7th ISCA Workshop on Speech and Language Technology in Education
25-26 August 2017, Stockholm, Sweden

Modeling Discourse Coherence for the Automated Scoring of Spontaneous Spoken Responses

Xinhao Wang[†], *Keelan Evanini*[‡], *Klaus Zechner*[‡], *Matthew Mulholland*[‡]

Educational Testing Service R&D
[†]90 New Montgomery Street, Suite 1500, San Francisco, CA, USA
[‡]660 Rosedale Rd., Princeton, NJ, USA
{xwang002, kevanini, kzechner, mmulholland}@ets.org

Abstract

This study describes an approach for modeling the discourse coherence of spontaneous spoken responses in the context of automated assessment of non-native speech. Although the measurement of discourse coherence is typically a key metric in human scoring rubrics for assessments of spontaneous spoken language, little prior research has been done to assess a speaker's coherence in the context of automated speech scoring. To address this, we first present a corpus of spoken responses drawn from an assessment of English proficiency that has been annotated for discourse coherence. When adding these discourse annotations as features to an automated speech scoring system, the accuracy in predicting human proficiency scores is improved by 7.8% relative, thus demonstrating the effectiveness of including coherence information in the task of automated scoring of spontaneous speech. We further investigate the use of two different sets of features to automatically model the coherence quality of spontaneous speech, including a set of features originally designed to measure text complexity and a set of surface-based features describing the speaker's use of nouns, pronouns, conjunctions, and discourse connectives in the spoken response. Additional experiments demonstrate that an automated speech scoring system can benefit from coherence scores that are generated automatically using these feature sets.
Index Terms: discourse coherence, spontaneous spoken English proficiency, automated speech scoring

1. Introduction

In recent years, much research has been conducted into developing automated assessment systems to automatically score spontaneous speech from non-native speakers with the goals of reducing the burden on human raters, improving reliability, and generating feedback that can be used by language learners [1, 2]. Various features related to different aspects of speaking proficiency have been explored, such as features for pronunciation, prosody, and fluency [3, 4, 5, 2], features for vocabulary and grammar, as well as content features [6, 7, 8, 9, 10]. However, discourse-level features have, to our knowledge, not been investigated in the context of automated speech scoring, except for our own previous work [11]. This is despite the fact that an important criterion in the human scoring rubrics for a standardized international English language speaking assessment is the evaluation of coherence, which refers to the conceptual relations between different units within a response [12].

This study focuses on modeling coherence cues for spontaneous speech, and its main contributions can be summarized as follows:

- We extended a corpus of coherence annotations for spoken responses drawn from a large-scale standardized assessment of English for the academic domain from 600 to 1440 responses (600 responses are double-annotated, 840 responses are single-annotated).
- We used these coherence annotations on a corpus of spontaneous spoken responses to demonstrate the effectiveness of using coherence cues in the task of automated speech scoring by (1) evaluating correlations between human discourse coherence scores and human holistic spoken proficiency scores, and (2) evaluating the extent to which human discourse coherence scores can improve a baseline scoring model for spontaneous speech.
- We examined two different types of feature classes to automatically model these coherence scores from human annotators, namely features from an NLP system designed to measure multiple aspects of text complexity as well as several surface-based features which were designed to represent the use of nouns, pronouns, conjunctions, and discourse connectives in a spoken response.
- We used these features for predicting human discourse coherence and evaluated (1) the prediction accuracy compared to the human coherence annotations, and (2) the improvement of a baseline speech scoring system when including these automatically predicted discourse features.

2. Related Work

Methods for automatically assessing discourse coherence in text documents have been widely studied in the context of applications such as natural language generation, document summarization, and assessment of text readability. For example, [13] measured the overall coherence of a text by utilizing Latent Semantic Analysis (LSA) to calculate the semantic relatedness between adjacent sentences. [14] introduced a model for the document-level analysis of topics and topic transitions based on Hidden Markov Models. [15] presented an approach for coherence modeling focused on the entities in the text and their grammatical transitions between adjacent sentences, and calculated the entity transition probabilities on the document level. [16] provided a summary of the performance of several different types of features for automated coherence evaluation, including features based on cohesive devices, measurements of adjacent sentence similarity, Coh-Metrix [17], word co-occurrence patterns, and entity transitions. [18] proposed a graph-based approach for modeling transitions among entities in a text (in contrast to previous entity-grid approaches, which only examined transitions between adjacent sentences) and model coherence by calculating centrality measures on the nodes in the graph.

