UNIVERSITY OF
BATH

Link to publication

University of Bath

Alternative formats
If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

BERT Embeddings for Automatic Readability Assessment

Joseph Marvin Imperial
National University
Manila, Philippines
jrimperial@national-u.edu.ph

Abstract

Automatic readability assessment (ARA) is the task of evaluating the level of ease or difficulty of text documents for a target audience. For researchers, one of the many open problems in the field is to make such models trained for the task show efficacy even for low-resource languages. In this study, we propose an alternative way of utilizing the information-rich embeddings of BERT models with handcrafted linguistic features through a combined method for readability assessment. Results show that the proposed method outperforms classical approaches in readability assessment using English and Filipino datasets—obtaining as high as 12.4% increase in F1 performance. We also show that the general information encoded in BERT embeddings can be used as a substitute feature set for low-resource languages like Filipino with limited semantic and syntactic NLP tools to explicitly extract feature values for the task.

1 Introduction

Automatic readability assessment is the task of evaluating the level of ease or difficulty of text documents such as web articles, story and picture books, test materials, and medical prescriptions. Often readability levels can be expressed in many forms: discrete values with grade and age levels such as in the Common European Framework of Reference for Languages (CEFR)[1], or with continuous values from a given range such as in the famous Lexile Reading Framework[2]. In machine learning setting, this task is most often viewed as a classification task where an annotated set of corpora is trained with its corresponding gold-standard labels evaluated by an expert as mostly done in previous works (Vajjala, 2021; Chatzipanagiotidis et al., 2021; Weiß

and Meurers, 2018; Xia et al., 2016; Reynolds, 2016; Hancke et al., 2012; Vajjala and Meurers, 2012). Recent works have tried testing unexplored resources by utilizing large pre-trained language models such as Bidirectional Encoder Representations or **BERT** (Devlin et al., 2019) which is based on the attention-driven Transformer architecture by Vaswani et al. (2017) by (a) directly processing the data to the network (Martinc et al., 2021; Tseng et al., 2019) or by (b) using the discrete output of the network via transfer learning (Deutsch et al., 2020) as an additional feature. For these methods, however, evidence of efficacy are only seen in high-resources readability datasets in English. Thus, we propose an alternative way of incorporating the knowledge of large language models such as BERT by combining its information-rich sentence embeddings as a separate feature set for traditional machine learning algorithms with handcrafted linguistic features. We argue that this method is not only low-resource friendly but also preserves the semantic and syntactic information encoded by the attention heads of BERT since the embeddings itself will be used. We show that such information can act as a substitute for languages with limited tools for explicitly extracting semantic and syntactic features where results describe non-significance in difference of performances between models using semantic and syntactic features versus models using BERT embeddings.

2 Previous Work

The first generation of readability formulas and indices date as early as 1920-1940s with the works of Thorndike (1921), Dale and Chall (1948), and Flesch (1948) primarily using surface-based variables such as raw frequencies and average values of sentences and words per document. The process for using such indices requires manual computa-

---

[1]https://www.cambridgeenglish.org/exams-and-tests/cefr/

[2]https://lexile.com/

tion and plugging of values to formulas which can be tedious as the length of a document increases. Likewise, experts argue that considering narrow, surface-based features do not entirely capture the linguistic complexity of a given text (Macahilig, 2015; Collins-Thompson and Callan, 2004; Si and Callan, 2001). Thus, incorporation of deeper, linguistic variables such as a language's semantics, syntax, morphology, and discourse properties are imperative and worth exploring for the task. To answer this call, the use of handcrafted linguistic features remained the most popular type of input for training readability assessment models through the years. Handcrafted linguistic features are often represented as real-valued numbers serving as potential predictors of the difficulty of reading materials. These features span on a wide range of linguistically motivated factors that base on syntax, semantics, morphology, cohesion, and cognition to name a few. These features also serve as the input in the form of vectors for conventional readability assessment setups using traditional classification-based algorithms. To note, not all linguistic features can be applied or extracted for all languages as some have limited NLP tools suitable for use especially for low-resource languages. Notable works in various languages such as Greek (Chatzipanagiotidis et al., 2021), German (Weiss and Meurers, 2019; Weiß and Meurers, 2018; Hancke et al., 2012), Bangla (Sinha et al., 2012), and Filipino (Imperial and Ong, 2021a, 2020) have used this approach in combination with traditional machine learning algorithms such as Logistic Regression and Support Vector Machines. Likewise, another reason why studies have resorted to the classical approach of model building is that deep neural models are not practical for the task without a large amount of training data.

