1
enwiki
4.35
Readability score

Figure 1: Sketch of the readability scoring system for Wikipedia articles. Higher scores indicate more difficult-to-read text.

The concept of readability aims to capture how easy it is to read a given text, usually defined as the sum of all factors that affect a reader's understanding, reading speed, and level of interest (Dale and Chall, 1949). In practice, the goal is often to model and quantify the readability of a text on a pre-defined scale using linguistic features, known as Automatic Readability Assessment (ARA) (Vajjala, 2022). Common approaches are based on readability formulas such as the Flesch-Kincaid score (Kincaid et al., 1975), with a recent shift towards more complex computational models leveraging progress on language models in NLP (François, 2015). These readability scores are used to better serve readers' information needs in educational contexts for choosing appropriate reading materials to support, e.g., language learners (Xia et al., 2016) or readers with learning disabilities (Rello et al., 2012). Assessing the accessibility of content in terms of readability is also of relevance more broadly, as general textual information found on the web or in the news is often linguistically too complex for large fractions of the population (Stajner, 2021).

This use case for ARA is particularly relevant for Wikipedia, which has become the largest platform for open and freely accessible knowledge, read by millions of people worldwide with more than 60M articles across 300+ language versions (Wikistats). Unfortunately, this knowledge is believed to remain inaccessible to many readers because the text is written at a level above their reading ability – denoted as the readability gap (Redi et al., 2020). In fact, studies on English Wikipedia have concluded that "overall readability is poor" (Lucassen et al., 2012).

However, the state of readability in Wikipedia beyond English is unknown. Despite recent advances in ARA, there is no currently available system to systematically score articles across many languages due to several challenges (see also Vajjala (2022)). There is a lack of availability of ready-to-use multilingual approaches, as existing web interfaces such as Translated or Readable, support only a limited number of languages. At the

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6296–6311
August 11-16, 2024 ©2024 Association for Computational Linguistics

same time, there are no established readability formulas, such as the Flesch Reading Ease Formula, for most languages beyond English. Furthermore, models (or formulas) for ARA are often designed only for individual or pairs of languages, which makes it challenging to adapt existing models because *(i)* they are difficult to scale to hundreds of languages, and *(ii)* resulting scores cannot be easily compared across languages (Martinc et al., 2021). Furthermore, there is a general lack of ground-truth data. While there are many resources for English, the datasets are often small in size and some of the most commonly-used ones are not available under an open license (such as Newsela (Xu et al., 2015) or WeeBit (Vajjala and Meurers, 2012)), severely limiting their use in real-world applications. Beyond English, resources are scarce and scattered, such that there are no ready-to-use datasets in most languages.

In this paper, we develop a multilingual system to score the readability of articles in Wikipedia (see Figure 1). Specifically, we make the following contributions:

1. First, we compile a new multilingual dataset of pairs of encyclopedic articles with different readability levels covering 14 languages and make it publicly available under an open license.[1]

2. Second, we develop a single multilingual model for readability, demonstrating the effectiveness of the zero-shot cross-lingual transfer.[2]

3. Third, we apply the model to obtain the first systematic overview of the state of readability of Wikipedia articles beyond English and provide a public API endpoint of the model for use by readers, editors, and researchers.[3]

2 Related work

Traditional approaches Research on how to measure readability dates back to more than 100 years (DuBay, 2007). These early attempts focused mainly on developing vocabulary lists of common words and/or readability formulas, such as Flesch reading ease (Flesch, 1948), SMOG (Mc Laughlin, 1969), or the Dale-Chall readability formula (Dale and Chall, 1948), some of which are still commonly used today. In the past decades, there has been a shift towards computational models using approaches from NLP and machine learning; we point the interested reader to general overviews by Collins-Thompson (2014); François (2015); Vajjala (2022).

Language models for Automatic Readability Assessment (ARA) More recently, ARA has been dominated by approaches using language models based on deep neural networks (Martinc et al., 2021). A wide variety of architectures have been proposed based on, among others, word embeddings (Filighera et al., 2019), multiattentive recurrent neural networks (Azpiazu and Pera, 2019), and increasingly common transformers (Mohammadi and Khasteh, 2019) such as BERT (Devlin et al., 2019). These models have been combined with traditional linguistic features (Deutsch et al., 2020; Imperial, 2021). As an alternative to the common approach of modeling ARA as a classification task, the formulation as a ranking problem has been shown to perform better in cross-corpus and cross-lingual scenarios (Lee and Vajjala, 2022; Miliani et al., 2022). Also, recent work utilizes prompt-based seq2seq models to solve ARA as a text-to-text generative task (Lee and Lee, 2023b).

Multilingual ARA While most research on ARA is focused on English, there have been many efforts in the past years for a broad range of languages such as Arabic (Nassiri et al., 2023), Cebuano (Imperial et al., 2022), Dutch (Hobo et al., 2023), French (Wilkens et al., 2022), German (Blaneck et al., 2022), Greek (Chatzipanagiotidis et al., 2021), Spanish (Vásquez-Rodríguez et al., 2022), or Turkish (Uluslu and Schneider, 2023). However, most of these studies focus only on a single language, with few exceptions attempting to model several languages jointly (Madrazo Azpiazu and Pera, 2020b; Imperial and Kochmar, 2023). Some studies have demonstrated zero-shot cross-lingual transfer for individual pairs of languages (i.e., a model is trained only in one language and then evaluated on another) for English to French (Lee and Vajjala, 2022), Spanish to Catalan (Madrazo Azpiazu and Pera, 2020a), or English to German (Weiss et al., 2021). Our work extends these contributions beyond individual pairs, taking advantage of the more general findings that multilingual transformer models, such as mBERT (Devlin et al., 2019), per-

[1] https://zenodo.org/records/11371932

[2] https://gitlab.wikimedia.org/repos/research/readability-experiments

[3] https://api.wikimedia.org/wiki/Lift_Wing_API/Reference/Get_readability_prediction

form surprisingly well at zero-shot cross-lingual transfer learning for a wide range of tasks outside ARA (Pires et al., 2019).

Readability in Wikipedia There have been efforts to capture readability in Wikipedia, with most studies focusing on English and using traditional readability formulas. A comparison of Simple and English Wikipedia has shown that articles from Simple Wikipedia are easier to read (Yasseri et al., 2012), overall readability is insufficient for its target audience (even for Simple Wikipedia) (Lucassen et al., 2012), and readability in Wikipedia (both English and Simple) lags behind other encyclopedias such as Britannica (Jatowt and Tanaka, 2012). (Den Besten and Dalle, 2014) analyzed the temporal evolution of Simple Wikipedia, showing a gradual decline in readability. Some studies focus specifically on Wikipedia's health-related content finding that most articles remain written at a level above the reading ability of average adults (Reavley et al., 2012; Brezar and Heilman, 2019). Readability has also been discussed as a feature for article quality (Liu et al., 2021; Moás and Lopes, 2023).

