An Equitable Framework for Automatically Assessing Children's Oral Narrative Language Abilities

Alexander Johnson[1]*, Hariram Veeramani*[1]*, Natarajan Balaji Shankar*[1]*, Abeer Alwan*[1]

[1]University of California, Los Angeles, Department of Electrical and Computer Engineering

{ajohnson49, hariram, balaji1312 }@g.ucla.edu, alwan@ee.ucla.edu

Abstract

This work proposes a novel framework for automatically scoring children's oral narrative language abilities. We use audio recordings from 3rd-8th graders of the Atlanta, Georgia area as they take a portion of the Test of Narrative Language. We design a system which extracts linguistic features and fine-tuned BERT-based self-supervised learning representation from state-of-the-art ASR transcripts. We predict manual test scores from the extracted features. This framework significantly outperforms a deterministic method based on the assessment's scoring rubric. Last, we evaluate the system performance across student's reading level, dialect, and diagnosed learning/language disabilities to establish fairness across diverse demographics of students. Using this system, we achieve approximately 98% classification accuracy of student scores. We are also able to identify key areas of improvement for this type of system across demographic areas and reading ability.

Index Terms: Children's Speech, Spoken Language Assessments, Automatic Speech Recognition, Bias Mitigation

1. Introduction

Spoken language assessments (SLA) are a pivotal tool in measuring the development of children's oral language and narrative skills. These assessments are time-consuming and often require access to a highly-trained language specialist. Great strides have been made to automate oral language tests in order to serve a greater number of students. For example, [1] explores multi-task learning as an approach to overcoming the problem of limited data in automatic oral English proficiency SLAs for Mandarin speakers. In addition, [2] compares the performance of Wav2Vec2.0 [3] and Kaldi TDNN-based [4] grapheme embeddings as features for evaluating children's phonological working memory for nonwords. Similarly, the authors of [5] use hidden states from Wav2Vec2.0 [3] to predict mispronunciations and abnormalities in children's speech. Such methods that take advantage of large pre-trained automatic speech recognition (ASR) systems seem particularly promising given the recent advancements in training strategies for architectures like HuBERT [6], WavLM [7], and Whisper [8]. However, challenges remain in automatic SLA, especially for children. Children's developing language skills and growing speech articulators cause their speech to be highly variable [9], which in turn creates challenges in recognition and assessment [10, 11]. This paper deals with automatically assessing children's SLAs for the Test of Narrative Language (TNL) [12]. This assessment tests both children's speech pronunciation and language abilities such as the ability to correctly, coherently, and completely recall a story with correct grammar (including tense, word order, and subject verb agreement). This test is popular among clinicians because it does not call for full transcriptions of the children's speech, only scoring of pre-defined test items. Machine scoring must go beyond pronunciation-only models for SLA and incorporate aspects of natural language understanding (NLU). Studies in essay scoring have used NLU to score written essays for narrative language proficiency [13, 14]. Notably, [15] combines hand-crafted linguistic features which capture advanced semantics with soft label predictions from the language model, RoBERTa [16], in a hybrid model which achieves state-of-the-art-performance readability score classification. However, further work is needed to adapt such methods to spoken language systems. [17] shows good performance in assessing children's language abilities by using ASR transcripts of spoken sentences for downstream inference of oral proficiency. A needed next step is to apply more sophisticated NLU on ASR transcripts to measure narrative abilities both within and across spoken sentences. This problem is additionally complicated by the effects of bias towards speakers. The authors of [18] show that child speakers of minority dialects are more likely to be under-rated in language proficiency or misdiagnosed as having a language impairment. The authors of [19] further point out that ASR systems also typically produce higher word error rates for speakers of low-resource dialects. Therefore, when performing downstream NLU tasks on ASR transcripts, special attention must be given to ensure that rater and system bias are not compounded in the pipeline. In this work, we present a novel system for fair children's SLAs. The system leverages ASR transcripts from large pretrained ASR systems to score children's ability to pronounce words and correctly recall elements of a story with correct grammar. We compare a string similarity-based approach with a transfer learning-based neural network approach which utilizes a fusion of self-supervised learning representations from large language models and hand-crafted features extracted from ASR transcripts. Finally, we show system performance across children's reading ability, language-impairment, and dialect to ensure inclusive and explainable performance. The novelties are 1) selection of NLP methods for automatic assessment scoring that are robust to dialect and children's speech differences 2) offering insights on how cross-domain training can improve performance in the low-resource task, and 3) a providing a detailed analysis of the system across demographics.

