Developing a readability model for children's text using lexical and syntactic features

MOSTOFA NAJMUS SAKIB, Boise State University, USA

ACM Reference Format:
Mostofa Najmus Sakib. 2021. Developing a readability model for children's text using lexical and syntactic features. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 10 pages.

1 ABSTRACT

The importance of readability in any literature is an exhaustive conversation but a necessary one. Readability simply refers to the quality of reading ease of a piece of text. Readability assessment helps quantify the level of difficulty that a reader experiences in comprehending a particular text. Assessing text readability is a time-honored problem that has even more relevance in today's information-rich world. A text's readability is also a function of the readers themselves, their educational and social background, interests and expertise, and motivation to learn. The ability to quantify the readability of a text is achieved through the use of readability measures that take a text as input and estimate a numerical score or another form of prediction that indicates the level or degree of readability for a given population. This study examines the potential of applying advanced machine learning techniques to the educational problem of assessing text difficulty, more specifically focusing on the texts specialized for children. The combination of hierarchical machine learning and natural language processing (NLP) is leveraged to predict the difficulty of practice texts used in reading comprehension. Children being the nation's future, should have access to books and resources appropriate to their age. Our goal is to mold a robust readability model to offer such benefits to them.

2 INTRODUCTION

Readability, or the measurement of how difficult a passage of text is to read, is a section of Natural Language Processing (NLP) that has been researched over many decades, giving rise to many formulas that provide various scores or levels to quantify the reading difficulty of text [12]. While readability has applications in many areas (e.g., online medical infographics [33], privacy statements [1], and financial reports [23]), one of the least explored ones is education [37]. Over the years, teachers and language researchers have emphasized the importance of readability in the education sector. Text remains a crucial learning tool in most classrooms today [16]. In the classroom, students are often tasked to read texts and textbooks in order to learn new information. As such, many educators and text publishers attempt to level texts such that text difficulty is appropriate for the students. Readability formulas have been used for well over a century as a means to evaluate text difficulty. Indeed, teachers have long relied on age-old readability metrics to select classroom materials [3, 15]. Finding the texts that match students reading skills and course content is extremely

Manuscript submitted to ACM

challenging. However, Flesch–Kincaid and other common formulas, e.g., Dale-Chall, Spache, or Coleman-Liau, detect the readability level using shallow text features such as sentence length, number of words, etc., in a sentence.

There are ways of predicting how hard a piece of writing will be to understand (its textual difficulty). Research has shown that two main factors affect the ease with which texts are read [20]. The first one is **how difficult the words are**, this is the lexical difficulty. Rare words are less well-known than common words. Rare, difficult words are often longer than common, easy words. Second, **how difficult the sentences are**: this is the syntactic difficulty. Long, complicated sentences cause more difficulty than short, simple sentences. The lexical analysis in NLP deals with the study at the level of words with respect to their lexical meaning. Syntactic complexity metrics have proven to be indispensable research tools in various language-related research areas, including child language acquisition, language and aging, and second language acquisition, among others. For example, Ramer [31] examined the developmental trends of child syntactic acquisition during the early syntactic period using a simplicity-complexity dimension, which specified a sequence of acquisition observed in the data. With the advancement of natural language processing (NLP) technology in the past decade, there have been a few recent endeavors to develop computational systems for automatic syntactic complexity measurement. For example, Voss [42] implemented a heuristics-based system that incorporates very shallow syntactic analysis for rating the complexity of sentences using the D-Level scale. Graesser et al. [17] developed a software package, Coh-Metrix, which calculates the coherence of texts on a wide range of measure.

Readability is influenced by many factors, such as a degree of lexical and syntactic sophistication, discourse cohesion, and background knowledge [10]. Recently, machine learning strategies to assess readability have been introduced that go beyond shallow text features [9]. The current research works stress the importance of combining both the lexical and syntactic features and developing a model that would be ideally suited for classyfying children's text. Existing literature has lackings in labeling the texts suitable for children. Taking advantage of deep syntactic and insightful lexical features, we have worked on a readability assessment project that will help us better understand the ways to assess the degree of difficulty of a text written for children. However, we have narrowed the scope of our work to young readers, namely children aged between 6-11 years. With the advancement in deep learning, neural network architectures like recurrent neural networks (RNN and LSTM) and convolutional neural networks (CNN) have shown a decent improvement in solving text classification problems. BERT (Bidirectional Encoder Representations from Transformers) [11] is a big neural network architecture. It is one of the most popular transformer-based models and in this research, we fine-tuned BERT for text classification.

