Using Rhetorical Structure Theory to Assess Discourse Coherence for Non-native Spontaneous Speech

Xinhao Wang[1], **Binod Gyawali**[2], **James V. Bruno**[2], **Hillary R. Molloy**[1],
Keelan Evanini[2], **Klaus Zechner**[2]
Educational Testing Service
[1]90 New Montgomery St #1500, San Francisco, CA 94105, USA
[2]660 Rosedale Road, Princeton, NJ 08541, USA
{xwang002, bgyawali, jbruno, hmolloy}@ets.org
{kevanini, kzechner}@ets.org

Abstract

This study aims to model the discourse structure of spontaneous spoken responses within the context of an assessment of English speaking proficiency for non-native speakers. Rhetorical Structure Theory (RST) has been commonly used in the analysis of discourse organization of written texts; however, limited research has been conducted to date on RST annotation and parsing of spoken language, in particular, non-native spontaneous speech. Due to the fact that the measurement of discourse coherence is typically a key metric in human scoring rubrics for assessments of spoken language, we conducted research to obtain RST annotations on non-native spoken responses from a standardized assessment of academic English proficiency. Subsequently, automatic parsers were trained on these annotations to process non-native spontaneous speech. Finally, a set of features were extracted from automatically generated RST trees to evaluate the discourse structure of non-native spontaneous speech, which were then employed to further improve the validity of an automated speech scoring system.

1 Introduction

The spread of English as the main global language for education and commerce is continuing, and there is a strong interest in developing assessment systems that can 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 (Zechner et al., 2009; Higgins et al., 2011). Various features related to different aspects of speaking proficiency have been explored, such as features for pronunciation, prosody, and fluency (Cucchiarini et al., 2002; Chen et al., 2009; Cheng, 2011; Higgins et al., 2011), as well as features for vocabulary, grammar, and content (Yoon et al., 2012; Chen and Zechner, 2011; Yoon and Bhat, 2012; Chen and Zechner, 2011; Xie et al., 2012; Qian et al., 2016).

Discourse coherence, which refers to how well a text or speech is organized to convey information, is an important aspect of communicative competence, as is reflected in human scoring rubrics for assessments of non-native English (ETS, 2012). However, discourse-level features have rarely been investigated in the context of automated speech scoring. In order to address this deficiency, this study aims to explore effective means to automate the analysis of discourse and the measurement of coherence in non-native spoken responses, thereby improving the validity of an automated scoring system.

Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) is one of the most influential approaches for document-level discourse analysis. It can represent a document's discourse structure using a hierarchical tree in which nodes are recursively linked with rhetorical relations and labeled with *nucleus* or *satellite* tags to depict the importance of the child nodes in a relation. In our previous study (Wang et al., 2017a), RST-based discourse annotations were obtained on a corpus of 600 spontaneous spoken responses provided by non-native English speakers in the context of an English speaking proficiency assessment. In this paper, we continued this line of research, and made further contributions as follows:

- A larger annotated corpus consisting of 1440 non-native spontaneous spoken responses was obtained using an annotation scheme based on the RST framework. In addition to the previously annotated 600 responses (Wang et al., 2017a), annotations on additional 840 responses were obtained to enlarge the data set that can be used to train

Proceedings of Discourse Relation Parsing and Treebanking (DISRPT2019), pages 153–162
Minneapolis, MN, June 6, 2019. ©2019 Association for Computational Linguistics

an automatic RST parser. When comparing the annotations from two independent human experts on 120 responses, the resulting micro-averaged F1 scores on the three different levels of span, nuclearity, and relation[1] are 86.8%, 72.2%, and 58.2%, respectively.

- Based on all these manual annotations, automatic RST parsers were trained and evaluated. When comparing the automatically generated trees with double annotations from each of the two human experts separately, the F1 scores on the three levels of span, nuclearity, and relation are 76.1%/77.0%, 57.6%/59.7%, and 42.6%/44.4%, respectively.

- A set of RST-based features were introduced to measure the discourse structure of non-native spontaneous speech, where 1) an automatic speech recognizer (ASR) was used to transcribe the speech into text; 2) the aforementioned automatic parsers were applied to build RST trees based on the ASR output; 3) a set of features extracted from the automatic trees were explored, and the results show that these discourse features can predict holistic proficiency scores with an accuracy of 55.9%. Finally, these features were used in combination with other types of features to enhance the validity of an automated speech scoring system.