The advent of large and complex pre-trained language models such as BERT and its variations spawned a handful of studies on how these models fare with the readability assessment tasks. The work of Martinc et al. (2021) on the supervised experiment setup explored directly using English benchmark corpus such as Weebit and OneStopEnglish as input for BERT via transfer learning while Deutsch et al. (2020) explored using the final discrete output of BERT as a feature for the same datasets. Results from both studies show effectiveness of BERT for English data as direct input while no significant improvement is seen when the

discrete output itself is used as a feature. While these results are remarkable, BERT's effectiveness remain a gray area for low-resource languages.

3 Task Definition

We define our task at hand as a supervised learning setup. Given a text document d where a feature vector $x = [x_1, x_2 \ldots, x_n]$ is extracted, a model M is trained using said collection of features X along with the gold label Y or expert-identified readability level. The label is relative in form (discrete or continuous) based on how readability levels are categorized for each corpus.

Data	Doc Count	Sent Count	Vocab
OSE	567	4,890	17,818
CCE	168	20,945	78,965
Adarna House	265	10,018	16,058

Table 1: Data distribution for English and Filipino corpus.

4 Corpus

We describe each corpus used in the study below as well as the statistics and breakdown in Table 1

OneStopEnglish. The OSE corpus is a collection of 567 texts in three different reading levels (beginner, intermediate, and advanced) for adult ESL learners from the MacMillan Education website[3]. This corpus was first used in the work of Vajjala and Lučić (2018) and has become one of the most-used benchmark datasets for readability assessment and text simplification in English.

Common Core Exemplars. The CCE dataset contains 168 prose texts from the Appendix B of the Common Core State Standards Initiative (CCSS)[4] for English Language studies and first used by Flor et al. (2013) for readability assessment. The initiative was a project of the National Governors Association and the Council of Chief State School Officers in USA[5]. The dataset is divided into three age-range categories: 2-5, 6-7, and 9-12.

Adarna House. The Adarna House corpus is a collection of 265 story books for grades 1-3 from Adarna House Inc.[6], the largest children's litera-

---

[3] https://www.onestopenglish.com/
[4] http://www.corestandards.org/assets/Appendix_B.pdf
[5] http://www.ccsso.org
[6] https://adarna.com.ph/

ture publisher in the Philippines. This corpus has been used by Imperial et al. (2019); Imperial and Ong (2020, 2021a) for readability assessment in Filipino[7]

5 BERT Embeddings + Handcrafted Linguistic Features

BERT's efficacy on a wide range of NLP tasks stems from its implicit capability to encode linguistic knowledge such as hierarchical parse trees (Hewitt and Manning, 2019), parts of speech and syntactic chunks (Liu et al., 2019; Tenney et al., 2019), semantic roles (Ettinger, 2019) as well as entity types and relations (Tenney et al., 2019) to name a few. In view with this, we find such amount of knowledge an extremely valuable resource which can potentially improve performances of readability assessment models especially for low-resource languages if used correctly. Thus, to maximize the potential of BERT for low-resource readability assessment, we propose a combined training of its raw embeddings with handcrafted linguistic feature sets through a *concatenation process* and feeding them to traditional machine learning algorithms. The embeddings of BERT generated by the multi-head attention layers are *information-rich*, specifically on semantic and syntactic knowledge (Rogers et al., 2020), due to the nature of its training. We describe our proposed architecture in Figure 1 with a sample Filipino sentence for context.

