Automatic Scoring of Non-native Children's Spoken Language Proficiency

Khairun-nisa Hassanali[1], *Su-Youn Yoon*[2], *Lei Chen*[2]

[1]University of Texas at Dallas
[2]Educational Testing Service

khairunnisa.hassanali@gmail.com, syoon@ets.org, lchen@ets.org

Abstract

In this study, we aim to automatically score the spoken responses from an international English assessment targeted to non-native English-speaking children aged 8 years and above. In contrast to most previous studies focusing on scoring of adult non-native English speech, we explored automated scoring of child language assessment. We developed automated scoring models based on a large set of features covering delivery (pronunciation and fluency), language use (grammar and vocabulary), and topic development (coherence). In particular, in order to assess the level of grammatical development, we used a child language metric that measures syntactic proficiency in emerging language in children.

Due to acoustic and linguistic differences between child and adult speech, the automated speech recognition (ASR) of child speech has been a challenging task. This problem may increase difficulty of automated scoring. In order to investigate the impact of ASR errors on automated scores, we compared scoring models based on features from ASR transcriptions with ones based on human transcriptions. Our results show that there is potential for the automatic scoring of spoken non-native child language. The best performing model based on ASR transcriptions achieved a correlation of 0.86 with human-rated scores.

Index Terms: automated speech proficiency scoring, child language proficiency assessment, grammatical development index

1. Introduction

As English becomes a global language of communication for both education and the workplace, children from non-English speaking countries are exposed to English from an early age. Many children in the world start to learn English as a foreign language while they are elementary and middle school students. This global trend creates a strong demand to develop an objective and reliable English assessment for young learners. To address this need, an international English assessment designed to measure English language skills of non-native English-speaking children aged 8 years and above (TOEFL® Primary^{TM}) was developed. Our study presents the first effort to develop automated scoring models for responses of speaking tests in the TOEFL Primary test.

There are many unique challenges that one faces with processing and automatically scoring child language. Child speech is shorter and has more disfluencies when compared to adult speech. Due to the premature articulatory organs, child speech may have different speech patterns in pronunciation and prosody and also include more pronunciation errors. These characteristics themselves are critical issues which increase the difficulty of automated speech scoring. Furthermore, they are also critical issues for automated speech recognition (ASR) systems. These issues result in frequent speech recognition errors and reduce the reliability of the automated scores derived from the speech recognition output in turn.

Most studies on automated speech scoring have focused on adult non-native English speech. Only recently a few studies have started to develop automated scoring systems for young learners. In addition, in contrast to frequently studied topics, such as fluency [1, 2], pronunciation [3, 4, 5, 6], and intonation [7], relatively limited research has been conducted on development of the automated measurement of grammatical proficiency. To the best of the authors' knowledge, no study provides automated measures that consider the grammatical proficiency of non-native children's language.

This is our initial attempt to automatically score the narration item on the TOEFL Primary test. We explore a wide range of features covering grammar, vocabulary, coherence, pronunciation and prosody and identify features that are significantly associated with non-native children's oral proficiency. In particular, for grammar, we explore the use of the index of productive syntax, a child language metric that measures syntactic proficiency in emerging language in children. Next, we develop scoring models and evaluate their performance using small data. Our results are encouraging and show that there is potential for the automatic scoring of spoken child language.

2. Related Work

Most studies in automated scoring of non-native speech have focused on adult speech, and little work has been done in the domain of automated scoring of non-adult speech. Recently, a few studies have developed automated speech scoring systems for young learners. [8] developed an automated speech scoring system for non-native middle school students. [9] developed an automated speech scoring system for a large-scale operational test for students from kindergarten up to grade 12 (K-12) who had been previously identified as English learners (ELs). They developed systems for diverse items such as items that elicited constrained speech (e.g., reading sentences or words) or items that elicited unconstrained speech (e.g., providing a summary of audio stimuli). The systems in both studies had broad construct-coverage, but they had limited coverage for some key traits such as grammar and coherence.