10.21437/SLaTE.2017-23

[19] examined the use of Recurrent and Recursive Neural Network models based on word embeddings for modeling coherence and showed that they outperform other approaches on the tasks of sentence ordering and readability assessment.

In addition to these studies on well-formed text, prior research has also investigated the task of evaluating coherence in student essays, which may contain multiple spelling, vocabulary, and grammar errors, especially when produced by non-native speakers of English. Utilizing LSA and Random Indexing methods, [20] measured the global coherence of students' essays by calculating the semantic relatedness between sentences and the corresponding prompts. [21] combined entity-grid features with writing quality features produced by an automated essay assessment system to predict the coherence scores of student essays. [22] systematically analyzed a variety of coherence modeling methods within the framework of an automated assessment system for non-native free text responses and indicated that features based on Incremental Semantic Analysis (ISA), local histograms of words, the part-of-speech IBM model, and word length were the most effective. [23] applied ideas from Centering Theory to model coherence in short argumentative essays by calculating the percentage of sentences within each paragraph that have the same centers. [24] investigated the use of lexical chains (sequences of related words within a text that contribute to the continuity of meaning) through features related to the length, frequency, location, and quality of the lexical chains in an essay, and demonstrated that these features can effectively model writing proficiency scores for essays written by native and non-native speakers of English.

In contrast to these previous studies involving well-formed written text or learners written texts containing errors, our previous work provided a corpus of coherence annotations on 600 spoken responses from an English language assessment, and examined several features that have been used in the context of learners written essays based on human transcriptions of these spoken responses in the task of automatic prediction of human coherence scores [11]. The current study extends this research effort in the following three ways: 1) obtaining a larger set of spoken responses annotated with coherence scores (N=1,440); 2) extending the construct coverage of an automated speech scoring system by modeling the coherence quality scores using two new classes of features; and 3) employing these two new classes of features based on automatic speech recognition output in an attempt to measure the coherence quality of spontaneous speech.

In a related study, Hassanali et al. investigated coherence modeling for spoken language in the context of a story retelling task for the automated diagnosis of children with language impairment [25]. They annotated transcriptions of children's narratives with coherence scores as well as markers of narrative structure and narrative quality; furthermore they built models to predict the coherence scores based on Coh-Metrix features and the manually annotated narrative features. The current study differs from this one in that it deals with free spontaneous spoken responses provided by students at a university level; these responses therefore contain more varied and more complex information than the child narratives.

3. Data

3.1. Spoken language corpus

The data used in this study was drawn from the TOEFL® Internet-based test (TOEFL® iBT), a large-scale standardized assessment of English for non-native speakers, which assesses English communication skills for academic purposes [12]. The Speaking section of TOEFL iBT contains six tasks, each of which requires the test taker to provide an extended response containing spontaneous speech. In total, we collected 1,440 spoken responses from the TOEFL iBT assessment, including 240 responses from each of six different test questions. These six test questions comprise two different speaking tasks: 1) providing an opinion based on personal experience (N = 480 responses) and 2) summarizing or discussing material provided in a reading and/or listening passage (N = 960 responses). The spoken responses were all manually transcribed, and the transcriptions include standard punctuation and capitalization. The average number of words per response was 113.8 (SD = 33.6) and the average number of sentences was 4.8 (SD = 2.1).

The spoken responses were all provided with holistic English proficiency scores on a scale of 1 - 4 by expert human raters in the context of operational scoring for the spoken language assessment. The scoring rubrics[1] address the following three main aspects of speaking proficiency: delivery (pronunciation, fluency, prosody), language use (grammar and lexical choice), and topic development (content and coherence). In order to ensure a sufficient quantity of responses from each proficiency level for training and evaluating the coherence prediction features, the spoken responses selected for this study were balanced based on the human scores as follows: 60 responses were selected randomly from each of the 4 score points (1 - 4) for each of the 6 test questions.