6 Experiment Setup

For the OSE and CCE corpus in English, we extracted over 155 linguistic features covering lexical diversity and density features, syntactic features based on parse trees, morphosyntactic properties of lemmas, and word-level psycholinguistic features. For the Adarna House corpus in Filipino, we extracted over 54 linguistic features covering traditional surface-based features, lexical features based on POS tags, language model features, morphology based on verb inflection, and orthographic features based on syllable pattern. The size of the BERT embeddings for all datasets remain equal with a fixed dimension of $H = 768$ since the base version of BERT for English (Devlin et al., 2019) and Filipino (Cruz et al., 2020c; Cruz and Cheng, 2020, 2019) were used. The embeddings and extracted linguistic feature sets were concatenated, for a total of 923

dimensions for combined features for both English datasets and 823 for the Filipino dataset. Recipes for feature extraction were obtained from the studies of Vajjala and Meurers (2016, 2014) for English and Imperial and Ong (2020, 2021a,b) for Filipino. We used the `sentence-transformers` library by Reimers and Gurevych (2019) with mean pooling option to extract BERT embedding representations for the readability corpora[8].

For the traditional machine learning algorithms, we used three of the commonly utilized in previous works: Logistic Regression, Support Vector Machines, and Random Forest. Models for each dataset were trained on a 5-fold cross validation procedure. We used weighted F1 as the overall metric for performance evaluation.

7 Results

7.1 Ablation

We compared performances of models on three different setups, (a) linguistic features only, (b) BERT sentence embeddings only, and (c) combined training of the two feature embeddings to gauge the efficacy of the proposed framework.

As described in Table 2, generally speaking, models trained using the proposed combined training of handcrafted linguistic feature sets with contexual BERT embeddings outperform both performances of only using each exclusively on English and Filipino datasets. On average, we note an increase of performance of 2.63% for OSE, 6.23% for CCE, and 12.4% in weighted F1 score for Adarna House across all algorithms. From this, we infer that extracting and incorporating the information-rich embeddings of any readability dataset using BERT to commonly-used linguistic feature sets can substantially improve model performance.

Interestingly, there are a few notable cases reported in Table 2 where BERT embeddings alone outperformed the traditional method of using handcrafted linguistic feature sets as primary input. These cases are evident in the all models utilizing the Adarna House dataset in Filipino with an average increase of 9.5% weighted F1 scores. From this, we infer that the general semantic and syntactic knowledge implicitly encoded in BERT embeddings as detailed in probing tasks from previous works (Rogers et al., 2020; Hewitt and Manning,
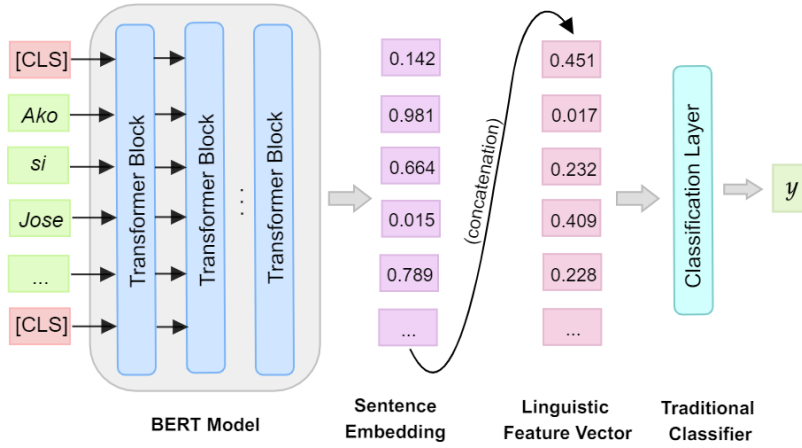

Figure 1: The proposed combined training approach using sentence embeddings from BERT model and extracted handcrafted linguistic feature sets.