3 Data

We generate a new multilingual dataset of document-aligned pairs of encyclopedic articles, where each pair contains the same article in two levels of readability (easy/hard). The pairs are obtained by matching Wikipedia articles (hard) with the corresponding version from different simplified or children's encyclopedias (WMF-a) (easy). The latter encyclopedias are purposefully designed with the goal of creating articles using simpler language (e.g. vocabulary, grammar, and sentence structure) (WMF-b). Additionally, past research has shown that articles from Simple English Wikipedia are easier to read than articles from English Wikipedia, using traditional readability formulas for English (Lucassen et al., 2012).

The proposed approach yields a dataset covering 14 languages summarized in Table 1.

While some of the same sources have already been used previously, e.g. Simple English Wikipedia (Zhu et al., 2010), Klexikon (Aumiller and Gertz, 2022), Vikidia (Lee and Vajjala, 2022), our dataset provides substantial improvements: i) two new data sources (Txikipedia; Wikikids)); ii) 11 new languages (Armenian, Basque, Catalan, Dutch, Greek, Italian, Occitan, Portuguese, Russian, Sicilian, Spanish); iii) improved extraction of the plain text from articles by parsing the HTML version instead of wikitext; iv) the dataset is publicly available under an open license in contrast to some of the most commonly used resources in ARA, such as Newsela (Xu et al., 2015).

Dataset	#Pairs	Avg. #Sen.	Avg. #Char.
simplewiki-en	112,342	6.2/7.9	84.6/130.9
vikidia-en	1,991	6.4/14.3	83.3/142.8
vikidia-ca	234	5.2/9.7	79.3/145.2
vikidia-de	260	6.4/11.2	75.8/131.0
vikidia-el	39	6.0/11.8	96.8/134.9
vikidia-es	1,915	5.7/7.7	109.0/179.4
vikidia-eu	571	6.5/8.7	114.6/129.5
vikidia-fr	12,221	5.7/7.3	106.9/152.1
vikidia-hy	485	14.3/11.4	105.3/115.1
vikidia-it	1,662	4.5/6.0	84.6/152.6
vikidia-oc	12	4.2/7.1	77.0/105.6
vikidia-pt	809	5.7/11.8	97.3/157.9
vikidia-ru	125	5.8/11.2	83.8/110.6
vikidia-scn	10	3.8/4.7	50.9/86.3
klexikon-de	2,255	17.7/8.9	73.9/136.9
txikipedia-eu	1,162	7.3/8.4	107.4/126.4
wikikids-nl	12,090	8.0/7.5	83.7/112.0

Table 1: Dataset summary statistics (easy/hard).

3.1 Dataset sources

Simple Wikipedia is a simplified version of English Wikipedia written in basic and learning English targeted towards children, non-native speakers, and people with learning disabilities. It is a commonly-used resource for large-scale text simplification datasets, such as PWKP (Zhu et al., 2010), SEW (Coster and Kauchak, 2011) WikiLarge (Zhang and Lapata, 2017), WikiAuto (Jiang et al., 2020), DWikipedia (Sun et al., 2021), or SWiPE (Laban et al., 2023).

Txikipedia is a children encyclopedia contained in the Basque Wikipedia. Similar to article talk pages, the children's version of an article is stored under a different namespace and available to readers as a separate tab at the top of the page.

Vikidia, Klexikon, and Wikikids are wiki-based encyclopedias for children independent from the language versions of Wikipedia hosted by the Wikimedia Foundation (WMF). Vikidia exists in 11 different languages. Azpiazu and Pera (2019) considered six of the languages in their experiments, but the article-aligned data is not publicly available. Lee and Vajjala (2022) compiled data for English and French. Klexikon is available in German and was utilized by Aumiller and Gertz (2022) to create text simplification datasets, and Wikikids is available in Dutch.

3.2 Pre-processing

We match articles from Wikipedia with the corresponding article in the simplified/children encyclopedia either via the Wikidata item id or their page titles. We extract the text of each article directly from their parsed HTML version instead of using the original wikitext (Wikipedia), the markup language in which Wikipedia is edited. Previous studies show that this approach provides a more accurate representation of the content of Wikipedia articles as seen by its readers (Mitrevski et al., 2020). For example, using wikitext as a source often results in missing important information.[4] In order to limit systematic differences in length, we only consider the text from the first (lead) section. We only keep pairs of articles in which both versions of the text have three or more sentences. For more details about data processing, see Appendix B.

4 Model

4.1 Design requirements

In this section, we describe the requirements for the system to score the readability of Wikipedia articles across languages, as they influence the architecture of the underlying model.

First, our aim is to score the readability of articles in Wikipedia on a continuous scale. Typically, ARA is modeled as a classification problem in NLP research with the goal to predict the labels of the ground-truth data corresponding to a few readability levels (e.g., five in Newsela or three in OneStopEnglish). In our case, the intended use-case is not to predict the label corresponding to the article's source (Wikipedia or simplified/children encyclopedia) but to score articles on a more fine-grained scale.

Second, we aim to develop a single multilingual model. This will not only allow us to compare scores across different languages, but also reduce infrastructure costs related to scaling and maintaining the model for many languages.

Third, the model should require no or only little language-specific fine-tuning (i.e. zero-shot or few-shot scenario) as there is little to no annotated ground-truth data on readability available for almost all of the 300+ languages in Wikipedia.

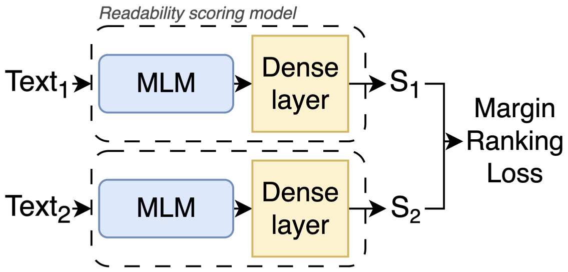

Figure 2: Sketch of the model architecture consisting of two joint readability scoring models trained using a Margin Ranking Loss. S_1 and S_2 refer to the predicted scores of $Text_1$ and $Text_2$, respectively.

4.2 Model architecture

We train a ranking-based model that can perform text scoring of individual texts during inference (see Figure 2 for a sketch of the training procedure). For this, we adapt the Neural Pairwise Ranking Model (NPRM) introduced by Lee and Vajjala (2022) to overcome its main limitation with respect to scoring the readability of articles, i.e. that the model requires the input of at least two texts and only provides a relative ranking as output.

We first build a readability scoring model composed of a multilingual Masked Language Model (MLM) (Devlin et al., 2019) and a Dense layer. The MLM takes the text as input and encodes it into a numerical representation. The Dense Layer then performs a linear transformation into a single real number, which serves as a readability score. We then use a Siamese architecture (Bromley et al., 1993) composed of two joint readability scoring models that share their weights. We apply a Margin Ranking Loss (MRL), also known as Pairwise Hinge Loss, that is commonly used for ranking models, such as Ranking SVM (Herbrich et al., 1999), given by:

$$\mathrm{MRL}(S_1, S_2, y) = \max(0, -y(S_1 - S_2) + \mathrm{m}),$$

where m is a hyperparameter. The MRL is a function of a pair of the same text annotated with readability levels, where S_1 and S_2 are the predicted scores of the texts and $y = -1$ or 1 depending on whether the first or second text should have a lower or higher score based on the annotation. During inference, we pass each individual text to the readability scoring model to obtain the score that is used for readability assessment.