In order to perform our experiments, we have taken advantage of labeled texts for children developed by Reading A-Z to train classifiers to measure the readability of any given text. Reading A-Z is a website that offers affordable, online teaching resources for differentiated offline reading instruction. We relied heavily on their assigned levels to measure the performance of our task. We tried a handful amount of experiments for our analysis. Lexical and syntactic features were applied together and separately on predicting the A-Z levels. We also measured the prediction performance of the features on a group of levels and feed the A-Z texts to BERT for capturing the levels. With this preliminary exploration, we have fixed a list of questions that we will try to answer throughout this research.

- **RQ1** – Which classifier was more persistent? Was it consistent/effective for all/certain levels?
- **RQ2** – Which features had the most impact?
- **RQ3** – Is there a general pattern between the misclassified texts?
- **RQ4** – Can we apply the selected feature sets/model to score children's text?

Our findings revealed that lexical features had an edge over the syntactical features determining readability. This seemed to be consistent even when tested for a group of levels or all levels. There were more misclassified texts when syntactic features were used for evaluation separately. Lexical features hit the top of the list when feature importance was checked. Transfer learning experiment or BERT fine-tuning didn't have a promising result on the A-Z text.

3 RELATED WORK

The quantitative analysis of English text readability started with L.A. Sherman in 1880 [34]. To date, English has over 200 readability metrics. Now there are formulas for Spanish, French, German, Dutch, Swedish, Russian, Hebrew, Chinese, Vietnamese, and Korean [30]. The existing quantitative approaches towards predicting the readability of a text can be broadly classified into three categories [2]: **traditional methods** incorporate the easy-to-compute features of a text like a sentence length, paragraph length, etc. The examples are Flesch Reading Ease Score [13], FOG index [18], Fry graph [14], SMOG [28] etc. The chronologically newer formulas like new Dale-Chall index [4], lexile framework [36], ATOS-TASA [21], Read-X [29] considers the readers' background and text semantics; **cognitively motivated methods** use high level text parameters like cohesion and cognitive aspects of the reader. Proposition and inference model, prototype theory, latent semantic analysis, semantic networks are examples of this category. This type of approach introduced text levelling or text revising methods. Two distinguished instances of this class are Coh-metrix [17], and the DeLite software [41]; the third class of approaches incorporate the power of **machine learning methods** and probabilistic analysis. They are useful in determining online readability based on user queries [22] and predicting the readability of web texts [6, 7, 35]. Sophisticated machine learning methods like support vector machines have been used to identify grammatical patterns within a text and classification based on it [19]. A successful classification approach to readability was proposed by Vajjala and Meurers [40] Their multi-layer perceptron classifier is trained on the WeeBit corpus which contains articles from WeeklyReader and BBCBitesize The texts were classified into five classes according to the age group they are targeting. They report 93.3% accuracy on the test set. For our analysis, we also envisioned a similar strategy but extended the scope by exercising multiple classifiers. Beyond the empirical analysis, we also explored the applicability of BERT, to detect the readability of English text. Tseng et al. [38] used BERT to assess the readability of Chinese texts. We explored the potential of following the research process introduced therein to adapt BERT to English texts for children, using the HuggingFace library [43].

4 DESCRIPTION OF THE DATA SET

We received around 2400 annotated doc files from A-Z. All the documents had their names and corresponding levels written at the start. Figure 1 represents one of the sample doc files where the first line indicates the name of the doc file and the second line has its corresponding level. As seen from figure 1, each document contains some bolded lines which represent the link to the audio files for each paragraph. The only necessary part of the documents was the unbolded plain texts. A-Z has categorized its learning system into 29 levels. Human annotators have classified the received documents into those groups based on the difficulty level. A graphic representation of the frequency distribution of the documents is shown in Figure 2. Likewise level, all the documents had corresponding grade levels associated with them. Table 1 summarizes the levels and their equivalent grades.