Method	OSE	CCE	Adarna
Linguistic Features	0.676	0.774	0.389
BERT Embeddings	0.620	0.747	0.505
Combined Features (Ling + BERT)	**0.732**	**0.778**	**0.554**

(a) Logistic Regression

Method	OSE	CCE	Adarna
Linguistic Features	0.691	0.732	0.414
BERT Embeddings	0.611	0.826	0.487
Combined Features (Ling + BERT)	**0.704**	**0.893**	**0.571**

(b) Support Vector Machines

Method	OSE	CCE	Adarna
Linguistic Features	0.683	0.842	0.423
BERT Embeddings	0.439	0.770	**0.504**
Combined Features (Ling + BERT)	**0.690**	**0.861**	0.467

(c) Random Forest

Table 2: F1 performance via training with (a) Logistic Regression, (b) Support Vector Machines, and (c) Random Forest using handcrafted linguistic features, BERT sentence embeddings, and combined training of both.

2019; Liu et al., 2019; Tenney et al., 2019) may be significantly more informative than the traditional handcrafted linguistic features for discriminating reading difficulty. Consequently, this poses as a probable and alternative solution for low-resource languages with little to no NLP tools such as a good part-of-speech tagger, stemmer, syntactic parse tree extractor, and morphological analyzer to name a few for manually extracting linguistic information from documents. Since BERT models are trained in an self-supervised manner, the overhead of developing these tools from scratch can be disregarded, at least for readability assessment. We discuss further experiments on this inference in the next section.

7.2 Substituting Semantic and Syntactic Features for BERT Embeddings

To empirically test if BERT embeddings can act as substitute for semantic and syntactic linguistic features for readability assessment, we removed features from the three datasets that assume semantic and syntactic knowledge. For OSE and CCE, we removed 56 features covering part-of-speech densities, lexical richness, type-token densities, and general parse-tree based features. For Adarna, we removed 22 features covering part-of-speech densities, type-token densities, and verb inflection densities. There are no parse-tree based features for Adarna House as there are currently no NLP tools for extracting such feature set for Filipino. The rest of the linguistic features from the datasets denoting other aspects of reading difficulty measurement such as frequency-based features and syllable patterns remain unchanged. Models were retrained using the three selected machine learning algorithms for comparison.

Results of substitution experiments can be found in Table 3. Generally speaking, it is evident that models trained using the combined method still outperforms models using the reduced feature set on the account of CCE and Adarna data. However, we note the 1.2% increase in F1 score on the OSE data. Stemming from this observation, we also note small differences in performances of using the combined features against decreased features. In the CCE corpus, the highest performing model using decreased features obtained 86.9% F1 score which is less 2.4% than using the model with combined features. For the Adarna data, the difference is 6.4%.

To identify if such difference is significant, we used a two-tailed test of difference via Mann-Whitney U using the performance scores of models with combined features and models with decreased features for all datasets. We arrived at a p-value of 0.522 ($p >0.5$), meaning that the difference of the scores between two groups is not significant[9]. Thus, we conclude that BERT embeddings can be fully used as a substitute for semantic and syntactic features if such information cannot be explicitly extracted from readability data due to the lack of NLP tools and low-resourceness of other languages. To add, since BERT models are trained in an self-supervised manner and there are over 3,000 pretrained models from online repositories[10], these resources and the proposed combined training method as a viable option.

7.3 Feature Decomposition for Performance Boost

In extending the effort to improve the performance of BERT-enriched readability assessment models and reduce feature size or dimensionality, we resorted to the use of feature decomposition to the large feature vector sizes (BERT + linguistic features) via Principal Components Analysis (PCA). PCA works by projecting the overall feature set (often large) to a lower dimensional property while preserving quality and information of features (Hotelling, 1933). We experimented on differing values of variance percentages: 25, 50, 75, 95, and 100 (full, no feature removed). Results of feature decomposition via PCA for each machine learning model are described in Figure 2.