Research on measurement related to grammar usage is relatively nascent in automated speech scoring. [7] includes a normalized language model score of the speech recognizer as a grammatical measure. This measures the similarity between word distributions in the response and in the language model, rather than the accuracy and diversity in grammatical expressions. More recently, based on the reliable performance in the essay scoring, syntactic complexity measures have been proposed [10, 11, 12]. Some features showed promising performance in the assessment of the adult speech, but their performance of automated measuring children's grammatical proficiency was unknown. Furthermore, children's speech tends to exhibit different patterns (e.g., use of simpler and fewer syntactic structures than adults). This leads us to the need to find new measures which consider the child's grammatical development.

Grammatical development is also an important aspect of first language acquisition. Studies in this area have developed multiple grammatical developmental indices that represent the grammatical levels reached at various stages of language acquisition. [13] proposed a revision to the D-level scale which was originally studied by [14]. The D-Level Scale categorizes grammatical development into 8 levels according to the presence of a set of diverse grammatical expressions varying in difficulty. For example, level 0 consists of simple sentences, while level 5 consists of sentences joined by a subordinating conjunction. Similarly, [15] proposed the Index of Productive Syntax (IPSyn), according to which, the presence of grammatical structures from a total of 60 structures (ranging from simple ones such as including only subjects and verbs, to more complex constructs such as conjoined sentences) is evidence of language acquisition milestones. Although these indices have been applied to broader areas (e.g., clinical research such as language impairment), limited research has been conducted in the area of second language acquisition.

[16, 17] showed that cohesion and coherence are important constructs for assessing children's language ability. They created a large set of features covering vocabulary, grammar, and cohesion using the Coh-Metrix toolkit [18] which computes cohesion and coherence metrics for written and spoken texts. They showed that these features were useful in the automatic prediction of coherence in child language narratives. Moreover, [19] explored the potential for automated indices related to speech delivery, language use, and topic development to model human judgments of the TOEFL speaking proficiency in second language samples. In their study, they used 244 transcribed TOEFL speech samples and analyzed these samples using automated indices from the Coh-Metrix toolkit, the Linguistic Inquiry Word Count (LIWC) tool [20], and the Computerized Propositional Idea Density Rater (CPIDR) tool [21]. They selected a total of 14 features and used a linear regression model to automatically predict the score. Their study showed that automated indices related to breadth of vocabulary such as word count and lexical diversity and cohesion were useful in automatic scoring. These studies suggested that the use of automated indices present in the Coh-Metrix toolkit is a promising direction for scoring of spontaneous speech from both native children and non-native adults.

Based on the promising performance of these grammar and coherence measures in assessing the native language development of children, we will apply them in the scoring of speech from young EFL (English as a Foreign Language) learners. In particular, our work is one of the first automated applications of IPSyn for scoring speech of non-native English child speakers.

In order to create grammar features, we conducted a deep syntactic analysis based on NLP technology such as tagging and parsing. The correct sentence boundary is essential for the accuracy of these NLP tools. Due to lack of sentence boundary in the speech recognition output, many researchers such as [22, 23, 24] have explored automated clause boundary detection using both lexical and prosodic cues. [11] developed clause boundary detection system in the context of automated speech scoring. Despite the frequent ungrammatical sentences and disfluencies in non-native speakers, the system achieved a comparable performance as the studies based on native data. In this study, we will use an automated clause boundary detection system in automated feature generation and discuss the impact on the accuracy of grammar features.

3. Data

3.1. Data set

The data in this study came from a pilot administration of the TOEFL Primary test. The speaking test in TOEFL Primary consists of 14 items from 6 types, and speakers are prompted to provide responses lasting between 15 and 30 seconds per item. The data included responses from a total of 463 speakers from 11 countries and 7 native languages. The speakers were aged between 8 and 12 years and had exposure to English from anywhere between less than one year to more than 6 years. The daily exposure to English could be from less than 1 hour to more than 5 hours a day in an after-school or outside school context.

In our work, we focused on one narration item. Here, the speaker was given a sequence of pictures and was asked to describe the events depicted. The communication goal of this item was to measure how well the speaker explains and sequences simple events.