[4] https://huggingface.co/datasets/wikimedia/wikipedia/discussions/51

4.3 Fine-tuning strategy

We fine-tune the model for ARA using the dataset consisting of pairs of encyclopedic articles available in two readability levels (Sec. 3). More precisely, we use only the *simplewiki-en* dataset for fine-tuning, splitting the data randomly into a training (80%) and testing (20%) dataset. All other datasets are only used for testing.

This is motivated by the fact that the main goal is to apply the model without language-specific fine-tuning. The *simplewiki-en* dataset is by far the largest available annotated dataset for ARA, providing a large volume and wide spectrum of training examples. Using a multilingual MLM as the backbone of our model, which has been pretrained in an unsupervised setting in more than 100 languages, we expect zero-shot cross-lingual transfer learning (Pires et al., 2019).

4.4 Technical implementation

We utilize the transformers package (Wolf et al., 2020) to fine-tune the model. We implement the ranking-based readability model in two different flavors.

The text-based model (TRank) takes as input the full text of each article at once. For this, we use the *xlm-roberta-longformer-base* (Sagen, 2021) as a backbone, as it allows us to process long inputs of up to 4096 tokens.

As an alternative, we also implement a sentence-based version (SRank), where the text is split into sentences that are passed to the model sequentially, and as a result, leading to much smaller input lengths. For this, we use the *bert-base-multilingual-cased* (Devlin et al., 2019) as a backbone. The main motivation for adding SRank experiments alongside the TRank model is that the model is smaller (i.e. requiring fewer computational resources during inference) and that it can, in principle, process articles of any length without truncation (addressing a key limitation of the TRank model).

Details about the hyperparameters and computational resources can be found in Appendix C.

5 Experimental evaluation

5.1 Task setup

Ranking task We evaluate the model in a pair-wise ranking task following the approach by (Lee and Vajjala, 2022). That is, for each pair of articles, we assume that the model's readability score should be lower for the easy text (from the simplified/children encyclopedia) than for the hard text (from Wikipedia). This approach has several advantages over a binary classification task aiming to predict the readability label: *(i)* we take advantage of document-level alignment of the same text in different readability levels instead of predicting labels of individual documents; *(ii)* pair-wise ranking directly evaluates the model's readability scores by checking whether the easy version receives a lower score. As an evaluation metric, we use ranking accuracy (RA), that is the percentage of pairs that are ranked correctly. We use bootstrapping to compute confidence intervals (see Appendix D).

Baselines In order to evaluate our model's performance, we consider a set of strong baselines that are representative of the most common approaches.

NS constitutes a naive baseline that calculates the number of sentences in each text.

FRE calculates the Flesch reading ease of the text. Using TextStat, we obtain the language-specific reading ease formula when available and English-specific otherwise.

LFR and LFC constitute a ranker and classifier, respectively, which are based on linguistic features. We use the LFTK tool (Lee and Lee, 2023a) to extract all language-agnostic features (via the parameter *language="general"*). Using these features, we then train a ranker and classifier model using CatBoost (Dorogush et al., 2017) with default parameters for 5K iterations with a learning rate of 0.01.

LMC uses a multilingual MLM for a classification-based approach (instead of ranking-based). We fine-tune the *bert-base-multilingual-cased* model Devlin et al. (2019), splitting the texts into sentences and applying mean pooling to get aggregated readability scores.

5.2 Results

Multilingual benchmark We evaluate the performance of our model and the baselines on our new multilingual benchmark dataset (Table 2). Overall, our model substantially outperforms all baselines, yielding an RA above 0.8 across all datasets, and generally with TRank being slightly better than the SRank model. In the following, we inspect these results in more detail.

First, we observe that our model (TRank) yields an almost perfect RA of 0.976 on the test set of *simplewiki-en*. Taking into account that we fine-

Dataset	NS	±CI	FRE	±CI	LFC	±CI	LFR	±CI	LMC	±CI	TRank	±CI	SRank	±CI
simplewiki-en	0.543	0.007	0.868	0.005	0.937	0.003	0.945	0.003	0.965	0.002	**0.976**	0.002	0.972	0.002
vikidia-en	0.814	0.017	0.935	0.011	0.979	0.006	0.981	0.006	0.979	0.006	**0.991**	0.004	0.985	0.005
vikidia-ca	0.782	0.054	0.906	0.038	0.94	0.031	0.932	0.033	0.936	0.032	**0.962**	0.025	0.936	0.032
vikidia-de	0.735	0.054	0.815	0.048	0.888	0.039	0.869	0.042	0.908	0.036	**0.938**	0.03	0.919	0.034
vikidia-el	0.718	0.144	0.718	0.144	0.744	0.14	0.795	0.129	0.897	0.096	**0.923**	0.086	0.897	0.097
vikidia-es	0.573	0.023	0.842	0.017	0.883	0.015	0.892	0.014	0.879	0.015	**0.911**	0.013	0.909	0.013
vikidia-eu	0.541	0.042	0.673	0.04	0.639	0.04	0.623	0.041	0.63	0.04	**0.818**	0.032	0.736	0.037
vikidia-fr	0.553	0.009	0.84	0.007	0.82	0.007	0.845	0.006	0.849	0.007	**0.923**	0.005	0.918	0.005
vikidia-hy	0.394	0.045	0.594	0.045	0.534	0.045	0.598	0.044	0.637	0.044	**0.802**	0.036	0.761	0.039
vikidia-it	0.569	0.024	0.83	0.019	0.919	0.013	0.94	0.012	0.925	0.013	**0.958**	0.01	0.952	0.01
vikidia-oc	0.667	0.273	0.667	0.271	0.75	0.25	0.667	0.27	0.917	0.159	**1.0**	0.0	0.917	0.161
vikidia-pt	0.761	0.03	0.869	0.024	0.938	0.017	0.934	0.017	0.921	0.019	**0.960**	0.014	0.938	0.017
vikidia-ru	0.728	0.08	0.608	0.087	0.736	0.078	0.776	0.074	0.736	0.079	**0.880**	0.058	0.760	0.077
vikidia-scn	0.4	0.314	0.6	0.309	0.6	0.308	0.8	0.254	0.6	0.31	0.9	0.191	**1.0**	0.0
klexikon-de	0.114	0.013	0.984	0.005	0.999	0.002	0.995	0.003	0.991	0.004	**0.999**	0.002	0.996	0.003
txikipedia-eu	0.512	0.029	0.707	0.027	0.689	0.027	0.698	0.027	0.67	0.027	**0.81**	0.023	0.762	0.025
wikikids-nl	0.427	0.009	0.795	0.007	0.831	0.007	0.834	0.007	0.85	0.007	0.897	0.006	**0.907**	0.005

Table 2: Ranking accuracy on different test datasets (zero-shot scenario for all datasets except simplewiki-en). Confidence intervals (CI) denote two standard deviations from bootstrapping (Appendix D).