3.2. Annotation

The coherence annotation guidelines used for the spoken responses in this study were modified based on the annotation guidelines developed for written essays described in [21]. According to these guidelines, expert annotators provided each response with a score on a scale of 1 - 3. The three score points were defined as follows: 3 = highly coherent (contains no instances of confusing arguments or examples), 2 = somewhat coherent (contains some awkward points in which the speaker's line of argument is unclear), 1 = barely coherent (the entire response was confusing and hard to follow; it was intuitively incoherent as a whole and the annotators had difficulties in identifying specific weak points). For responses receiving a coherence score of 2, the annotators were requested to highlight the specific awkward points in the response. In addition, the annotators were specifically required to ignore disfluencies and grammatical errors as much as possible; thus, they were instructed not to label sentences or clauses as awkward points solely because of the presence of disfluent or ungrammatical speech [11].

Two annotators (not drawn from the pool of expert human raters who provided the holistic scores) first made independent coherence annotations for 600 spoken responses, including 25 samples from each of the 4 score levels of speaking proficiency for each of the 6 test questions. The distribution of annotations across the three score points is presented in Table 1 (numbers repeated here from [11] for convenience). The two annotators achieved a moderate inter-annotator agreement [26] of κ = 0.68 on the 3-point scale of coherence scores. Subsequently, the same two annotators provided coherence annotations for the remaining 840 responses in the corpus using the following approach: each annotator provided a single annotation for 420 responses from 3 test questions, i.e., 35 responses from each score

[1] https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf

Table 1: *Distribution of coherence scores from two annotators, where 600 responses receive double coherence scores and 840 responses receive single scores.*

	Annotator	Coherence Scores		
		1	2	3
Double Annotation	# 1	160 (26.7%)	278 (46.3%)	162 (27%)
	# 2	125 (20.8%)	251 (41.8%)	224 (37.3%)
Single Annotation	# 1	112 (26.7%)	156 (37.1%)	152 (36.2%)
	# 2	95 (22.6%)	162 (38.6%)	163 (38.8%)

level for each of 3 test questions.

4. Method

In order to verify the effectiveness of the proposed coherence cues in the assessment of speaking proficiency, we conducted the following four experiments:

- correlating human discourse coherence scores with human holistic scores (for each spoken response);
- comparing a baseline scoring model for predicting human holistic scores that does not use coherence information with an extended scoring model using such coherence information (human coherence scores);
- predicting human coherence scores using two classes of low-level features; and
- comparing a baseline scoring model for predicting human holistic scores that does not use coherence information with an extended scoring model using such coherence information which is generated based on two classes of coherence features.

4.1. Correlation between human coherence scores and human holistic speaking scores

For the first experiment, we extracted two types of features based on the human annotations of the spoken responses, namely the coherence scores and the number of awkward points identified. These two features are then correlated with the holistic proficiency scores for each response. For the double annotation set (600 responses), the average coherence scores from two annotators were used; and the union set of awkward points identified by either annotator on each response was used.

4.2. Comparison between baseline scoring model and extended scoring model using human coherence scores

For the second experiment, these coherence cues from human annotations were further investigated within a context of an automated spoken language assessment system, SpeechRater^{SM} [27, 1]. SpeechRater can automatically generate various features to assess different aspects of spontaneous speech, including pronunciation, prosody, fluency, vocabulary, grammar , as well as content. In this study, we employed 12 of these features to measure different aspects of delivery and language use. These features were either extracted directly from the speech signal or were based on the output of an automatic speech recognition system within SpeechRater. Both the training and evaluation sets that were used to develop the speech recognizer consist of similar spoken responses drawn from the same assessment and do not overlap with the data sets included in this study; its word error rate on a held-out ASR evaluation set is around 28%..

This evaluation was conducted on the double-annotated responses only. Classification models were built to automatically predict the holistic speaking proficiency scores. 10-fold cross-validation was performed using the Random Forest classifier from SKLL[2], a Python toolkit that simplifies the running of common Scikit-Learn experiments [28]. The classification accuracy, i.e., the percentage of correctly predicted holistic scores, and the Pearson correlation coefficient (r) between the predicted scores and the human scores were used as evaluation metrics. Specifically, we compared (1) the baseline model consisting of the 12 SpeechRater features only; and (2) an extended model where average human coherence scores and the number of merged awkward points were added to the baseline feature set separately.