5 FEATURE DESCRIPTION

We have used two types of features for our experiments e.g. lexical and syntactic features. As part of the feature generation, we have used a lexical complexity analyzer [24, 25] which generates 33 lexical features (see Appendix for

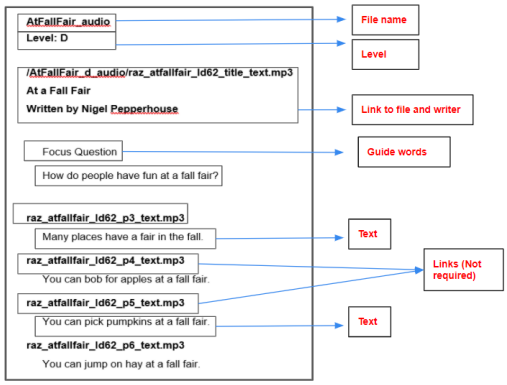

Fig. 1. A-Z: Sample text file.

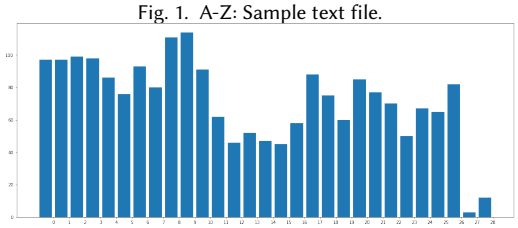

Fig. 2. Frequency distribution of A-Z Level.

Table 1. Level and corresponding grades for A-Z datasets

Level	Grade	Level	Grade	Level	Grade	Level	Grade
aa, A - C	K	D - J	1	K - P	2	Q - T	3
U - W	4	X - Z	5	Z1- Z2	5+		

description) listed in Table 2. We also used a heuristics-based system for automatic measurement of syntactic complexity using the revised Developmental Level (D-Level) scale [8, 32]. The system takes raw text as input and assigns it to an appropriate developmental level on the scale. The raw inputs are annotated through a parts of speech tagger developed by Stanford university and later parsed using Collins parser[5]. The D-level scale is designed with child language acquisition and psycho-linguistic research in mind and is therefore developed and evaluated using both written data from the Penn Treebank [27] and spoken child language acquisition data from the CHILDES database [26]. We have a total of 9 d-level features (Level 0-7 and Mean level).

6 EXPERIMENTS

6.1 Experiment 1

In this experiment, we evaluated how the traditional machine learning classifiers perform in detecting text labels while the above described lexical and syntactical feature sets were used. We used syntactic features in two ways (i) number of sentences on each level (ii) relative frequency on each level. Lexical feature values were exactly same in both case. We also fine-tuned BERT on the A-Z datasets for detecting both levels and grades separately. With classification approach we placed features directly in the classifiers but with BERT, we passed texts and the model generated it's own set of features for detecting both levels and grades separately.

Table 2. Lexical feature sets

Feature name (Acronym)	Full form	Feature name (Acronym)	Full form
LD	Lexical Density	LV	Lexical variation
LS I & II	Lexical sophistication I & II	VV 1 & 2	Verb variation 1 & 2
VS1	Verb Sophistication	SVV 1 & 2	Squared verb variation 1 & 2
CVS1	Corrected verb sophistication	CVV 1 & 2	Corrected verb variation 1 & 2
NDW	Number of different words	NV	Noun variation
TTr	Type token ratio	Adjv & advv	Adjective & Adverb variation
msttr	Mean segmented type token ratio	modv	Modifier variation
cttr	Corrected type token ratio	Word tokens	Word tokens
uber	Uber index	Lextokens	Lexical tokens
Rttr	Root type token ratio	Lextypes	Lexical types
Logttr	Bi-logarithmic type token ratio	Wordtypes	Word types

Task and procedure For this experiment, we fed all the features generated from the lexical and D-level analyzer into the classifiers and measured their overall performance to predict the annotated class levels. A 5 fold stratified cross-validation was done in order to keep all the levels representation in both the test and train set. All the features were normalized before feeding to the classifiers. In order to perform the text analysis with BERT, we fed the model with the previously cleaned texts from the A-Z documents. Due to computation limitations, we couldn't increase the maximum length of character processing beyond 128 or vary the batch size.