2. Methods

2.1. Data

This paper uses audio recordings of 184 3rd-8th grade students from the Atlanta, Georgia, area as they perform the "Test of Narrative Language (TNL) - Task 1, Story Retelling" assessment (data collected in [20]) where students are read a story by

the test administerer. The students are then asked to retell the story and graded on their ability to use the set of pre-determined test keywords from the original story-telling. These keywords contain story elements (eg. character names, locations, times, important objects, and action verbs) that must be retold in the same verb tense and order to receive credit in the test scoring. For example, if a test item contained the sentence, "**Tim eats** his lunch while **Matt** plays **football**" where the bolded words are the scored keywords, the child will receive points for two of the four keywords if they retell it as "**Tim** played **football** while **Matt ate** lunch," as the word order or tense of the other two keywords are incorrect. Each child's assessment was administered and audio recorded by a trained member of the project staff according to the TNL standardization manual protocols. The recordings were then independently scored by a speech-language pathologist and a second trained speech-language staff member. If disagreements occurred in scoring, the two scorers reviewed the audio discussed differences to come to a consensus. Each child's score was an integer value between 0 and the total number of test keywords. Recordings were taken at the child's school. Audio was recorded in stereo at a sampling rate of 48kHz. All recordings were resampled to mono audio with a sampling rate of 16kHz for experimentation. Each of the children gave a response with an average length of about 5min, resulting in approximately 16 total hours of speech. Although not necessary to the TNL protocol or the training procedure below, the project team additionally transcribed ground truth transcriptions for each audio recording. The dataset additionally contains demographic metadata on the students in the following categories: 1) the presence of reading/language impairment, 2) the student's reading ability (good or poor) as rated by a team of experienced teachers and learning specialists as a selection criterion for the study, and 3) the speaker's dialect (either African American English (AAE) or Southern American English) as labeled by the authors according to the procedure in [19].

2.2. Experiments

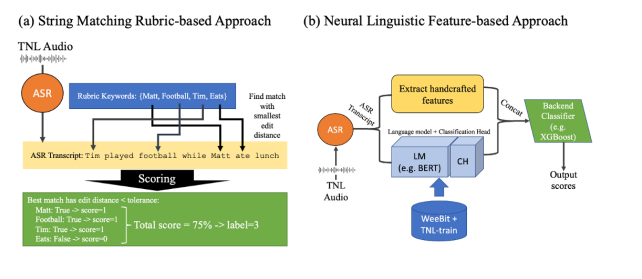

Figure 1: *Overview of a) the string-search rubric-based approach and b) the neural linguistic feature-based approach.*

We first map the test scores to discrete labels in order to formulate the assessment scoring problem as a multi-class classification task. This is intended to reduce over-fitting to negligibly small differences between scores. We sort the scores into five equally spaced histogram bins and then assign a class to each sample based on its bin, resulting in a five-class classification task. We predict the class automatically from ASR-generated transcripts. We consider ASR transcripts from Whisper, Wav2Vec2.0, and HuBERT. A 4-gram KenLM language model [21] trained on the LibriSpeech Corpus [22] is applied to the the output transcripts of each ASR system, and we report scoring accuracy both with and without the effects of the external language model. We also fine-tune HuBERT on the MyST Database [23] to explore possible improvements from training on additional children's speech data.

String-Search Rubric Matching-based scoring (SSRM) The TNL provides a comprehensive scoring rubric which assigns points to each keyword given. As a preliminary approach (shown in Figure 1a), we apply fuzzy string matching with a similarity ratio of 85% to the ASR transcripts to identify close matches to the specified keywords. This method then awards points to a student's predicted score if a word whose Levenshtein edit distance with a test keyword is less than or equal to 15% of the word length appears in the ASR transcript. We then present the accuracy and root mean square error (RMSE) of this method for each ASR systems used.