For SVM and Random Forest, all datasets have the highest performances if all features are retained (100 variance percentage). While for Logistic Regression, 75 variance percentage obtained the highest performance with 82.7 F1 score for OSE, 95 variance percentage obtained the highest performance with 83.8% F1 score for CCE, and 100% or full features for Adarna. Thus, we infer that there is no need to perform feature decomposition to find the principal components as the highest-performing models for OSE, CCE, and Adarna use 100% of the combined feature set (BERT + linguistic features).

---

[9]The distribution of the two groups are of equal variances with p-value of 0.619.
[10]https://huggingface.co/models

Model w/ Removed Features	OSE	CCE	Adarna
Logistic Regression	**0.744**	0.865	0.492
Support Vector Machines	0.615	0.869	0.507
Random Forest	0.669	0.791	0.431
Full Model (Ling + BERT)	0.732	**0.893**	**0.571**

Table 3: Performances of models via F1 score after retraining with semantic and syntactic handcrafed linguistic features removed to test if information-rich BERT embeddings can act as substitution for such features. Best performing model utilizing combined features from Table 2 appended for comparison.
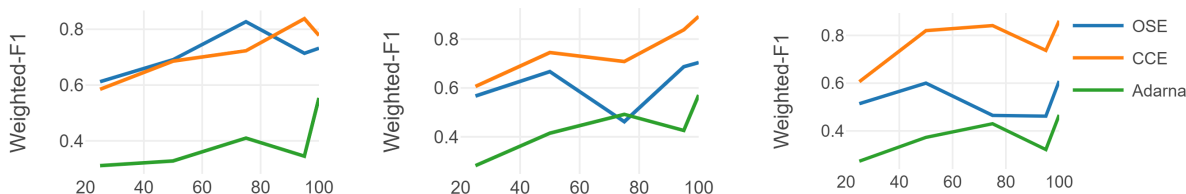

Figure 2: Decomposing large feature sets on a 25%, 50%, 75%, 95%, and 100% (full) variance percentages using PCA for Logistic Regression, Support Vector Machines, and Random Forest (left to right).

8 Conclusion

In this study, we proposed an alternative way of combining information-rich BERT embeddings with handcrafted linguistic features for the readability assessment task. Results from our experiments showed that the method outperforms classical, vanilla approaches in readability assessment using English (OSE and CCE) and Filipino (Adarna) datasets in various machine learning algorithms such as Logistic Regression, Support Vector Machines, and Random Forest. We also demonstrated that the knowledge implicitly encoded in BERT embeddings (semantic and syntactic information) can be used as a full substitute feature set for low-resource languages like Filipino with limited NLP tools to explicitly extract feature values for the task. We are looking forward to the application of our proposed method to other languages struggling with the extraction deep linguistic features to trained readability assessment models. Future directions of the study include deeper exploration of BERT such as isolating extracted embeddings for each of the twelve attention layers.

9 Acknowledgments

10 Ethical Considerations

We report that there are no major ethical concerns in the study as it involves no human subjects nor discriminate any identifiable group of people. As for the dataset, for Adarna House, permission was obtained from the publishing house while for the OSE and CCE datasets, it remains open-sourced. As for energy consumption, the study only uses pre-trained BERT models and the authors did not actually perform the pre-training phase itself.

References

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 and James P. Callan. 2004. A language modeling approach to predicting reading difficulty. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 193–200, Boston, Massachusetts, USA. Association for Computational Linguistics.

Hilaria Cruz, Antonios Anastasopoulos, and Gregory Stump. 2020a. A resource for studying chatino verbal morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2827–2831, Marseille, France. European Language Resources Association.