Dataset	NS	±CI	FRE	±CI	LFC	±CI	LFR	±CI	LMC	±CI	TRank	±CI	$NPRM$	±CI
VikidiaEn	0.966	0.005	0.948	0.006	0.888	0.008	0.946	0.006	0.965	0.005	**0.984**	0.003	0.975	0.004
VikidiaFr	0.952	0.005	0.899	0.008	0.878	0.008	0.888	0.008	0.75	0.011	**0.978**	0.004	0.811	0.010
OSE	0.794	0.059	0.915	0.04	0.889	0.046	0.873	0.048	0.942	0.034	**0.974**	0.023	0.878	0.048

Table 3: Ranking accuracy on previous reference datasets (zero-shot scenario). Results of the $NPRM$ model taken from (Lee and Vajjala, 2022). Confidence intervals (CI) denote two standard deviations from bootstrapping (Appendix D).

tuned the model on the corresponding training set, this performance might not be surprising. In fact, many of the baselines yield RA of around 0.9 or higher.

Second, we consider performance on a different dataset not used for training but still in the same language as the training data (*vikidia-en*). We observe that most models yield RA above 0.979, demonstrating that the models generalize well beyond the specific training data.

Third, we consider performance in languages that were not used for fine-tuning the model (zero-shot scenario). Our model (TRank) yields an RA higher than 0.8 for all datasets and higher than 0.9 for 10 out of 15 non-English datasets. Notably, the SRank model performs substantially worse in some of the languages (e.g., Basque). In comparison, the performance of the baseline model varies substantially across languages. The naive *NS* baseline yields generally poor results across most languages. The *FRE* baseline performs well in English, but RA in other languages is substantially lower than for our models. The RA for the *LFC, LFR, LMC* baselines is similar to our models in some cases (e.g. *vikidia-it*) but much lower for others, most notably languages with non-Latin scripts (*vikidia-el*, *vikidia-ru*, *vikidia-hy*).

As the TRank model outperforms SRank across almost all datasets and languages, we will consider only the TRank model for all experiments in the remainder of the paper.

Reference datasets In order to directly compare our results with previous work, we also evaluate our model on three open reference datasets considered in the experiments by Lee and Vajjala (2022), who introduced the $NPRM$ as one of the most recent SOTA approaches in multilingual ARA: Vikidia-En, Vikidia-FR, and OneStopEnglish (OSE) (Vajjala and Lucic, 2018). For the latter dataset with 3 readability levels, RA is the fraction of triples where predicted scores for all three versions are ranked correctly.

In Table 3, we show the performance of our model on the corresponding test datasets in a zero-shot scenario, i.e. without any additional training or fine-tuning. The results show that TRank substantially outperforms not only the baselines but also the $NPRM$, especially in French. Surprisingly, many of the baselines yield high RA (e.g. simplistic features using the number of sentences in VikidiaEn and VikidiaFr). This further highlights the usefulness of our proposed multilingual bench-

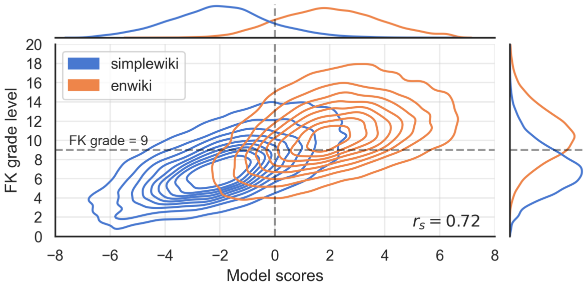

Figure 3: Distribution of model scores vs. FKGL for articles from the test set of simplewiki-en stratified by readability level: simplewiki (easy) and enwiki (hard).

mark, as it seems to constitute a more challenging dataset for ARA tasks.

5.3 Interpreting the model's readability scores

Our model yields a readability score on a continuous scale, which can take positive and negative values and higher scores indicate that the text is more difficult to read.

We find that the readability scores are strongly and statistically significantly correlated (p-value $< 10^{-12}$) with existing readability formulas adapted for different languages. Specifically, we calculate the Spearman rank correlation between the model's readability scores and the language-specific Flesch reading ease (FRE) (TextStat) for articles in the corresponding languages: -0.63 (simplewiki-en) and -0.72 (vikidia-en), -0.67 (vikidia-de) and -0.81 (klexikon-de), -0.67 (vikidia-es), -0.65 (vikidia-fr), -0.76 (vikidia-fr), -0.62 (wikikids-nl), and -0.43 (vikidia-ru).[5] These results demonstrate that the readability scores of our multilingual model correspond to existing and well-founded notions of readability across languages.

Focusing on English, we associate the model's readability score with the Flesch-Kincaid grade level (FKGL) (Kincaid et al., 1975). The main advantage of FKGL is its interpretability in terms of U.S. grade level or, loosely speaking, the number of years of education required to understand a text. Higher grade levels indicate more difficult text. Using articles from simplewiki-en, we find that scores are significantly and strongly correlated ($\rho = 0.72$, p-value $< 10^{-12}$). In Figure 3, we show that, as expected, a model score of ≈ 0 separates texts from simplewiki (easy) and enwiki (hard). We find that this separation corresponds to an FKGL ≈ 9. This allows us to roughly map the model's reference point to a corresponding grade level and interpret the readability of articles that are above or below.

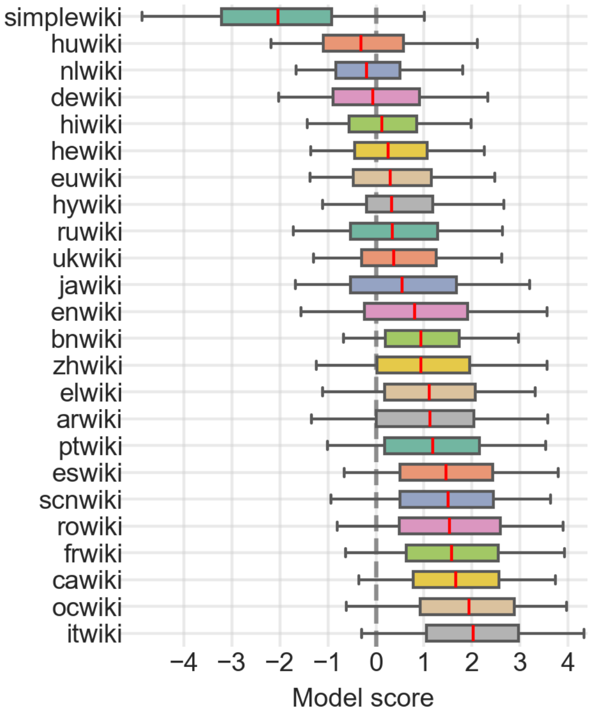

Figure 4: Distribution of readability scores (from the TRank model) across different language editions of Wikipedia. Boxplots show median (red line) and 25- and 75-percentiles with whiskers ranging from 2.5- to 97.5-percentile.

6 Application

6.1 State of readability in Wikipedia

We use the TRank model to get an overview on the state of readability in Wikipedia beyond English. For this analysis, we consider overall 24 different Wikipedias covering all languages from our dataset (Sec.3), in addition to a set of 10 languages considered in a prior study covering different language families and taking into account the number and distribution of speakers worldwide (Lemmerich et al., 2019). For each language, we select a random subsample of 10K articles and extract the text following the same methodology as in Sec. 3.2.