Each response was rated by two trained human raters using a 5-point scoring scale, where 1 indicates a low speaking proficiency and 5 indicates a high speaking proficiency. We removed all responses that received a score of 0 or could not be scored due to a technical difficulty. The human-human Pearson's correlation for the narration item type was 0.73. The data size and the distribution of $rater_1$ scores are summarized in Table 1.

Score	1	2	3	4	5	Total
Num. of speakers	38	66	138	105	32	379
%	10%	17%	36%	28%	8%	100%

Table 1: Data size and proficiency score distribution

3.2. Transcription generation

For our experiments, we used both human transcriptions and ASR-based transcriptions. For ASR-based transcriptions, we used a high performing HMM-based speech recognizer trained on approximately 733 hours of non-native adult speech collected from 7,872 speakers. A gender independent triphone acoustic model and combination of bigram, trigram, and four-gram language models were used. A word error rate (WER) of 27% on the held-out adult speech was observed. In order to improve the ASR accuracy for the child language data, we adapted both the acoustic model (AM) and the language model (LM). For the AM adaptation, we used 137 hours of speech data from 1,625 non-native English children aged between 11 and 15 years. For the LM adaptation, we used human transcriptions of 20 sample responses chosen at random from the TOEFL Primary data. The adapted ASR system achieved 48% word error rate for the narration item type. The high word error rate can be attributed to the small size of LM adaptation data and age mismatch between AM adaptation data and TOEFL Primary data. Our expectation is that adding more responses to the training data will result in a speech recognizer with a lower error rate.

The ASR transcriptions did not have sentence boundaries, but the presence of sentence boundaries was essential for generation of automated grammar features. To address this gap, we used an automated clause boundary detection system, described in [11]. The system was trained based on the maximum entropy model and lexical and acoustic features such as word bigrams, POS tag bigrams, and pause features. The method achieved an F-score of 0.60 on the adult non-native speakers' ASR hypotheses.

4. Features

Our main goal is to find significant features in predicting proficiency levels of non-native children in narration task. For this goal, we explored a large set of features from three automated systems: the automated speech scoring system, SpeechRater^{SM} [7], for pronunciation, fluency, vocabulary, grammar; the AC-IPSyn system for grammar; and the automated text-complexity evaluation system, TextEvaluator^{TM} [25], for grammar, vocabulary, and coherence. Despite construct overlaps among systems (grammar and vocabulary), we used features from all three systems since each feature in the same construct was based on a different algorithm and did not assess the same sub-construct.

First, over 400 features were created from these three systems. Next, we performed an initial feature selection based on the correlation analysis with $rater_1$ score. In this process, we removed all the features that were not significantly correlated to the human score. From the remaining features, we then selected final feature sets using the feature selection method in the WEKA toolkit [26]. Finally, we combined all the shortlisted features from the SpeechRater, AC-IPSyn and TextEvaluator subsets and created a final feature set by performing WEKA-based automated feature selection once again.

Due to a large number of available features (over 400) and limited size of TOEFL Primary data (a total of 379 responses), we selected features using the entire data without separate training/evaluation partition. This is non-standard procedure and may result in the overestimation of model performance.

4.1. Features from the automated speech scoring system (SpeechRater)

In this study, we used features from SpeechRater, the automated speech proficiency scoring of non-native speakers. The overall structure of SpeechRater is as follows. First, it performs automated speech recognition (ASR) and yields a hypothesized sentence for a given spoken response. Next, it computes over 100 features which cover the delivery (fluency, pronunciation, prosody, and rhythm), language use (vocabulary sophistication, grammar accuracy, and complexity) and content accuracy aspects of speech. Finally, it generates a score using a scoring model. A detailed description of the system and features are available from [7, 27, 28, 29].

For our experiments, we did not use all of the features generated by the SpeechRater, but selected the features that were relevant for scoring the TOEFL Primary. For instance, we did not use features that were related to number of words, length of response, and acoustic features related to audio quality or amount of energy since these features did not directly measure proficiency.

Table 2 gives the SpeechRater features that were shortlisted through the feature selection process. The selected features assess fluency, pronunciation, prosody, grammar and vocabulary features.