Neural Linguistic Feature-based Scoring (NLF)

A weakness of the SSRM scoring is that it only considers whether or not a close match to a keyword appears in the transcripts. It does not consider, for example, whether words were used in reference to the right characters or appeared in the correct order. These tasks require a neural network-based approach to capture finer scoring details. For this, we split the samples from the TNL into a 45-15-40, train-val-test split. We arrived at this split by starting with a 70-10-20 train-val-test split and reducing the amount of data in the training set until performance significantly worsened. This was done to best simulate the low-resource data scenario found in many children's speech responses. With this data, we employ methods used in readability assessment from [15]. We first apply transfer learning to large language models to generate soft label features from the transcripts for downstream scoring. Here, we experiment with BERT [24], RoBERTa, BART [25], and XLNET [26] using Huggingface. A 5-dim fully-connected layer is appended to the output of the large language model, and then, we fine-tuned this extended model on a combination the WeeBit Corpus [27] and the training set of the TNL data for more task-specific text-scoring. A parameter grid search over the validation set using the AdamW Optimizer found a linear learning schedule (beginning with a learning rate of 2e-5 and a weighted decay of 0.01 after 10% of the total training steps were used in warmup) and a batch size of 8 as the best system hyperparameters. The other model hyperparameters were not changed from their original implementations. The WeeBit Corpus contains short news article-style texts used for children's reading comprehension tasks. These texts are each labeled with an integer difficulty rating between 1 and 5 with difficulty 1 meant for children ages 7-8 and difficulty 5 meant for children ages 14-16. By having the network simultaneously learn to both predict difficulty levels of children's texts and scores for narrative language proficiency on spoken transcripts, we create a multitask learning framework in which the machine must learn to combine both knowledge of age-appropriate reading texts (WeeBit) and knowledge of oral language abilities (TNL). As education literature shows that children's reading proficiency and comprehension abilities are directly correlated with their oral language proficiency [28], we use this strategy to make the machine use the same weights to jointly predict both tasks. We report the accuracy of this system in predicting the TNL scores of the test set. Next, we extract the subset of the 255 hand-crafted linguistic features found optimal for the WeeBit corpus in [15] and try additional features from that study in the "Discourse", "Semantic", and "Traditional" categories which capture several measures used to score essay quality (eg. ease of identifying main topics, density of predicted entities, lexical difficulty of words used) as proposed by [15]. Finally, we concatenate the hand-

crafted features with the soft-labels given by the large language model for input to a backend classifier trained to perform score prediction (as shown in Figure 1b). We experiment with logistic regression, support vector machines, Random Forest, and XGBoost [29] and report the accuracy of each system.

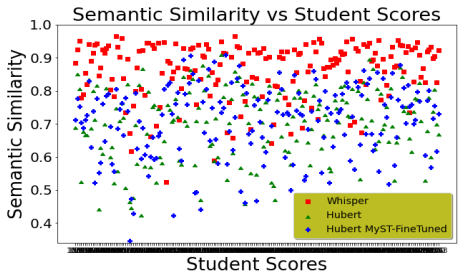

Figure 2: *Semantic Similarity between each student's ASR and ground truth transcript. ASR transcripts generated with Whisper, Hubert, and Hubert fine-tuned on MyST.*

Evaluation of Fairness: As no training is necessary for the SSRM scoring, we report the accuracy over the entire set of speakers. For the NLF scoring, we report metrics as averaged over 5 train-test splits. To ensure that the model performs fairly for diverse students, we also report test accuracy for the three demographic categories listed in Section 2.1.

3. Results and Discussion

Table 1 shows the results of the SSRM method for the different ASR transcripts in comparison to the performance on the ground truth transcripts. We report ASR WERs and the 5-class classification accuracy and RMSE. In addition to lowest overall WER, Whisper also had 93.6% precision and 93.7% recall in correctly detecting and transcribing the test keywords with a detection threshold of 85% string matching. Table 2 shows the performance of the NLF method while only using the language model with a final classification layer (no linguistic features) after being fine-tuned on the WeeBit corpus and training set of the TNL transcripts. To better understand the effects of different steps in this training pipeline, we perform an ablation study in which we remove stages from the training. Fine-tuning the best performing language model on only the TNL with no WeeBit text data gives a maximum classification accuracy of approximately 60%. Likewise, fine-tuning this language model on only the WeeBit text without using the TNL transcripts gives a maximum classification accuracy of about 58%. Next, Table 3 shows the performance of backend classifiers using a concatenation of the soft-labels from the best system in Table 2 and hand-crafted linguistic features. Table 2 shows that BERT performs equally well with either the Whisper ASR transcripts or the Hubert-finetuned on MyST transcripts. We proceed with the Whisper transcripts because of their higher semantic similarity with the groundtruth transcripts (depicted in Figure 2). Finally, we divide the student samples into the three demographic groups listed in Section 2.1 and show the performance of the best system for each group in Table 4.