In Figure 4, we show the distribution of readability scores across articles in each language. For English Wikipedia, we observe a median score around 1, with the majority of articles above a score of 0 (roughly corresponding to a Flesch-Kincaid grade level of 9 (Sec. 5.3). For most languages, the distribution of scores is similar or shifted towards much higher values, such as Italian, with a median of

[5] A negative correlation is expected as higher FRE scores indicate easier-to-read texts.

2.02. Some languages show slightly lower scores, such as Hungarian, with a median of -0.32. Only Simple Wikipedia, as expected, shows substantially lower readability scores with a median of -2.04 and the 75-percentile below 0. Overall, this demonstrates that previous findings about the poor overall readability of English Wikipedia (Lucassen et al., 2012) can be generalized to most other language editions of Wikipedia.

6.2 Productionization as a ready-to-use tool

In order to facilitate the use of our model, we provide a public API endpoint to directly access readability scores for articles in Wikipedia from the model. The deployed model is an end-to-end system, including the articles' text collection using the MediaWiki API, processing, and scoring.

We measure efficiency on inference by selecting a random sample of 1K articles and passing them sequentially to the readability model. We observe the median response time of 0.5 seconds, and the $75\%/95\%$ percentiles are $0.83/2.05$ seconds when limiting resources to 1 thread CPU. The same measurements are $0.02/0.03/0.05$ when GPU is enabled. It should be mentioned that inference time is influenced by the length of text, as long articles require more processing by the language model.

7 Discussion

7.1 Summary of findings

We created a new dataset for multilingual ARA with pairs of aligned articles in two readability levels from Wikipedia and a corresponding simplified/children encyclopedia. The advantage of our dataset is that it i) contains new sources (such as Txikipedia) and languages (such as Basque), ii) provides cleaner plain text from processing HTML sources, and iii) is publicly available under an open license.

We developed a new multilingual model for ARA, adapting a ranking-based architecture to score individual texts across languages. We demonstrate that the model performs well in the zero-shot scenario across all languages with a ranking accuracy > 0.8, substantially outperforming all baselines, including traditional readability formulas. This suggests that the model can be applied at scale to languages in which we do not have ground-truth data for additional fine-tuning. We provide additional insights into the interpretability of this model's readability scores by showing that they correlate with hand-crafted readability formulas available for individual languages. In order to ensure reproducibility, we make the code for training and evaluating the model available in a public repository.

We apply our model to get the first state of readability in Wikipedia beyond English. We reproduce previous findings in English Wikipedia: most of the content is written at a reading level above the average (American) adult (Lucassen et al., 2012; Brezar and Heilman, 2019). We find that the readability of most analyzed languages in Wikipedia is at a similar (or even more difficult) level than English Wikipedia. In order to facilitate the use of our model, we provide a publicly available API endpoint to the trained model for researchers and contributors as a ready-to-use tool for ARA in Wikipedia.

7.2 Implications and broader impact

Children encyclopedia communities Our work shows that the simplified and children encyclopedias provide an invaluable resource for multilingual research on readability. More generally, it highlights that these encyclopedias, which are targeted towards specific audiences, play an important role in the open online knowledge ecosystem. However, very little is known about these projects. In order to better understand its content (e.g. quality, reliability, suitability for different readability levels), more research is needed about its audience (who is using it) and contributors (who is creating it) and their motivations.

Improving the state of multilingual ARA Our results improve the state of ARA by directly addressing some of its main limitations according to one of the most recent reviews on the topic (Vajjala, 2022). First, we address the lack of publicly available multilingual corpora by providing a new, high-quality, multilingual dataset under an open license. Second, our model addresses the lack of availability of ready-to-use tools: We not only provide the code in a public repository together with documentation in the form of a model card, but also make available a public API endpoint for users. Third, our results address the lack of well-defined SOTA: We provide reproducible new benchmarks for 14 languages (datasets and training/evaluation code).

The extent of the readability gap in Wikipedia Our tool provides a systematic approach to quantify readability as a knowledge gap in Wikipedia (Redi et al., 2020). We start from the observation in English Wikipedia that the readability scores of the majority of articles exceed a reading level corresponding to a Flesch-Kincaid grade level of 9 (Sec. 5.3). This means that much of its content is not accessible to the larger population in terms of readability when taking into account that the average reading ability of adult Americans is estimated at grade 7-8 (Mcinnes and Haglund, 2011) (matching recommendations for readability levels of public resources such as for patient education material by the U.S. National Institutes of Health (Brezar and Heilman, 2019)). Our results show that these insights can be generalized to many other language versions, as the distribution of articles' readability scores is similar or shifted towards higher difficulty.

More generally, this expands previous research on motivations of readers (Singer et al., 2017; Lemmerich et al., 2019), which has shown that information needs vary substantially with demographics such as gender (Johnson et al., 2021). Measuring the readability of articles describes the suitability of content for readers with different educational and/or literacy backgrounds. In this way, it is possible to identify misalignment between supply (readability of existing content) and demand (education levels of the potential reader population) in Wikipedia, similar to previous studies focusing on information quality (Warncke-Wang et al., 2015).

Text simplification Our model provides a starting point for systematically approaching the task of text simplification in Wikipedia in order to make content more accessible to different audiences. The use of ARA for the automatic evaluation of text simplification approaches (Alva-Manchego et al., 2020) can now be applied across languages. Also, ARA can identify those articles that are the most difficult to read (and thus, the most in need of simplification). Surfacing those to contributors would enable them to make data-informed editorial decisions taking into account readability (WMF-e).

Limitations

We tried two variants of MLMs (SRank and TRank), and found them to have similar performances. Similar larger models with more parameters, such as mLongT5 (Uthus et al., 2023), could yield even better performance. However, the necessary infrastructure (especially in terms of GPUs) required for training and inference makes it challenging to provide the model as a ready-to-use tool.

Multilingual models based on transformer architectures support many languages (e.g., multilingual BERT was trained on 104 languages). However, among those supported, the performance on low-resource languages is still considered unsatisfactory (Wu and Dredze, 2020). More severely, the majority of the more than 300 languages in Wikipedia is still not explicitly represented in the training data of these models. Thus, if unaddressed, the use of such models could lead to a language gap constituting a substantial barrier towards knowledge equity (Redi et al., 2020).

We evaluated our multilingual model on only 14 languages for which we were able to compile a ground truth dataset of encyclopedic articles available at different readability levels. It should be noted that the models were trained only on English texts, so the scores for unseen languages constitute approximations. Additional validation in an applied scenario (Vajjala, 2022), beyond showing statistically significant correlations with commonly-used language-specific readability formulas, would be desirable for future research using, e.g., comprehension tests such as Cloze tests (Redmiles et al., 2019).

Our models and experiments focus only on document-level readability assessment, evaluating the overall readability of entire articles. This approach differs from other forms of ARA that target finer-grained levels, such as sentence-level or phrase-level readability. By concentrating on document-level assessments, we aim to provide a general readability score for Wikipedia articles, though this may overlook variations in readability within smaller sections of text.