4.2. Index of Productive Syntax Features

The Index of Productive Syntax (IPSyn) is a child language metric that was developed by Scarborough in 1990 [15]. IPSyn tries to measure the syntactic proficiency of a child's emerging language.

The IPSyn scores structures across the Noun Phrase (NP), Verb Phrase (VP), question and negation, and sentence categories, described briefly below:

- **Noun phrase:** Consists of structures such as adjectives, modifiers, nouns, plural nouns, two word Noun Phrases (NP) and three word NP.
- **Verb phrase:** Consists of structures such as prepositional phrases and different forms of verbs and adverbs.
- **Question and negation:** Consists of structures such as intonational questions, wh-questions, and negations.
- **Sentence:** Consists of structures that look at later-developing syntactic abilities such as the use of relative clauses, passive constructs, and tag questions.

In total, there are 56 grammatical structures (11 noun, 16 verb, 10 question and negation, and 19 sentence). Scarborough [15] gives a listing of structures defined by the IPSyn standard.

IPSyn directly samples structures whereby a given structure can receive 0 points (never occurred), 1 point (occurred once in sample), or 2 points (occurred twice or more). It requires the clinician to consider only 100 consecutive utterances in a sample and look for at most 2 unique occurrences of a structure. Since IPSyn measures language emergence, two occurrences are considered enough for this purpose. Since a poor score can be attributed to specific structures the child performed poorly on, it allows for measuring the child's progress relative to these structures.

Several of the test takers of the TOEFL Primary are beginning learners of English who are children aged 8 years and above. Since the IPSyn considers several structures with varying complexity, we felt it would be appropriate to use the scores of individual structures, the scores for each of the IPSyn categories, and the IPSyn scores as grammatical features. We hypothesized that the more complex features in the IPSyn noun, verb and sentence categories would be useful features for scoring the narration item type, since the narration item type describes a sequence of events.

We used the AC-IPSyn system, described in [30], to generate 65 grammar related features. For each of the structures in the IPSyn specification, we created a feature corresponding to each structure. Additionally, we summed up the scores for each of the noun, verb, question and negation, and sentence categories and used these as features. Finally, we used the IPSyn score which is the sum of the scores of the structures.

We did the same with the raw counts of the structures but further analysis showed that use of IPSyn structure scores as features was more promising, perhaps due to the short length of the responses. For the most part, in a short sample of less than 5 utterances there would be less than 3 exemplars of most IPSyn structures.

Table 3 gives subset of IPSyn features selected through the feature selection process. More complex structures such as two verb sentences can be found from this list. This seems apt, since describing a sequence of events corresponding to several pictures would require more than one verb.

Construct	Feature
Fluency	Number of silences per word
	Mean of silence duration in seconds
	Number of words per second
	Average over all absolute differences between each silence duration and the mean of silence duration
Grammar	A similarity score between a response and responses with score of 1 in grammatical expressions. The similarity was estimated based on the Part-of-Speech bigrams. High score means high similarity with score of 1 in grammatical accuracy and range
	A similarity score between a response and responses with score of 2 in grammatical expressions
	A similarity score between a response and responses with score of 3 in grammatical expressions
Pronunciation	Acoustic model score that compares pronunciation to reference model
	The mean of absolute differences between the test takers' normalized vowel duration and native speakers' normalized vowel durations computed from a large native speech corpus
	Acoustic model score per second
Vocabulary	The proportion of types that occurred in both a response and a reference list (most frequent 100 word types in T2K-SWAL corpus)

Table 2: Selected SpeechRater features

Construct	Feature
Grammar	Proper, mass or count noun
	Two word Noun Phrase (NP) after verb or preposition
	Three word NP
	Adverb modifying adjective or nominal
	Sum of scores of structures in the noun category
	Particle or preposition
	Prepositional phrase
	Adverb
	Regular past tense suffix
	Verb-object sequence
	Two verb sentence
	Infinitive with to
	Sentence with three or more VPs

Table 3: Selected IPSyn features

4.3. Features from the automated text complexity evaluation system (TextEvaluator)

Finally, we used the automated system which measures the complexity of any written text in English (except for poems and plays). It efficiently evaluates complexity characteristics of reading materials that are selected for use in instruction and assessment. Additionally, TextEvaluator identifies specific sources of comprehension difficulty in text. TextEvaluator reports the complexity in terms of US grade levels.