A comparison of the string-matching (SSRM) approach and neural (NLF) approach shows that the machine learning method far outperforms the rubric-based baseline. The proposed system (BERT soft labels + hand-crafted linguistic features + XGBoost) achieves a classification accuracy of 98.5% using the Whisper ASR transcripts. In comparison, the rubric-based approach achieves only about 87% classification accuracy and sees marginal improvement even with the ground truth transcripts. Since the rubric-based method only considers whether or not test keywords appeared in the story and not whether they're used coherently, the performance difference between the two approaches suggests that the proposed system is able to capture grammar and logic rules used in scoring the assessments that cannot be assessed with a simple fuzzy string search. The ablation study shows a significant degradation in performance of the proposed approach when either the test set or the added WeeBit set is removed from the fine-tuning process. This further suggests that the language model only performs well given a sufficient amount of in-domain data but can make use of the correlation between reading proficiency measures (with WeeBit readability scores) and oral proficiency measures (with the TNL training set) in order to learn relationships in children's language well. Given that we only use 45% of the 184 samples in training, this approach appears to successfully deal with the low-resource data problem in children's language assessments. The demographic splits in Table 4 imply that our method performs fairly across language ability (or presence of disability), reading ability, and dialect. Across the Reading/Language Impairment demographics, the NLF method matches or outperforms the rubric-based approach in all cases. We note, however, that the rubric-based approach performs more fairly across these categories, with scores from the ASR transcript varying by less than 3% absolute value from the control students (no impairment) to the students with both a reading disorder and language impairment. The NLF method achieves nearly perfect accu-

Table 1: *ASR Word Error Rate (WER) , Classification Accuracy (C. Acc), and classification RMSE for the fuzzy string matching approach for each system*

Transcripts	WER	C. Acc	RMSE
Ground Truth	-	88.0%	0.120
Wav2Vec2	37.0%	61.4%	0.402
HuBert	46.7%	62.0%	0.413
HuBert Finetuned	42.5%	64.1%	0.407
HuBert XL	43.9%	73.3%	0.282
HuBert XL w/ 4-gram LM	38.9%	76.0%	0.282
Whisper Large	26.8%	86.4%	0.136
Whisper Large w/ 4-gram LM	**26.3%**	**87.0%**	**0.130**

Table 3: *System performance using a backend classifier to predict assessment scores from an input concatenation of hand-crafted linguistic features and soft labels from the best-performing large language model (BERT) extracted from the best ASR transcripts (Whisper). Backend classifiers tested are: Support Vector Machines (SVM), Logistic Regression (LogReg), Random Forest (RandFor), and XGBoost.*

BERT Soft Labels + Linguistic Features				
Transcripts	Classifier	C. Acc	F1 Score	RMSE
Ground Truth	SVM	96.9%	0.96	0.034
	LogReg	97.0%	0.96	0.032
	RandFor	98.5%	0.97	0.029
	XGBoost	99.2%	0.99	0.025
Whisper	SVM	96.5%	0.95	0.038
	LogReg	96.0%	0.96	0.039
	RandFor	97.6%	0.97	0.032
	XGBoost	**98.5%**	**0.98**	**0.030**

Table 2: *The classification metrics (C. Accuracy, F1-score, and RMSE) of each of the fine-tuned language models considered when predicting scores. Numbers reported are the average of 5 trials of random hold out.*

	BERT			ROBERTA			BART			XLNET		
Transcript	C. Acc	F1	RMSE	C. Acc	F1	RMSE	C. Acc	F1	RMSE	C. Acc	F1	RMSE
Ground Truth	96%	0.95	0.04	91%	0.90	0.15	80%	0.78	0.40	84%	0.83	0.16
Whisper	**96**%	**0.95**	**0.04**	90%	0.89	0.16	78%	0.77	0.43	82%	0.82	0.18
HuBert Large	95%	0.94	0.04	86%	0.85	0.27	75%	0.60	0.50	80%	0.79	0.24
HuBert Base	89%	0.88	0.11	83%	0.83	0.16	71%	0.69	0.29	80%	0.77	0.35
HuBert Base Ft	**96**%	**0.95**	**0.04**	93%	0.93	0.09	82%	0.82	0.18	82%	0.81	0.19
Wav2Vec2	85%	0.83	0.15	87%	0.84	0.25	70%	0.70	0.40	85%	0.83	0.30