Ethics statement

We develop multilingual datasets and models for measuring the readability of Wikipedia articles to better understand the state of readability on Wikipedia, across its many language editions. By pinpointing articles with low readability scores, we support editors and researchers in identifying and addressing these gaps in order to make knowledge on Wikipedia more accessible.

Dataset Quality We contribute a novel and openly available dataset of encyclopedic articles covering different readability levels. It is the largest of its kind and covers 11 more languages compared to past work. By making it available under an open license, we provide a valuable resource for NLP researchers, especially those working on ARA.

Wikipedia articles have been used to create a wide variety of NLP datasets, especially those used to train large language models (Gao et al., 2020). Our dataset consists of encyclopedic articles from Wikipedia and other online encyclopedias. While the dataset contains some metadata about the articles (e.g., link to Wikidata), it does not contain any details about the author(s) of the articles. Therefore, the dataset does not divulge any private information about the articles' authors or readers. Furthermore, some of the online encyclopedias, especially those hosted by WMF, have robust community-driven moderation that ensures that the content is reliable (Yasseri and Menczer, 2023). We also take further steps in processing and filtering (cf. Section 3.2), to improve data quality, a crucial issue in multilingual NLP research (Kreutzer et al., 2022).

In terms of language variety, while we do not have fine-grained information about the editors writing the articles in this dataset, previous research on the demographics of the editors of English Wikipedia have revealed that they are primarily men from North America and Europe (WMF-f). However, Wikipedia is *read* by a more diverse group of people (WMF-g). So, we expect this dataset and the models built using it to be useful for this more diverse global audience.

Intended Use of Models We also develop a model for scoring the readability of Wikipedia articles. Since there are few resources for assessing the readability of non-English text, even less so for low-resourced languages, one of the main strengths of our model is its promising zero-shot cross-lingual transfer capabilities. This model is hosted and publicly deployed so that it can be easily used in an off-the-shelf manner, without investing effort in training the model from scratch. We intend for this model to be used not only by researchers interested in investigating the state of readability in Wikipedia articles across languages but also by other stakeholders such as readers and editors. By assessing the current status of readability in Wikipedia, editors can flag articles needing further simplification.

We do not intend for this dataset and model to be used to increase or reinforce biases, e.g., to discriminate against people whose writing is automatically scored with lower readability scores, or to profile or censor people based on their writing. This model was tested for encyclopedia-like text and might not generalize to other forms of writing, such as academic assignments. In future research, we hope to use the readability scores from our multilingual model in tandem with other metrics of knowledge accessibility, such as visual content and citations, to meet the needs of Wikipedia readers.

Acknowledgements

The work of Mykola Trokhymovych is funded by MCIN/AEI /10.13039/501100011033 under the Maria de Maeztu Units of Excellence Programme (CEX2021-001195-M).

References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational linguistics (Association for Computational Linguistics)*, 46(1):135–187.

Giusepppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

Dennis Aumiller and Michael Gertz. 2022. Klexikon: A German Dataset for Joint Summarization and Simplification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2693–2701, Marseille, France. European Language Resources Association.

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

Patrick Gustav Blaneck, Tobias Bornheim, Niklas Grieger, and Stephan Bialonski. 2022. Automatic readability assessment of German sentences with transformer ensembles. In *Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text*, pages 57–62, Potsdam, Germany. Association for Computational Linguistics.

Aleksandar Brezar and James Heilman. 2019. Readability of English Wikipedia's health information over time. *WikiJournal of Medicine*, 6(1):7.

Jane Bromley, James Bentz, Leon Bottou, Isabelle Guyon, Yann Lecun, Cliff Moore, Eduard Sackinger, and Rookpak Shah. 1993. Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7:25.

Debarshi Chanda. 2021. Jigsaw starter. https://www.kaggle.com/code/debarshichanda/pytorch-w-b-jigsaw-starter/notebook. Accessed on 15 Feb 2024.

Savvas Chatzipanagiotidis, Maria Giagkou, and Detmar Meurers. 2021. Broad linguistic complexity analysis for Greek readability classification. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 48–58, Online. Association for Computational Linguistics.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.

Will Coster and David Kauchak. 2011. Learning to simplify sentences using Wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 1–9, Portland, Oregon. Association for Computational Linguistics.

Edgar Dale and Jeanne Chall. 1948. A formula for predicting readability. *Educational research bulletin*, 27(1):11–28.

Edgar Dale and Jeanne S Chall. 1949. The concept of readability. *Elementary English*, 26(1):19–26.

Matthijs Den Besten and Jean-Michel Dalle. 2014. Keep it simple: A companion for simple wikipedia? In *Online Communities and Open Innovation*, pages 55–64. Routledge.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online. Association for Computational Linguistics.

Jacob Devlin, 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, Minnesota. Association for Computational Linguistics.

Anna Veronika Dorogush, Andrey Gulin, Gleb Gusev, Nikita Kazeev, Liudmila Ostroumova, and Aleksandr Vorobev. 2017. Fighting biases with dynamic boosting. *arXiv:1706.09516*.

William H DuBay. 2007. *Unlocking Language: The Classic Readability Studies*. Impact Information.

Bradley Efron and R.J. Tibshirani. 1994. *An Introduction to the Bootstrap*, 1st edition edition. Chapman and Hall/CRC, New York.

Wikimedia Enterprise. Html dumps. `https://dumps.wikimedia.org/other/enterprise_html/`. Accessed: 2024-2-9.

Anna Filighera, Tim Steuer, and Christoph Rensing. 2019. Automatic text difficulty estimation using embeddings and neural networks. In *Transforming Learning with Meaningful Technologies*, pages 335–348. Springer International Publishing.

R Flesch. 1948. A new readability yardstick. *The Journal of applied psychology*, 32(3):221–233.

Thomas François. 2015. When readability meets computational linguistics: a new paradigm in readability. *Revue francaise de linguistique appliquee*, Vol. XX(2):79–97.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Ralf Herbrich, Thore Graepel, and Klaus Obermayer. 1999. Support vector learning for ordinal regression. In *Proceedings of the 9th International Conference on Artificial Neural Networks*, page 97–102.

Eliza Hobo, Charlotte Pouw, and Lisa Beinborn. 2023. "geen makkie": Interpretable classification and simplification of dutch text complexity. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 503–517.

Joseph Marvin Imperial. 2021. BERT embeddings for automatic readability assessment. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 611–618, Held Online. INCOMA Ltd.

Joseph Marvin Imperial and Ekaterina Kochmar. 2023. Automatic readability assessment for closely related languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5371–5386, Toronto, Canada. Association for Computational Linguistics.

Joseph Marvin Imperial, Lloyd Lois Antonie Reyes, Michael Antonio Ibanez, Ranz Sapinit, and Mohammed Hussien. 2022. A baseline readability model for Cebuano. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 27–32, Seattle, Washington. Association for Computational Linguistics.

Adam Jatowt and Katsumi Tanaka. 2012. Is wikipedia too difficult? comparative analysis of readability of wikipedia, simple wikipedia and britannica. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 2607–2610, New York, NY, USA. Association for Computing Machinery.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Online. Association for Computational Linguistics.