Model We have explored a total of 8 classification algorithms e.g., Random forest, Logistic regression, Support vector machine, Extra trees classifier, Gaussian naive bayes, XGBoost, Decision tree, and Multilayer perceptron to map from features to a distribution over class labels. For Support Vector Machine algorithm, we used a linear kernel. Incase of Multilayer perceptron, we used the "lbfgs" solver and L2 regularization. We set the hyperparamter "number of trees in the forest" to 200 and 100 for random forest and decision tree. For text input, we used the pre-trained BERT model as BERT has shown to improve over traditional recurrent models. We trained using a cross-entropy loss function and the adam optimizer, for a period of over 3 epochs.

Metrics For every classifier, Accuracy, Precision, Recall, and F1 score were calculated and an average value of those metrics was considered for evaluation. F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively, the F1 score gives a better picture for evaluation purposes. We have reported two types of baseline e.g., random baseline and most common baseline. Random baseline is the probability of a level being randomly selected out of all the available levels. The most common baseline is for the level with maximum number of samples, divided by the total number of samples. For this scenario, the random baseline was 0.034 and the most common baseline was 0.054. For fine tuning with BERT, We reported the accuracy and loss for the first epoch.

Results Overall "**extra-tree classifier**" had the best performance out of the tested ones in terms of F1 score, so we focused on reporting the results for that algorithm. We had an exact similar result for both type of syntactic feature input. The reason behind such similarity was the normalization of features before fitting in the classifiers. So the number of events or frequency doesn't have any difference in classification. The lower levels (1-10) tend to be more accurately measured by the extra tree classifier (See appendix for confusion matrix). On the other hand, the middle region (Level 11-16) seems to be distributed to different levels making the prediction levels inaccurate for most documents of this level. We had a relatively lower amount of text samples available for this region providing fewer examples for training

purposes. The last region (Level 17-29) had a relatively better performance in comparison to the middle regions although still there was a handful amount of misclassification. With limited computation power, the BERT model generally failed to recognize the different levels. The model fitted the documents only to a few class levels. With the availability of a better computation setup in the future, we plan to progress this analysis beyond the current representation. A similar setup was used to fine-tune BERT for the grade level as well. As expected, tagging the texts at a grade level had better accuracy compared to the previous analysis as we only had fewer grades to predict.

Table 3. Results from Experiment 1 (All features)

Best classifier	Accuracy	Precision	Recall	F1 score
ExtraTree	36.6	32.9	33.7	33.5

Table 4. Performance of BERT in text classification

Level/Grade	Accuracy	Loss
Level	14.35	3.37
Grade	49.76	1.62

6.2 Experiment 2

Since the performance of the classifiers varied along with levels, for this experiment, we examined the classification performance separately for different levels to check if the classifiers were more effective for certain levels.

Task and procedure For the first step, we separated the samples set into three groups. Level 1-10 belonged to group 1, level 11-16 to group 2, and level 17-29 to group 3. All the features were trained and tested on all three groups separately. Five-fold stratified cross-validation and normalization of the features were done for all the groups.

Model This is the same as experiment 1

Metrics This experiment also has similar metrics as experiment 1 for evaluation. However, due to the variation of sample size and levels, we had a different random and most common baseline which are listed in table 5.

Table 5. Baseline values (Experiment 2)

Group (Level)	1 (1-10)	2 (11-16)	3 (17-29)
Random baseline	0.1	0.16	0.08
Most common baseline	0.12	0.2	0.13

Results We weighed a slight variation of results compared to the combined level testing of experiment 1. Group 1 had similar classification performance from our testing but the last group had a more sparse distribution of samples compared to the combined testing of levels. On the other hand, for group 2, the extra tree classifier placed the sample texts more accurately into their corresponding levels. Those facts are also evident from the value of the metrics enlisted below. The best-performing classifier also changed for groups 1 and 3.