TextEvaluator uses over 270 features that are grouped into 26 feature groups. About 158 of the features are based on counts of word phrases from several word lists. The different feature groups include several categories such as adjectives, adverbs, conjuncts, determiners, modals, negation, nouns/nominalizations, pronouns, quantifiers, reflexive pronouns, verbs and wh. Additionally, there are other feature groups such as cohesion, concreteness, imageability, listenability and situation model cohesion.

Prior work by [16] and [17] showed that cohesion and coherence features were useful in the automatic prediction of coherence in child language narratives. Since TextEvaluator uses a rich feature set that includes grammar, coherence, and vocabulary features, we decided to explore the use of TextEvaluator features in building a scoring model. From the features generated by TextEvaluator, we discarded the number of words feature and those features that were highly correlated to number of words (correlation $\geq$ 0.8). We did so because these features did not necessarily measure the proficiency of the test taker.

Table 4 gives selected TextEvaluator features. We observe that the cohesion features are important in the prediction of the proficiency in the narration task. This is appropriate since the test taker describes a story based on a sequence of pictures.

Construct	Feature
Cohesion	Situational model cohesion measure
	Number of words that relate to spatial situational model cohesion
Cohesion-Grammar	Number of verbs from the verbs conversation list
	Number of ordinals
	Number of intensifiers
	Number of first person singular pronouns
	Number of third person plural pronouns
Grammar	Number of past tense verbs
Vocabulary	Flesch-kincaid grade level score
	Number of collocations
	Number of phrasal verb collocations
	Type to token ratio
	Number of words greater than 8 characters
	Number of idioms per clause
	Number of idioms per sentence

Table 4: Selected TextEvaluator features

5. Experiments

We built a linear regression model using the WEKA toolkit [26]. We built scoring models using the following two types of transcriptions:

- Human transcriptions

Transcriptions	model ID	**Feature Set**	r
ASR	model1	Speech (SpeechRater)	0.69
	model2	Text(IPSyn + TextEvaluator)	0.86
	model3	All (SpeechRater+ IPSyn + TextEvaluator)	**0.85**
Human	model4	Text (IPSyn + TextEvaluator)	0.77
	model5	All	**0.86**

Table 5: Automatic scoring of narration item type

- ASR transcriptions with sentence boundaries based on an automated sentence boundary detection system

In order to investigate the impact of features, we developed scoring models with different sets of features; speech-based features listed in Table 2, text-based features (IPSyn+TextEvaluator) which were combination of features listed in Table 3 and Table 4, and all features which were based on the final feature set by applying WEKA feature selection algorithm to the combination of all above three sets.

For all of our experiments, we used leave-one-out cross validation. In each round, we excluded one speaker from the training dataset and used him/her in the evaluation.

6. Results

In Table 5, we report the results of our experiments. We report the Pearson's correlation coefficient of the scores predicted by the scoring model and the human score. In our experiments, we use the $rater_1$ score. As we can observe from Table 5, using only SpeechRater features (model1) gives us a correlation of 0.69 for the narration item type. The best performance was obtained using the IPSyn+TextEvaluator features generated using human transcripts and SpeechRater features from the speech sample (model5). In this case, the Pearson's correlation was 0.86, which is better than the human-to-human correlation (r=0.73). When we used ASR transcriptions and IPSyn+TextEvaluator feature (model2), the Pearson's correlation was comparable to the results based on human transcripts (r=0.86).