Isaac Johnson, Florian Lemmerich, Diego Sáez-Trumper, Robert West, Markus Strohmaier, and Leila Zia. 2021. Global gender differences in wikipedia readership. In *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM '21)*.

Peter Kincaid, Robert P. Fishburne, 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*.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Philippe Laban, Jesse Vig, Wojciech Kryscinski, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023. SWiPE: A dataset for Document-Level simplification of Wikipedia pages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10674–10695, Toronto, Canada. Association for Computational Linguistics.

Bruce W Lee and Jason Lee. 2023a. LFTK: Handcrafted features in computational linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19, Toronto, Canada. Association for Computational Linguistics.

Bruce W. Lee and Jason Lee. 2023b. Prompt-based learning for text readability assessment. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1819–1824, Dubrovnik, Croatia. Association for Computational Linguistics.

Justin Lee and Sowmya Vajjala. 2022. A neural pairwise ranking model for readability assessment. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3802–3813, Dublin, Ireland. Association for Computational Linguistics.

Florian Lemmerich, Diego Sáez-Trumper, Robert West, and Leila Zia. 2019. Why the world reads wikipedia: Beyond english speakers. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, pages 618–626, New York, NY, USA. ACM.

Yang Liu, Alan Medlar, and Dorota Glowacka. 2021. Can language models identify wikipedia articles with readability and style issues? In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '21, pages 113–117, New York, NY, USA. Association for Computing Machinery.

Teun Lucassen, Roald Dijkstra, and Jan Maarten Schraagen. 2012. Readability of wikipedia. *First Monday*.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2020a. An analysis of transfer learning methods for multilingual readability assessment. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20 Adjunct, pages 95–100, New York, NY, USA. Association for Computing Machinery.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2020b. Is cross-lingual readability assessment possible? *Journal of the Association for Information Science and Technology*, 71(6):644–656.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

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

Nicholas Mcinnes and Bo J A Haglund. 2011. Readability of online health information: implications for health literacy. *Informatics for health & social care*, 36(4):173–189.

Martina Miliani, Serena Auriemma, Fernando Alva-Manchego, and Alessandro Lenci. 2022. Neural readability pairwise ranking for sentences in Italian administrative language. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 849–866, Online only. Association for Computational Linguistics.

Blagoj Mitrevski, Tiziano Piccardi, and Robert West. 2020. WikiHist.html: English wikipedia's full revision history in HTML format. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:878–884.

Pedro Miguel Moás and Carla Teixeira Lopes. 2023. Automatic quality assessment of wikipedia articles - a systematic literature review. *ACM Comput. Surv.*

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

Naoual Nassiri, Violetta Cavalli-Sforza, and Abdelhak Lakhouaja. 2023. Approaches, methods, and resources for assessing the readability of arabic texts. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(4):1–30.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Readable. Your readability toolkit. https://readable.com/. Accessed: 2024-2-9.

N J Reavley, A J Mackinnon, A J Morgan, M Alvarez-Jimenez, S E Hetrick, E Killackey, B Nelson, R Purcell, M B H Yap, and A F Jorm. 2012. Quality of information sources about mental disorders: a comparison of wikipedia with centrally controlled web and printed sources. *Psychological medicine*, 42(8):1753–1762.

Miriam Redi, Martin Gerlach, Isaac Johnson, Jonathan Morgan, and Leila Zia. 2020. A taxonomy of knowledge gaps for wikimedia projects (second draft). *arXiv:2008.12314*.

Elissa Redmiles, Lisa Maszkiewicz, Emily Hwang, Dhruv Kuchhal, Everest Liu, Miraida Morales, Denis Peskov, Sudha Rao, Rock Stevens, Kristina Gligorić, Sean Kross, Michelle Mazurek, and Hal Daumé, III. 2019. Comparing and developing tools to measure the readability of Domain-Specific texts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4831–4842, Hong Kong, China. Association for Computational Linguistics.

Luz Rello, Horacio Saggion, Ricardo Baeza-Yates, and Eduardo Graells. 2012. Graphical schemes may improve readability but not understandability for people with dyslexia. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 25–32, Montréal, Canada. Association for Computational Linguistics.

Markus Sagen. 2021. Large-context question answering with cross-lingual transfer. Master's thesis, Uppsala University, Department of Information Technology.

Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. 2017. Why we read wikipedia. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1591–1600, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Sanja Stajner. 2021. Automatic text simplification for social good: Progress and challenges. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652, Online. Association for Computational Linguistics.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-Level text simplification: Dataset, criteria and baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7997–8013, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

TextStat. Calculate statistical features from text. https://github.com/textstat/textstat. Accessed: 2024-2-9.

Translated. Readability analyzer. https://translatedlabs.com/text-readability. Accessed: 2024-2-9.

Txikipedia. Basque encyclopedia for children. https://eu.wikipedia.org/wiki/Txikipedia:Azala. Accessed: 2024-2-9.

Ahmet Yavuz Uluslu and Gerold Schneider. 2023. Exploring hybrid linguistic features for turkish text readability. *arXiv:2306.03774*.

David Uthus, Santiago Ontañón, Joshua Ainslie, and Mandy Guo. 2023. mLongT5: A multilingual and efficient Text-To-Text transformer for longer sequences. *arXiv:2305.11129*.

Sowmya Vajjala. 2022. Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377, Marseille, France. European Language Resources Association.

Sowmya Vajjala and Ivana Lucic. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification.

Sowmya Vajjala 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.

Laura Vásquez-Rodríguez, Pedro-Manuel Cuenca-Jiménez, Sergio Morales-Esquivel, and Fernando Alva-Manchego. 2022. A benchmark for neural readability assessment of texts in Spanish. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 188–198, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Morten Warncke-Wang, Vivek Ranjan, Loren Terveen, and Brent Hecht. 2015. Misalignment between supply and demand of quality content in peer production communities. In *Ninth International AAAI Conference on Web and Social Media*.

Zarah Weiss, Xiaobin Chen, and Detmar Meurers. 2021. Using broad linguistic complexity modeling for Cross-Lingual readability assessment. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 38–54, Online. LiU Electronic Press.

Wikikids. Dutch encyclopedia for children. https://wikikids.nl/. Accessed: 2024-2-9.

Wikipedia. Help:wikitext. https://en.wikipedia.org/wiki/Help:Wikitext. Accessed: 2024-2-9.

Wikistats. Statistics for wikimedia projects. https://stats.wikimedia.org/. Accessed: 2024-1-9.

Rodrigo Wilkens, David Alfter, Xiaoou Wang, Alice Pintard, Anaïs Tack, Kevin P Yancey, and Thomas François. 2022. FABRA: French Aggregator-Based readability assessment toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1217–1233, Marseille, France. European Language Resources Association.

WMF-a. List of wikis for children. https://meta.wikimedia.org/wiki/List_of_wikis_for_children. Accessed: 2024-2-9.

WMF-b. Wikipedia:how to write simple english pages. https://simple.wikipedia.org/wiki/Wikipedia:How_to_write_Simple_English_pages. Accessed: 2024-5-28.