Table 6. Results from Experiment 2 (Grouped features)

Group (Level)	Best classifier	Accuracy	Precision	Recall	F1 score
1 (1-10)	Random forest	48.6	48.6	50.56	48.62
2 (11-16)	ExtraTree	40.32	39.70	41.95	40.15
3 (17-29)	Random forest	31.87	28.52	29.76	29.87

Table 7. Results from Experiment 3 (With important features)

Feature selection method	Top features	Best classifier	Accuracy	Precision	Recall	F1 score
Extratree	word tokens, number of different words, lexical tokens	XGBoost	32.02	27.78	28.57	28.46
RFE	word tokens, number of different words, type token ratio, mean segmented type-token ratio, ndwerz	Random forest	32.79	29.50	30.06	29.98
Univariate	corrected type token ratio, root type-token ratio, lexical types, word types, number of different words	XGBoost	28.86	26.12	27.13	26.79

6.3 Experiment 3

For this experiment, we did a feature importance check to measure the most impactful features. Feature importance refers to techniques that assign a score to input features based on how useful they are at predicting a target variable. Feature importance scores play an important role in a predictive modeling project, including providing insight into the data and model, and the basis for dimensionality reduction. For the feature importance, we applied multiple techniques e.g., extra tree classifier, recursive feature elimination, and Univariate selection. The purpose of the extra tree classifier is to fit a number of randomized decision trees to the data, and in this regard is a form of ensemble learning. Recursive Feature Elimination, or RFE for short, is effective at selecting those features (columns) in a training dataset that are more or most relevant in predicting the target variable. Another way of selecting the important features is Univariate feature selection. It works by selecting the best features based on univariate statistical tests.

Task and procedure This is the same as experiment 1.

Model This is the same as experiment 1.

Metrics This is the same as experiment 1.

Results Throughout the approaches we examined, the lexical features stand out as the most important features out of the lot. However, this better performance of lexical features was mainly due to the addition of the traditionally used features e.g., number of different words, word tokens. So, it is no wonder that these shallow features have been used in the traditional readability formulae for such a longtime. The top features varied for different approaches. We separately used the best features for training and testing across all the levels. Overall the prediction performance drastically drops in comparison to all the previous experiments. So, eliminating features costed the performance as they lose important contribution from them.

6.4 Experiment 4

An important aspect of our analysis was to check the misclassified texts and unearth the reasons behind them.

Task and procedure For misclassification analysis, we trained and tested the lexical and d-level features separately using the Extra tree classifier as it yielded best results (experiment 1). All the levels were tested for classification. A 5 fold stratified cross-validation was done and the features were normalized before feeding to the classifiers.

Model We used the Extra tree classifier.

Metrics Once trained separately for both the lexical and the d-level features separately, they were tested for prediction accuracy and the percentage of misclassification above and below their certain level was also computed.

Results Overall, **34%** and **15%** of the samples were correctly classified while the lexical and d-level features were used individually. In general, majority of the samples were placed to a lower level when lexical features were used. On the contrary, more samples were put to a higher level when d-level features were used. For example, all the samples that originally had a level of 21, 65 percent were classified to an upper level in the case of lexical features. On the other hand, 67 percent were classified to a lower level when d-level features were used.

7 LIMITATION, PITFALLS, AND FUTURE RESEARCH

Although the project started with the aim of developing a readability model for children, significant time was spent on cleaning the data and generating the features. High computation requirements also made the path difficult as d-level features take a long time to process. We started processing each of the document files but even if a single line was not processed on each document the whole feature set resulted as zero for that document. We had to change our strategy from a document level feature generation to a sentence level feature generation and removing the unprocessed sentences. Performing a sentence-level feature generation made the task way more difficult as the computation time now increased significantly.

We have listed a couple of task orders still needed to accomplish for the model development. The very first work would be tracking the misclassified texts. We have planned to allow a tolerance of 1 for readability. It means even if the model/classifier assigns a level of 4/6 to a particular text and it originally had a level of 5 then that should be accepted. Ideally, syntactic features should play a more vital role in determining the readability levels. But from our analysis, we had a contrasting result. We would dive deep into the feature values of d-level for further reasoning. A grade-level analysis is also on the horizon to finalize a model capable of classifying children's text into different levels/grades, which will immensely benefit teachers, students, and language professionals.