4.3. Discourse coherence features

In the third and fourth experiments, we describe how we explored ways to model and automatically predict the human discourse coherence scores and then verified that the automatic scoring system for speaking proficiency prediction can also benefit from these automatically produced coherence scores.

In summary, we use two classes of features for the prediction of human discourse coherence scores: (1) features from TextEvaluator (TE) [29], a system originally designed for evaluating the complexity of reading passages, and (2) surface-based features looking at distributions of certain words and word classes that may indicate discourse coherence.

For these experiments, the 600 double-annotated responses were used as a training set and the 838 single-annotated responses were used as the evaluation set for the task of automatic prediction of human coherence scores.[3]

4.3.1. TextEvaluator features

TextEvaluator[4] is a tool that employs a variety of natural language processing techniques, as well as linguistic resources such as word lists, to generate more than 300 features measuring multiple aspects of sentence structure, vocabulary difficulty, connections across ideas, and organization [29]. In this study, we explore the use of these features to model discourse coherence in spontaneous, non-native speech.

TextEvaluator features were first examined on the training set of 600 double-annotated responses, which were extracted based on both the human transcriptions and the automated transcriptions of the spoken responses (ASR output). The Pearson correlation coefficients (r) of these features with the average human coherence scores were used as the evaluation metric. Table 2 shows the distributions of the absolute values of feature correlations, which can be grouped into 6 bins. There are around 150 features with very low correlations, $r < 0.1$, and around 30 features with moderate correlations, $r \geqslant 0.4$, which indicates that these features can potentially contribute to the automatic

[2]Downloaded from `https://github.com/EducationalTestingService/skll`

[3]Two responses from the single annotation set of 840 responses were removed from the experimental set due to speech recognition failures.

[4]`https://textevaluator.ets.org`

Table 2: *Distribution of the absolute values of Pearson correlation coefficients (r) of TextEvaluator features with the average human coherence scores. Features were separately extracted on the human transcriptions and the ASR output.*

	Transcriptions	ASR
$r < 0.1$	143	153
$0.1 \leqslant r < 0.2$	92	84
$0.2 \leqslant r < 0.3$	36	33
$0.3 \leqslant r < 0.4$	31	33
$0.4 \leqslant r < 0.5$	27	27
$0.5 \leqslant r < 0.55$	4	3

Table 3: *Pearson correlation coefficients (r) of surface-based features (word-based features and connective chain features) with the average human coherence scores, extracted on the human transcriptions and the ASR output separately N=600).*

Features	Transcription	ASR
num_pronouns	0.204	0.186
ratio_pronoun_nountype	-0.128	-0.106
num_conjunctions	0.174	0.209
num_connective_types	0.381	0.352
num_connective_tokens	0.337	0.33
connective_chain_bleu_item	0.068	0.155
connective_chain_ed_generic	0.282	0.268

prediction of coherence scores. We chose to extract 162 features for further experiments whose correlations were greater than or equal to 0.1 for both the human transcriptions and the ASR output.

4.3.2. Surface-based features

In addition to using the NLP-based features in TextEvaluator, we also developed a set of simple features which were designed to be more robust towards errors from the ASR system (as well as ungrammatical speech produced by test takers); these capture the use of nouns, pronouns, conjunctions, and discourse connectives in a test taker's spoken response. For this purpose, the discourse connective list from the Penn Discourse Treebank (PDTB) [30] was used. Various basic features were computed based on occurrence counts, such as the number of nouns, pronouns, conjunctions, as well as discourse connectives (counted based on both word types and word tokens), and the ratios between these counts were also extracted. A preliminary evaluation was conducted to examine the correlations of these features with the average human coherence scores on the 600 double-annotated responses. And only features with absolute correlations greater than 0.1 on both the human transcriptions and the ASR output were included for further experiments. The first 5 features shown in Table 3 are based on word counts.

Loosely inspired by related work [15], we also designed additional features that capture the global coherence, represented by the use of pronouns, conjunctions and discourse connectives across an entire test response. In order to obtain these features, first, a reference corpus containing high-scoring responses was collected, and then a connective chain was extracted from each reference response, where only the pronouns, conjunctions, and discourse connectives were retained and all other words were removed from the response. Given a test response, a similar connective chain can be also extracted. Then by comparing the similarity of the test chain with each of the reference chains, the maximum similarity or the minimum distance can be computed as a feature to measure the proper use of the connective sequence in a test response. The following three evaluation metrics were investigated to evaluate the similarity between two chains: BLEU score [31], edit distance, and word error rate (which is a normalized edit distance).