Table 4: *Results for both the SSRM and NLF approaches across different student demographics. We present a breakdown of best performing ASR system (Whisper) word error rate, the classification C. Accuracy and RMSE of the system on the ground truth (GT) transcripts, and those metrics on the Whisper ASR transcripts for the following three demographic splits: 1) Type of Reading or Language Impairment from i) control - no impairment, ii) RD Only- student has reading disorder like dyslexia that does not occur with or as a secondary effect of a primary learning or language impairment or other condition, iii) RD + LI - A reading disorder that occurs with a primary Language impairment 2) Reading status from i) Poor - the student is evaluated to read at a level below their appropriate grade level or ii) Good - the student reads at or above their appropriate grade level, and 3) Dialect from i) African American English (AAE) or ii) Non-AAE - a mix of characteristics of General American English and Southern American English native to the Atlanta, Georgia Area. Note that the number of students in the Reading/Language Impairment and Reading Status demographic categories do not sum to the full 184. For this analysis, we excluded children with other disorders like ADHD that may complicate the test taking and children who were not able to be assessed for reading status into either the Poor or Good category.*

Demographic			String-Search Rubric (SSRM) Approach				Neural Linguistic Feature (NLF) Approach			
Reading/Language Impairment	WER	# of students	GT RMSE	GT C. Acc	Whisper RMSE	Whisper C. Acc	GT RMSE	GT C. Acc	Whisper RMSE	Whisper C. Acc
Control	21.1%	32	0.125	87.5%	0.125	87.5%	0.020	99.9%	0.025	99.9%
RD only	26.8%	60	0.116	88.3%	0.133	86.6%	0.061	87.5%	0.064	87.5%
RD+LI	34.5%	27	0.111	88.8%	0.148	85.0%	0.110	87.8%	0.130	85.0%
Reading Status	WER	# of students	GT RMSE	GT C. Acc	Whisper RMSE	Whisper C. Acc	GT RMSE	GT C. Acc	Whisper RMSE	Whisper C. Acc
Poor	27.0%	142	0.119	88.0%	0.140	85.9%	0.022	95.0%	0.033	89.5%
Good	21.1%	32	0.125	87.5%	0.125	87.5%	0.086	97.5%	0.094	96.9%
Dialect	WER	# of students	GT RMSE	GT C. Acc	Whisper RMSE	Whisper C. Acc	GT RMSE	GT C. Acc	Whisper RMSE	Whisper C. Acc
AAE	26.8%	116	0.119	87.9%	0.155	84.4%	0.062	94.0%	0.061	94.1%
Non-AAE	25.4%	68	0.108	88.2%	0.102	89.7%	0.013	99.2%	0.013	99.2%

racy for the control case. However, this system performs worse for students with impairments who may make non-standard language errors not present in the WeeBit training set. The complex nature and differing severities of these language disorders creates high variability in the narrative language abilities of the students in these groups. This suggests the need for additional data or data augmentation strategies to model disordered children's language in order to improve performance more fairly across these demographics. We observe a similar trend in the poor vs good reading status demographics. However, the NLF approach performs better than the rubric-based approach across both the "good" and "poor" reading students. The almost 6% drop in performance of this system from the ground truth transcripts to the Whisper transcripts for the Poor reading status group may mean that further ASR improvements are needed for the speech differences that these children display. We observe relatively unbiased performance across dialect for the proposed system. While the work in [19] demonstrates that many commercially available ASR systems give worse performance for US East-coast AAE speakers than California General American English, little has been done to compare ASR performance for AAE vs other American dialects like Georgia's Southern American English. The comparably high WERs for these two Georgia dialects shown in Table 4 demonstrate that further work is needed in understanding and improving ASR for multiple types of children's regional dialectal speech.