8 CONCLUSION

The motivation behind our task was to help teachers and children with powerful technology that would help them to select books/articles, more appropriate for children who are the future of the world. We believe such work requires attention and care to accomplish. The average American is considered to have a readability level equivalent to a 7th/8th grader (12 to 14 years old). Learning also becomes meaningful at a very early age and children lose interest in learning and trying new things if the materials are not suitable for their age level. So, formally decomposing the ambiguous notion of text readability into a compact, transparent and logical score is discernible. With the prospect of suggesting age-appropriate reading materials, we stressed the importance of developing a better readability model throughout this research effort.

ACKNOWLEDGMENTS

I would like to thank the CAST research group for providing me the annotated A-Z data sets to perform the experiments.

REFERENCES

[1] Shmuel I Becher and Uri Benoliel. 2021. Law in books and law in action: the readability of privacy policies and the gdpr. In *Consumer Law and Economics*. Springer, 179–204.

[2] Rebekah George Benjamin. 2012. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review* 24, 1 (2012), 63–88.

[3] JS Chall. 1988. The beginning years. In. BL Zakalukand SJ Samuels (eds.), Readability: its past, present, and future, Newark. *DE: International Reading Association* (1988).

[4] Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability revisited: The new Dale-Chall readability formula.* Brookline Books.

[5] Michael Collins. 1999. Head-driven statistical models for natural language parsing [Ph. D. Dissertation]. *University of Pennsylvania* (1999).

[6] Kevyn Collins-Thompson and Jamie Callan. 2005. Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology* 56, 13 (2005), 1448–1462.

[7] 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*. 193–200.

[8] Michael A Covington, Congzhou He, Cati Brown, Lorina Naci, and John Brown. 2006. How complex is that sentence? A proposed revision of the Rosenberg and Abbeduto D-Level Scale. (2006).

[9] Scott A Crossley, Stephen Skalicky, and Mihai Dascalu. 2019. Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading* 42, 3-4 (2019), 541–561.

[10] Scott A Crossley, Stephen Skalicky, Mihai Dascalu, Danielle S McNamara, and Kristopher Kyle. 2017. Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes* 54, 5-6 (2017), 340–359.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[12] William H DuBay. 2004. The Principles of Readability. *Online Submission* (2004).

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

[14] Edward Fry. 1968. A readability formula that saves time. *Journal of reading* 11, 7 (1968), 513–578.

[15] Edward Fry. 2002. Readability versus leveling. *The reading teacher* 56, 3 (2002), 286–291.

[16] Eckhardt Fuchs, Inga Niehaus, Almut Stoletzki, et al. 2014. *Das Schulbuch in der Forschung. Analysen und Empfehlungen für die Bildungspraxis.* Göttingen: V&R unipress.

[17] Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers* 36, 2 (2004), 193–202.

[18] Robert Gunning et al. 1952. Technique of clear writing. (1952).

[19] Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the third workshop on innovative use of NLP for building educational applications*. 71–79.

[20] GR Klare. 1969. The measurement of readability, lowa State University Press. (1969).

[21] Renaissance Learning. 2001. The ATOS readability formula for books and how it compares to other formulas. *Madison, WI: School Renaissance Institute* (2001).

[22] Xiaoyong Liu, W Bruce Croft, Paul Oh, and David Hart. 2004. Automatic recognition of reading levels from user queries. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. 548–549.

[23] Tim Loughran and Bill McDonald. 2014. Measuring readability in financial disclosures. *the Journal of Finance* 69, 4 (2014), 1643–1671.

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

[25] Xiaofei Lu. 2012. The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal* 96, 2 (2012), 190–208.

[26] Brian MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk. transcription format and programs*. Vol. 1. Psychology Press.

[27] Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. (1993).

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

[29] Eleni Miltsakaki and Audrey Troutt. 2007. Read-x: Automatic evaluation of reading difficulty of web text. In *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*. Association for the Advancement of Computing in Education (AACE), 7280–7286.

[30] Annette T Rabin, B Zakaluk, and S Samuels. 1988. Determining difficulty levels of text written in languages other than English. *Readability: Its past, present & future. Newark DE: International Reading Association* (1988), 46–76.

[31] Andrya LH Ramer. 1977. The development of syntactic complexity. *Journal of Psycholinguistic Research* 6, 2 (1977), 145–161.