Furthermore, the reference chains can be built in either an item-specific or generic manner: item-specific references were drawn from responses to the same test question as the test response; generic references were drawn from multiple different test questions. In this work, the reference samples were extracted from a corpus that was used to train the speech recognizer in SpeechRater and did not overlap with the discourse coherence annotation corpus used for this study. Around 200 - 260 responses with the highest human holistic speaking scores were obtained for each test question; in total, 1,395 responses across 6 test questions were collected as references. A preliminary experiment indicated that the BLEU similarity with the item-specific models, i.e., connective_chain_bleu_item, and the edit distance with the generic models, i.e., connective_chain_ed_generic, can achieve moderate correlations that are higher than the other model configurations. The performance of these two features is shown in Table 3.

4.4. Automated prediction of human coherence scores

In order to model the human coherence scores, regression models were built with the TextEvaluator and surface-based features described above. In this experiment, the training set contained the 600 double-annotated responses, and the test set contained the 838 single-annotated responses. Since the average coherence scores were used for model training, a regression model was used instead of a classification model for this experiment, specifically, the Random Forest Regressor from SKLL [28]. The Pearson correlation coefficients of the automatically predicted scores with the human-annotated coherence scores was taken as the evaluation metric. Features were separately extracted on the human transcriptions and the ASR output.

4.5. Automated prediction of speaking proficiency scores

Since this work aims to improve the performance of an automated speech scoring system by modeling the discourse coherence of spontaneous speech, in the final experiment, we further investigated the integration of the automatically predicted coherence scores into the classification models for the automatic prediction of holistic speaking proficiency scores. This experiment was conducted on the 838 single-annotated responses, and 10-fold cross validation was performed. Both the classification accuracy and correlations of the automatically predicted holistic scores with human holistic proficiency scores were evaluated. The baseline system was built by only using the 12 SpeechRater features as described in Section 4.2.

As described in the above Section 4.4, regression models predicting the human coherence scores were trained on the 600 double-scored responses and then used to predict the coherence scores for the 838 responses included in this evaluation, where automatic speech recognition output was used for feature extraction. Finally, results from a system using both

Table 4: *Improvement to an automated speech scoring system by adding human-assigned coherence scores, labeled as Coh, and numbers of human-identified awkward points, labeled as nAwk. Both the classification accuracy and the Pearson correlation coefficient r between the experts' speaking proficiency scores and the automatic scores are reported.*

Features	Accuracy	*r*
SpeechRater	54%	0.728
SpeechRater + Coh	58.2%	0.779
SpeechRater + nAwk	58.5%	0.771
SpeechRater + Coh + nAwk	59%	0.782

the SpeechRater features and the human-annotated coherence scores were also reported for comparison.

5. Results

5.1. Correlation between human coherence scores and human holistic speaking scores

The average coherence scores on the set of 600 double-annotated responses correlate with the proficiency scores at r = 0.656, and the number of the union set of identified awkward points correlates with proficiency scores at r = -0.626. This indicates that the assessment of spoken language proficiency could benefit greatly from modeling the coherence cues proposed in this study. The correlations with proficiency scores for the coherence scores and the number of awkward points on the single-annotated set of 840 responses were r = 0.615 and r = -0.597, respectively.

5.2. Comparison between baseline scoring model and extended scoring model using human coherence scores

The average accuracy and correlation across 10 folds are reported in Table 4. The baseline system (with the 12 SpeechRater features) obtained an accuracy of 54%. Furthermore, by adding the average annotated coherence scores or the number of identified awkward points, the accuracy can be improved to 58.2% and 58.5% respectively. These experimental results demonstrate that the automatic scoring system can benefit from the coherence cues directly extracted from human annotations.

5.3. Automated prediction of human coherence scores

As shown in Table 5, the TextEvaluator feature set, the union of word-based and connective chain features, and their combination were examined in this regression task. The TextEvaluator and surface-based and connective chain features extracted from the human transcriptions can achieve correlations of 0.53 and 0.379 respectively. However, when the automatic speech recognition output was used, the correlations decreased to 0.498 and 0.299, respectively, due to the presence of recognition errors. The further combination of all feature sets cannot result in any correlation improvement, which may be due to the fact that the 7 surface-based features can potentially be subsumed in the much larger set of 162 TextEvaluator features.