Jan Christian Blaise Cruz and Charibeth Cheng. 2019. Evaluating language model finetuning techniques for low-resource languages. *arXiv preprint arXiv:1907.00409.*

Jan Christian Blaise Cruz and Charibeth Cheng. 2020. Establishing baselines for text classification in low-resource languages. *arXiv preprint arXiv:2005.02068.*

Jan Christian Blaise Cruz, Julianne Agatha Tan, and Charibeth Cheng. 2020b. Localization of fake news detection via multitask transfer learning. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2596–2604, Marseille, France. European Language Resources Association.

Jan Christian Blaise Cruz, Julianne Agatha Tan, and Charibeth Cheng. 2020c. Localization of Fake News Detection via Multitask Transfer Learning. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2589–2597.

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

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.

Allyson Ettinger. 2019. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *CoRR*, abs/1907.13528.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Michael Flor, 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, Georgia. Association for Computational Linguistics.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of COLING 2012*, pages 1063–1080, Mumbai, India. The COLING 2012 Organizing Committee.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

H. Hotelling. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:498–520.

Joseph Marvin Imperial and Ethel Ong. 2020. Exploring hybrid linguistic feature sets to measure filipino text readability. In *2020 International Conference on Asian Language Processing (IALP)*, pages 175–180. IEEE.

Joseph Marvin Imperial and Ethel Ong. 2021a. Application of lexical features towards improvement of filipino readability identification of children's literature. *arXiv preprint arXiv:2101.10537.*

Joseph Marvin Imperial and Ethel Ong. 2021b. A simple post-processing technique for improving readability assessment of texts using word mover's distance. *CoRR*, abs/2103.07277.

Joseph Marvin Imperial, Rachel Edita Roxas, Erica Mae Campos, Jemelee Oandasan, Reyniel Caraballo, Ferry Winsley Sabdani, and Ani Rosa Almaroi. 2019. Developing a machine learning-based grade level classifier for filipino children's literature. In *2019 International Conference on Asian Language Processing (IALP)*, pages 413–418. IEEE.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. 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 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Heidi B. Macahilig. 2015. A content-based readability formula for Filipino texts. *The Normal Lights*, 8(1).

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Robert Reynolds. 2016. Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories. In *Proceedings of the*

11th Workshop on Innovative Use of NLP for Building Educational Applications, pages 289–300, San Diego, CA. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, page 574–576, New York, NY, USA. Association for Computing Machinery.

Manjira Sinha, Sakshi Sharma, Tirthankar Dasgupta, and Anupam Basu. 2012. New readability measures for Bangla and Hindi texts. In *Proceedings of COLING 2012: Posters*, pages 1141–1150, Mumbai, India. The COLING 2012 Organizing Committee.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *CoRR*, abs/1905.06316.

Edward L Thorndike. 1921. The teacher's word book.

Hou-Chiang Tseng, Hsueh-Chih Chen, Kuo-En Chang, Yao-Ting Sung, and Berlin Chen. 2019. An innovative bert-based readability model. In *International Conference on Innovative Technologies and Learning*, pages 301–308. Springer.

Sowmya Vajjala. 2021. Trends, limitations and open challenges in automatic readability assessment research. *CoRR*, abs/2105.00973.

Sowmya Vajjala 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, Louisiana. Association for Computational Linguistics.

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, Montréal, Canada. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2016. Readability-based sentence ranking for evaluating text simplification. *CoRR*, abs/1603.06009.

Sowmya Vajjala and Walt Detmar Meurers. 2014. Readability assessment for text simplification: From analysing documents to identifying sentential simplifications. *ITL – International Journal of Applied Linguistics*, 165:194–222.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Zarah Weiß and Detmar Meurers. 2018. Modeling the readability of German targeting adults and children: An empirically broad analysis and its cross-corpus validation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 303–317, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Zarah Weiss and Detmar Meurers. 2019. Analyzing linguistic complexity and accuracy in academic language development of German across elementary and secondary school. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 380–393, Florence, Italy. 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.