[32] Sheldon Rosenberg and Leonard Abbeduto. 1987. Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults1. *Applied Psycholinguistics* 8, 1 (1987), 19–32.
[33] Kenneth D Royal and Kristan M Erdmann. 2018. Evaluating the readability levels of medical infographic materials for public consumption. *Journal of visual communication in medicine* 41, 3 (2018), 99–102.
[34] Lucius Adelno Sherman. 1893. *Analytics of literature: A manual for the objective study of English prose and poetry.* Ginn.
[35] Luo Si and Jamie Callan. 2003. A semisupervised learning method to merge search engine results. *ACM Transactions on Information Systems (TOIS)* 21, 4 (2003), 457–491.
[36] A Jackson Stenner. 1996. Measuring Reading Comprehension with the Lexile Framework. (1996).
[37] Susan Szabo and Becky Barton Sinclair. 2019. Readability of the STAAR Test is still misaligned. *Schooling* 10, 1 (2019), 1–12.
[38] 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*. Springer, 301–308.
[39] J Ure. 1971. Lexical density: A computational technique and some findings. *Talking about text* (1971), 27–48.
[40] 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*. 163–173.
[41] Tim vor der Brück, Hermann Helbig, Johannes Leveling, and Intelligente Informations-und Kommunikationssysteme. 2008. *The Readability Checker Delite: Technical Report.* FernUniv., Fak. für Mathematik und Informatik.
[42] Matthew J Voss. 2005. Determining syntactic complexity using very shallow parsing. (2005).
[43] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *CoRR* abs/1910.03771 (2019). arXiv:1910.03771 http://arxiv.org/abs/1910.03771

9 APPENDIX

Lexical density, originally coined by Ure [39], refers to the ratio of the number of lexical (as opposed to grammatical) words to the total number of words in a text. Lexical sophistication, also known as lexical rareness, measures "the proportion of relatively unusual or advanced words in the learner's text." Lexical variation, also labeled lexical diversity or lexical range, refers to the range of a learner's vocabulary as displayed in his or her language use. An intuitively straightforward measure of lexical variation is the number of different words (NDW) used in a language sample, which has proved to be a potentially useful measure of child language development. Type–token ratio (TTR), that is, the ratio of the number of word types (T) to the number of words (N) in a text. Mean segmental TTR (MSTTR) constitutes one way to improve TTR, which is computed by dividing a sample into successive segments of a given length and then calculating the average TTR of all segments. Other transformations of TTR include Corrected TTR (CTTR), Root TTR (RTTR), Bilogarithmic TTR (LogTTR), and the Uber Index. Verb variation is computed as the ratio of the number of verb types to the total number of verbs in a text.

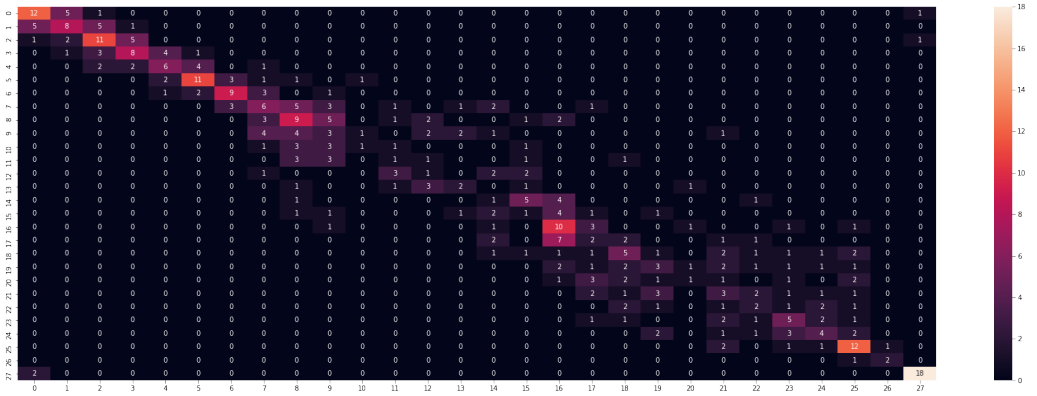

Fig. 3. Confusion matrix for A-Z: F1 score (All features)