7. Discussion

In this study, IPSyn+TextEvaluator model based on the ASR transcripts achieved better performance than the model based on the human transcripts. This result was surprising considering the ASR error rate. Based on our initial analysis, we think this may be relevant to the sentence boundary annotation issue. The human transcribers in this study tended not to mark sentence boundaries for the incomplete and ungrammatical sentences, while the automated sentence boundary detection system used in this study had the opposite tendency. As a result, there were large differences in the number of sentences; the mean number of sentences per each response was 1.85 in the human transcripts and 3.90 in the automated transcripts, respectively. This resulted in large differences between features generated from the human transcripts and features generated from the automated transcripts. This automated sentence boundary annotation is likely to be an important key factor of superior performance of the ASR-based system; there was a large performance drop (r=0.74) when the IPSyn+TextEvaluator model was trained based on the ASR transcripts without the automated sentence boundary annotation. This suggests the importance of an appropriate sentence boundary annotation system for assessment of the grammatical proficiency from spoken responses.

Due to limited size of available TOEFL Primary data, the features used in the scoring models were selected using entire data including evaluation partition, and this may result in inflation of the model performance. In future study, we will address this issue by using larger data with separate training and evaluation partition.

8. Conclusions

Automatically scoring child speech poses a lot of challenges. For one, child speech is not as developed as adult speech. The syntactic constructs used are shorter and the speech has a higher number of disfluencies when compared to adult speech. Speech recognizers that are trained on adult speech do not work as well on child speech. In our work, we built scoring models for the narration speaking item type, on the TOEFL Primary, an English assessment test aimed at elementary school children aged 8 years and above who are exposed to English as a foreign language.

We explored the use of grammar, speech and cohesion features generated using existing automated systems. We explored the use of a child language metric used to measure language acquisition skills, the Index of Productive Syntax, as features in the scoring model. The IPSyn score and scores of several of the IPSyn structures were highly correlated to the human score. We found the cohesion features extracted using TextEvaluator to be useful features. The speech features allowed for capturing characteristics of speech such as fluency and prosody.

A combination of speech, grammar and cohesion features resulted in a better performance for the narration item type than human-human correlation; 0.86 (model5) vs. 0.73 for the ASR transcriptions, and (0.86 vs. 0.73) for the human transcriptions. Our results are positive and show that building a scoring model for child language speech is promising. Since both IPSyn and TextEvaluator features provide a wide variety of features, we can use these features for scoring other item types on the TOEFL Primary, such as making a request or giving directions. Several verb and sentence structures were found to be useful features for scoring the narration item type. The grammar and cohesion features from TextEvaluator were useful features for scoring the narration item type.

9. Acknowledgement

We would like to thank Yeonsuk Cho, Diane Napolitano, Kathy Sheehan, Keelan Evanini, Klaus Zechner, Mathew Mullohand, Peter Ciemins, and Edward Getman for sharing TOEFL Primary data, TextEvaluator system, and comments.

10. References

[1] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology," *The Journal of the Acoustical Society of America*, vol. 107, no. 2, pp. 989–999, 2000.

[2] ——, "Quantitative assessment of second language learners' fluency: comparisons between read and spontaneous speech," *The Journal of the Acoustical Society of America*, vol. 111, no. 6, pp. 2862–2873, 2002.

[3] S. Witt and S. Young, "Performance measures for phone-level

pronunciation teaching in CALL," in *Proceedings of STiLL*, 1997, pp. 99–102.

[4] S. Witt, "Use of the speech recognition in computer-assisted language learning," Unpublished dissertation, Cambridge University Engineering department, Cambridge, U.K., 1999.

[5] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic pronunciation scoring for language instruction," in *Proceedings of ICASSP*, 1997, pp. 1471–1474.

[6] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic scoring of pronunciation quality," *Speech Communication*, pp. 88–93, 2000.

[7] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken english," *Speech Communication*, vol. 51, pp. 883–895, October 2009.

[8] K. Evanini and X. Wang, "Automated speech scoring for non-native middle school students with multiple task types," in *Proceedings of Interspeech 2013*, 2013.

[9] J. Cheng, Y. Z. DAntilio, X. Chen, and J. Bernstein, "Automatic assessment of the speech of young English learners," in *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, 2014.

[10] X. Lu, "Automatic analysis of syntactic complexity in second language writing," *International Journal of Corpus Linguistics*, vol. 15, no. 4, pp. 474–496, 2010.