4. Conclusions

This work presents a novel system for automatically assessing children's oral narrative language abilities from ASR transcripts. The neural network-based approach, which combines learning representations from BERT and hand-crafted measures of language proficiency, achieves high accuracy both across all students and when evaluated specifically on students in a minority demographic (reading ability, language impairment, or dialect spoken). Future steps include expanding this method to other types of SLAs and improving the investigated ASR systems for better recognition of diverse children's speech. [1]

[1]Work supported in part by the NSF

5. References

[1] J. H. M. Wong, H. Zhang, and N. Chen, "Variations of multi-task learning for spoken language assessment," in *Proc. Interspeech 2022*, 2022, pp. 4456–4460.

[2] I. Baumann, D. Wagner, S. Bayerl, and T. Bocklet, "Nonwords pronunciation classification in language development tests for preschool children," *Proc. Interspeech 2022*, 2022.

[3] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[4] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[5] Y. Getman, R. Al-Ghezi, K. Voskoboinik, T. Grósz, M. Kurimo, G. Salvi, T. Svendsen, and S. Strömbergsson, "wav2vec2-based Speech Rating System for Children with Speech Sound Disorder," in *Proc. Interspeech 2022*, 2022, pp. 3618–3622.

[6] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, p. 3451–3460, oct 2021. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3122291

[7] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[9] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[10] S. Dutta, S. A. Tao, J. C. Reyna, R. E. Hacker, D. W. Irvin, J. F. Buzhardt, and J. H. Hansen, "Challenges remain in Building ASR for Spontaneous Preschool Children Speech in Naturalistic Educational Environments," in *Proc. Interspeech 2022*, 2022, pp. 4322–4326.

[11] G. Yeung and A. Alwan, "On the difficulties of automatic speech recognition for kindergarten-aged children," *Interspeech 2018*, 2018.

[12] R. B. Gillam and N. A. Pearson, *Test of narrative language*. Proed, 2017.

[13] D. Ramesh and S. K. Sanampudi, "An automated essay scoring systems: a systematic literature review," *Artificial Intelligence Review*, vol. 55, no. 3, pp. 2495–2527, 2022.

[14] T. Shibata and M. Uto, "Analytic automated essay scoring based on deep neural networks integrating multidimensional item response theory," in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 2917–2926.

[15] B. W. Lee, Y. S. Jang, and J. Lee, "Pushing on text readability assessment: A transformer meets handcrafted linguistic features," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 10 669–10 686. [Online]. Available: https://aclanthology.org/2021.emnlp-main.834

[16] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[17] R. Gale, J. Dolata, E. Prud'hommeaux, J. van Santen, and M. Asgari, "Automatic assessment of language ability in children with and without typical development," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020, pp. 6111–6114.

[18] H. K. Craig and J. A. Washington, "An assessment battery for identifying language impairments in african american children," *Journal of Speech, Language, and Hearing Research*, vol. 43, no. 2, pp. 366–379, 2000.

[19] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[20] E. L. Fisher, A. Barton-Hulsey, C. Walters, R. A. Sevcik, and R. Morris, "Executive functioning and narrative language in children with dyslexia," *American journal of speech-language pathology*, vol. 28, no. 3, pp. 1127–1138, 2019.

[21] K. Heafield, "KenLM: Faster and smaller language model queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, Jul. 2011, pp. 187–197. [Online]. Available: https://aclanthology.org/W11-2123

[22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[23] W. Ward, R. Cole, D. Bolaños, C. Buchenroth-Martin, E. Svirsky, and T. Weston, "My science tutor: A conversational multimedia virtual tutor." *Journal of Educational Psychology*, vol. 105, no. 4, p. 1115, 2013.

[24] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: https://doi.org/10.18653/v1/n19-1423

[25] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Annual Meeting of the Association for Computational Linguistics*, 2019.

[26] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf

[27] S. Vajjala and D. Meurers, "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*, ser. NAACL HLT '12. USA: Association for Computational Linguistics, 2012, p. 163–173.

[28] F. Huettig and M. J. Pickering, "Literacy advantages beyond reading: Prediction of spoken language," *Trends in Cognitive Sciences*, vol. 23, no. 6, pp. 464–475, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364661319300907

[29] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou *et al.*, "Xgboost: extreme gradient boosting," vol. 1, no. 4, 2015, pp. 1–4.