5.4. Automated prediction of speaking proficiency scores

As shown in Table 6, in addition to the annotated coherence scores, labeled as AnnoCoh, three different sets of automatically predicted coherence scores were compared for classification model building, which were separately generated with TextEvaluator features, labeled as PredCoh_TextEvaluator, with surface-based features (including the connective chain features), labeled as PredCoh_Surface, and with the combination of TextEvaluator and surface-based features, labeled as TextEvaluator&Surface. The experimental results indicate that by adding the human-annotated coherence scores, the classification accuracy can be improved from 56% to 60.7%. In contrast, the automatically predicted coherence scores based on TextEvaluator features can improve the accuracy to 59.4%. When only using 7 surface-based and connective chain features to model the coherence quality, the generated automatic coherence scores can still improve the classification accuracy from 56% to 58.4% and correlation from 0.736 to 0.752. However, as also shown in previous results, when using the automatic coherence scores predicted with the combination of the TextEvaluator and the surface-based features, the system performance cannot be further improved.

Table 5: *Pearson correlation coefficients (r) of automatically predicted coherence scores with gold coherence scores from human annotations*

Features	Transcription	ASR
TextEvaluator	0.53	0.498
Surface-based	0.379	0.299
TextEvaluator + Surface-based	0.529	0.495

Table 6: *Improvement to a baseline automated speech scoring system by adding automatically predicted coherence scores, using automatically transcribed speech and various types of coherence features. A classification model built with SpeechRater features and human assigned coherence scores was also added for comparison.*

	Accuracy	r
SpeechRater baseline	56.0%	0.736
SpeechRater + AnnoCoh	60.7%	0.795
SpeechRater + PredCoh_TextEvaluator	59.4%	0.722
SpeechRater + PredCoh_Surface	58.4%	0.752
SpeechRater + TextEvaluator&Surface	59.4%	0.773

6. Discussion and Conclusion

In this study, we presented a corpus of coherence annotations for spontaneous spoken responses, and the analyses on these annotations showed that an automated speech scoring system can benefit from modeling the coherence of spoken responses. Based on this finding, two different sets of linguistic features were employed to model the discourse coherence of spontaneous speech, and the automatically generated coherence scores were further examined in the automatic prediction of holistic speaking proficiency scores. Experimental results showed that the performance of an automated speech scoring system can be improved by automatically modeling the coherence quality scores based on automatic speech recognition output and then introducing the generated coherence cues to measure the coher-

ence of spontaneous speech.

While TextEvaluator features and word-based features can be computed directly for each spoken response, the coherence chain features need a pre-scored corpus of spoken responses to build a reference model, thereby putting this feature class at a disadvantage compared to the other classes. In this study, we only use the first-best ASR hypotheses for further processing (feature generation); however, we could also look into obtaining additional information from the recognizer lattice and/or from an ASR N-best list, thereby potentially improving feature performance for ASR output.

Finally, since the correlations between human coherence scores and human holistic speaking proficiency scores are quite high, it is conceivable that the additional step of human annotation of discourse coherence could be skipped, and instead of using the coherence features to first predicting human coherence scores as an intermediate step, they might be used directly as part of the SpeechRater scoring model for predicting human holistic speaking proficiency scores. In future work, we will also attempt to develop more effective discourse-related features which are more robust to ASR recognition errors.

7. References

[1] 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(10), pp. 883–895, 2009.

[2] D. Higgins, , X. Xi, K. Zechner, and D. Williamson, "A three-stage approach to the automated scoring of spontaneous spoken responses," *Computer Speech and Language*, vol. 25, pp. 282–306, 2011.

[3] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech," *Acoustical Society of America*, vol. 111(6), pp. 2862–2873, 2002.

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

[5] C. J., "Automatic assessment of prosody in high-stakes English tests," in *Proceedings of Interspeech*, 2011, pp. 27–31.