WMF-c. Wikimedia rest api. https://www.mediawiki.org/wiki/Wikimedia_REST_API. Accessed: 2024-2-9.

WMF-d. Mediawiki action api. https://www.mediawiki.org/wiki/API:Main_page. Accessed: 2024-2-9.

WMF-e. Community wishlist survey 2022/bots and gadgets/readability scores gadget. https://meta.wikimedia.org/wiki/Community_Wishlist_Survey_2022/Bots_and_gadgets/Readability_scores_gadget. Accessed: 2024-2-9.

WMF-f. Community insights survey. https://meta.wikimedia.org/wiki/Community_Insights. Accessed: 2024-2-9.

WMF-g. Characterizing wikipedia reader behaviour: Demographics and wikipedia use cases. https://meta.wikimedia.org/wiki/Research:Characterizing_Wikipedia_Reader_Behaviour/Demographics_and_Wikipedia_use_cases. Accessed: 2024-2-9.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Menglin Xia, 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. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Taha Yasseri, András Kornai, and János Kertész. 2012. A practical approach to language complexity: A wikipedia case study. *PloS one*, 7(11):e48386.

Taha Yasseri and Filippo Menczer. 2023. Can crowdsourcing rescue the social marketplace of ideas? *Communications of the ACM*, 66(9):42–45.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China. Coling 2010 Organizing Committee.

A Additional data characteristics

The new dataset presented consists of pairs of hard and simple text of one article for multiple languages. However, we also mention that there are articles that are present in multiple languages. For example, we have more than 3K article pairs that are presented in two languages (Figure 5).

B Data preprocessing

We start by matching articles from Wikipedia with the corresponding article in the simplified/children encyclopedia. For simplewiki, we match all articles from English Wikipedia and Simple English Wikipedia via their Wikidata item ids. We remove all disambiguation and list pages as they often display only itemized lists without continuous text. For Txikipedia, we match page titles of all articles in namespace 0 (Main namespace) and 104 (Txikipedia), respectively; for example, the article "Klima" will have "Txikipedia:Klima" as the title for the children's version. For Vikidia, Klexikon, and Wikikids, we match the page title in Wikipedia with the corresponding article in the children encyclopedia. Due to different naming conventions and the fact that the two sources are not completely aligned, articles might have different titles (such as "Baby" and "Infant" in English Wikipedia and Vikidia, respectively). To address this, we also consider all redirects of an article as alternative titles and match pairs if we find one and only one match between all titles of an article.

We get the HTML-version of each article from the HTML dumps (Enterprise) or, alternatively for sources in which they are not available, the Wikimedia APIs (WMF-c,WMF-d).

We then parse the HTML of the article to extract the plain text. We first split the article into sections, only keeping the first section of each article (lead section) to limit extreme differences in length between the two versions. We then only consider text within <p>-tags to avoid text from, e.g., infoboxes, image captions, etc. Within each <p>-tag, we extract the plain text from each element, removing any formatting (e.g. from links) and removing text that is from references or in sub/super-script. This leads to much cleaner plain text than if one would parse the wikitext, mainly due to the wide usage of templates (Mitrevski et al., 2020). While there exist packages for expanding the content of templates in wikitext, such as WikiExtractor (Attardi, 2015), they require the full dump files, which are not publicly available for the children encyclopedias that are not WMF-hosted.

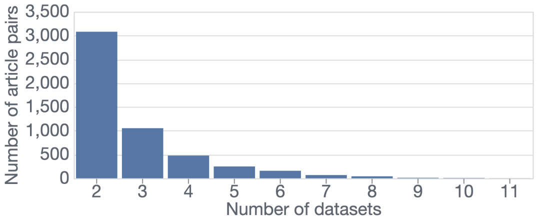

Figure 5: Number of articles that occur in two or more different datasets (single occurrence is skipped).

We only keep pairs of articles in which each version consists of three or more sentences to limit the number of low-quality or stub articles.

Note that, for each pair of articles, we keep the Wikidata item id associated with the corresponding Wikipedia article. This allows us to align pairs of articles across different datasets and, thus, also languages (see Figure 5 in the Appendix). For example, the Wikidata item id Q433 corresponds to the pair of articles with titles "Phyiscs" in Vikidia (English) and "Physik" in Klexikon (German).

C Additional modeling details

C.1 Model hyperparameters

For TRank, we utilize pretrained *xlm-roberta-longformer-base* ($\sim$281M parameters) model by Sagen (2021) as a backbone. This model is designed to process long text, accommodating up to 4096 tokens. However, we limit it to 1500 to fit into our available resource constraints, which allows us to tokenize 99.95% of our dataset without truncation. We concatenate all sentences to construct an article text and use it as model input. This input is then passed through a sequence of MLM and the Dense layer to generate the readability score.

We train a model for three epochs with an initial learning rate of 10^{-5} and weight decay equal to 10^{-7}. Also, we use the CosineAnnealingLR scheduler with hyperparameters $min\_lr = 10^{-7}$ and $T\_max = 256$. Due to memory constraints, the batch size during training is equal to one. Also, we fix the hyperparameter $m = 0.5$, which refers to the margin in the loss function. Also, we use the 1% sample from training data as the validation set. We track the loss and select the checkpoint created after the epoch when the model shows the best performance on validation data. As a result, we get a single model with the lowest loss. It takes $\sim$80 GPU hours for training and inference of the

TRank model on our dataset, and doing additional experiments presented in the paper.

For SRank, we use *bert-base-multilingual-cased* ($\sim$178M parameters) by (Devlin et al., 2019) as a backbone. We use the same hyperparameters for model training as for the TRank model, except for batch size, which is instead set to 16 for SRank. During the inference stage, each sentence from the article is individually passed through the model, and the scores are aggregated using mean pooling. We use the same training strategy as for the TRank model, but we prepare a custom sentence-based training dataset. To construct this dataset, we use Levenshtein Distance to select similar sentence pairs from the aligned articles available in different readability levels. It takes $\sim$30 GPU hours for training and inference of the SRank model on our dataset, and doing additional experiments presented in the paper.

The choice of values for the hyperparameters was motivated by previous approaches using similar models that have been shown to perform well in practice(Chanda, 2021).

C.2 Computational resources

The model choice is motivated by the available computational resources (1x AMD Radeon Pro WX 9100 16GB GPU). In total, $\sim$160 GPU hours are needed to reproduce the experiments presented in this paper.

D Confidence intervals

We estimate confidence intervals of the ranking accuracy metric via bootstrapping (Efron and Tibshirani, 1994). Specifically, we resample each test set 10K times by drawing randomly with replacement N samples from the test set, where N is the size of the test set. For example, if we have 1K observations in the testing data, we randomly choose 10K samples of size 1K (with replacement) from the given data. We then calculate the standard deviation of the ranking accuracy over the 10K bootstrap-samples. We report two standard deviations as the confidence interval, denoted as CI in Tables 2 and 3.

We calculate the CI for the $NPRM$ model taken from Lee and Vajjala (2022) without reproducing model inference. We use the scores for each sample of benchmark datasets published by the authors, which allows us to bootstrap their results without reproducing model inference.