[11] L. Chen and S.-Y. Yoon, "Detecting structural events for assessing non-native speech," in *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, 2011, pp. 38–45.

[12] M. Chen and K. Zechner, "Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech." in *Proceedings of ACL*, 2011, pp. 722–731.

[13] M. A. Covington, C. He, C. Brown, L. Naci, and J. Brown, "How complex is that sentence? A proposed revision of the Rosenberg and Abbeduto D-Level Scale," CASPR Research Report 2006-01, Athens, GA: The University of Georgia, Artificial Intelligence Center, Tech. Rep., 2006.

[14] S. Rosenberg and L. Abbeduto, "Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults," *Applied Psycholinguistics*, vol. 8, pp. 19–32, 1987.

[15] H. S. Scarborough, "Index of productive syntax," *Applied Psycholinguistics*, vol. 11, no. 01, pp. 1–22, 1990.

[16] K.-n. Hassanali, Y. Liu, and T. Solorio, "Evaluating NLP features for automatic prediction of language impairment using child speech transcripts," in *Proceedings of Interspeech 2012*, 2012.

[17] ——, "Coherence in child language narratives: A case study of annotation and automatic prediction of coherence," in *Proceedings of WOCCI 2012 - 3rd Workshop on Child, Computer and Interaction*, 2012.

[18] D. S. McNamara, M. M. Louwerse, P. M. McCarthy, and A. C. Graesser, "Coh-metrix: Capturing linguistic features of cohesion," *Discourse Processes*, vol. 47, no. 4, pp. 292–330, 2010.

[19] S. Crossley and D. McNamara, "Applications of text analysis tools for spoken response grading," *Language Learning & Technology*, vol. 17, no. 2, pp. 171–192, 2013.

[20] J. W. Pennebaker, M. E. Francis, and R. J. Booth, *Linguistic inquiry and word count: LIWC2001*. Mahway: Lawrence Erlbaum Associates, 2001.

[21] C. Brown, T. Snodgrass, S. J. Kemper, R. Herman, and M. A. Covington, "Automatic measurement of propositional idea density from part-of-speech tagging," *Behavior Research Methods*, vol. 40, no. 2, pp. 540–545, 2008.

[22] Y. Gotoh and S. Renals, "Sentence boundary detection in broadcast speech transcript," in *Proceedings of the International Speech Communication Association (ISCA) Workshop: Automatic Speech Recognition: Challenges for the new Millennium ASR-2000*, 2000.

[23] Y. Liu, "Structural event detection for rich transcription of speech," Ph.D. dissertation, Purdue University, 2004.

[24] M. Ostendorf, B. Favre, R. Grishman, D. Hakkani-Tur, M. Harper, D. Hillard, J. Hirschberg, H. Ji, J. Kahn, Y. Liu, S. Maskey, E. Matusov, H. Ney, A. Rosenberg, E. Shriberg, W. Wang, and C. Woofers, "Speech segmentation and spoken document processing," *Signal Processing Magazine, IEEE*, vol. 25, no. 3, pp. 59–69, May 2008.

[25] K. M. Sheehan, M. Flor, and D. Napolitano, "A two-stage approach for generating unbiased estimates of text complexity," in *Proceedings of the 2nd Workshop of Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, 2013, pp. 49–58.

[26] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[27] L. Chen, K. Zechner, and X. Xi, "Improved pronunciation features for construct-driven assessment of non-native spontaneous speech," in *Proceedings of HLT*, 2009, pp. 442–449.

[28] S.-Y. Yoon and S. Bhat, "Assessment of esl learners' syntactic competence based on similarity measures," in *Proceedings of EMNLP*, 2012, pp. 600–608.

[29] S.-Y. Yoon, S. Bhat, and K. Zechner, "Vocabulary profile as a measure of vocabulary sophistication," in *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, 2012, pp. 180–189.

[30] K.-n. Hassanali, Y. Liu, A. Iglesias, T. Solorio, and C. A. Dollaghan, "Automatic generation of index of productive syntax for child language transcripts," *Behavior Research Methods*, 2013.