[6] S.-Y. Yoon, S. Bhat, and K. Zechner, "Vocabulary profile as a measure of vocabulary sophistication," in *Proceedings of the 7th Workshop on the Innovative Use of NLP for Building Educational Applications*, 2012, pp. 180–189.

[7] 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.

[8] S.-Y. Yoon and S. Bhat, "Assessment of ESL learners syntactic competence based on similarity measures," in *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2012, pp. 600–608.

[9] M. Chen and K. Zechner, "Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics and the Human Language Technologies Conference*, 2011, pp. 722–731.

[10] S. Xie, K. Evanini, and K. Zechner, "Exploring content features for automated speech scoring," in *Proceedings of NAACL-HLT*, 2012, pp. 103–111.

[11] X. Wang, K. Evanini, and K. Zechner, "Coherence modeling for the automated assessment of spontaneous spoken responses," in *Proceedings of NAACL-HLT*, 2013, pp. 814–819.

[12] ETS, "The official guide to the TOEFL® test," *Fourth Edition, McGraw-Hill*, 2012.

[13] P. W. Foltz, W. Kintsch, and T. K. Landauer, "The measurement of textual coherence with latent semantic analysis," *Discourse Processes*, vol. 25(2-3), pp. 285–307, 1998.

[14] R. Barzilay and L. Lee, "Catching the drift: Probabilistic content models, with applications to generation and summarization," in *Proceedings of NAACL-HLT*, 2004, pp. 113–120.

[15] R. Barzilay and M. Lapata, "Modeling local coherence: An entity-based approach," in *Proceedings of ACL*, 2005, pp. 141–148.

[16] E. Pitler, A. Louis, and A. Nenkova, "Automatic evaluation of linguistic quality in multi-document summarization," in *Proceedings of ACL*, 2010, p. 544554.

[17] A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai, "Coh-metrix: Analysis of text on cohesion and language," *Behavior Research Methods, Instruments, & Computers*, vol. 36(2), pp. 193–202, 2004.

[18] C. Guinaudeau and M. Strube, "Graph-based local coherence modeling," in *Proceedings of ACL*, 2013, pp. 93–103.

[19] J. Li and E. Hovy, "A model of coherence based on distributed sentence representation," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013, pp. 2039–2048.

[20] D. Higgins, J. Burstein, D. Marcu, and C. Gentile, "Evaluating multiple aspects of coherence in student essays," in *Proceedings of NAACL-HLT*, 2004, pp. 185–192.

[21] J. Burstein, J. Tetreault, and S. Andreyev, "Using entity-based features to model coherence in student essays," in *Proceedings of NAACL-HLT*, 2010, pp. 681–684.

[22] H. Yannakoudakis and T. Briscoe, "Modeling coherence in esol learner texts," in *Proceedings of the 7th Workshop on the Innovative Use of NLP for Building Educational Applications*, 2012, pp. 33–43.

[23] V. Rus and N. Niraula, "Automated detection of local coherence in short argumentative essays based on centering theory," in *Proceedings of the 13th international conference on Computational Linguistics and Intelligent Text Processing (CICLing'12)*, 2012, pp. 450–461.

[24] S. Somasundaran, J. Burstein, and M. Chodorow, "Lexical chaining for measuring discourse coherence quality in test-taker essays," in *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, 2014, pp. 950–961.

[25] Y. L. Khairun-nisa Hassanali and T. Solorio, "Coherence in child language narratives: A case study of annotation and automatic prediction of coherence," in *Proceedings of the Interspeech Workshop on Child, Computer and Interaction*, 2012.

[26] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33(1), pp. 159–174, 1977.

[27] K. Zechner, D. Higgins, and X. Xi, "Speechrater^{SM}: A construct-driven approach to scoring spontaneous non-native speech," in *Proceedings of SLaTE*, 2007, pp. 128–131.

[28] D. Blanchard, N. Madnani, and M. Heilman, "SKLL: Scikit-learn laboratory," https://github.com/EducationalTestingService/skll/, May 2016.

[29] D. Napolitano, K. M. Sheehan, and R. Mundkowsky, "Online readability and text complexity analysis with TextEvaluator," in *Proceedings of NAACL-HLT*, 2015, pp. 96–100.

[30] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber, "The Penn Discourse Treebank 2.0," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2008.

[31] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of ACL*, 2002, pp. 311–318.