2 Previous Work

RST is a descriptive framework that has been widely used in the analysis of the discourse organization of written texts (Taboada and Mann, 2006b) and has also been applied to various natural language processing tasks, including language generation, text summarization, and machine translation (Taboada and Mann, 2006a). In particular, the availability of the RST Discourse Treebank (Carlson et al., 2001), with annotations on a selection of 385 Wall Street Journal articles from the Penn Treebank[2], has facilitated RST-based discourse analysis of written texts, since it provides a standard benchmark for comparing the performance of different parsers. A wide range of techniques have been applied to this task, and document-level discourse parsers are available (Marcu, 2000a; Sagae, 2009; Hernault et al., 2010; Joty et al., 2013; Feng and Hirst, 2014; Li et al., 2014; Ji and Eisenstein, 2014; Li et al., 2016; Liu and Lapata, 2017; Braud et al., 2017; Wang et al., 2017c). Morey et al. (2017) replicated the same evaluation procedure on 9 recent parsers, and indicated that the recent gains in discourse parsing can be attributed to the distributed representations.

Another important application of RST closely related to our research is the automated evaluation of discourse in student essays. For example, one study used features for each sentence in an essay to reflect the status of its parent node as well as its rhetorical relation based on automatically parsed RST trees, with the goal of providing feedback to students about the discourse structure in their essay (Burstein et al., 2003). Another study compared features derived from deep hierarchical discourse relations based on RST trees with features derived from shallow discourse relations based on Penn Discourse Treebank (PDTB) annotations (Prasad et al., 2008) and demonstrated the positive impact of using deep discourse structures to evaluate text coherence (Feng et al., 2014).

Related work has also been conducted to analyze discourse relations in spoken language, which is produced and processed differently from written texts (Rehbein et al., 2016), and often lacks explicit discourse connectives that are more frequent in written language. For example, RST has been used to analyze the semi-structured interviews of Alzheimer's patients (Paulino and Sierra, 2017; Paulino et al., 2018).

However, the annotation scheme with shallow discourse structure and relations from the PDTB (Prasad et al., 2008) has been generally used for spoken language (Demirsahin and Zeyrek, 2014; Stoyanchev and Bangalore, 2015) instead of the rooted-tree structure that is employed in RST. For example, Tonelli et al. (2010) adapted the PDTB annotation scheme to annotate discourse relations in spontaneous conversations in Italian, and Rehbein et al. (2016) compared two frameworks, PDTB and Cognitive approach to Coherence Relations (CCR) (Sanders et al., 1992), for the annotation of discourse relations in spoken language.

Regarding the measurement of discourse coherence in the automated assessment of spoken language, our previous work (Wang et al., 2013,

[1] In this paper, all the reported results on the relation level use the full labels of both nuclearity and relation for evaluation.

[2] https://catalog.ldc.upenn.edu/LDC2002T07

2017b) obtained an annotated corpus of non-native spontaneous speech in which each response was assigned a coherence score on a scale of 1 to 3, and several surface-based features were used to count the use of nouns, pronouns, conjunctions, and discourse connectives. However, that research did not investigate features that can actually represent the hierarchical discourse structure of spoken responses as described in the RST framework.

In contrast to previous studies, this study focuses on monologic spoken responses produced by non-native speakers within the context of a language proficiency assessment and aims to identify the discourse structure of spoken responses. The RST framework was selected due to the fact that it can effectively demonstrate the deep hierarchical discourse structure across an entire response, rather than focusing on the local coherence of adjacent units.

3 Data and Annotation

3.1 Data

This study obtained manual RST annotations on a corpus of 1440 spoken responses, where 600 of them were obtained in our previous work (Wang et al., 2017a), and the additional 840 responses were annotated more recently. All the responses were drawn from a large-scale, high-stakes standardized assessment of English for non-native speakers, the TOEFL® Internet-based Test (TOEFL® iBT), which assesses English communication skills for academic purposes (ETS, 2012). The speaking section of the TOEFL iBT assessment contains six tasks, each of which requires the test taker to provide an unscripted spoken response, 45 or 60 seconds in duration. The corpus used in this study includes 240 responses from each of six different test questions that comprise two different speaking tasks: 1) Independent questions, in which test takers provide an opinion based on personal experience (N = 480 responses) and 2) Integrated questions, in which test takers summarize or discuss material provided in a reading and/or listening passage (N = 960 responses). The spoken responses were all manually transcribed using standard punctuation and capitalization.

Responses were all provided with holistic English proficiency scores on a scale of 1 to 4 (weak to good) by expert human raters in the context of operational, high-stakes scoring for the spoken language assessment. The scoring rubrics 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). Responses were balanced for proficiency levels, i.e., 60 responses were included from each of the 4 score points from each of the 6 test questions.

In addition to the holistic proficiency scores, the transcription of each spoken response in this corpus was also provided with a global discourse coherence score by two expert annotators (not drawn from the pool of expert human raters who provided the holistic scores) in our previous study (Wang et al., 2013). The score scale for these coherence scores was from 1 to 3, and 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). A subset of 600 responses were double annotated, and the inter-annotator agreement for these coherence scores was with a quadratic weighted kappa of 0.68.

3.2 Annotation Guidelines

This study used the same annotation guidelines as in our previous work Wang et al. (2017a), which is a modified version of the tagging reference manual from the RST Discourse Treebank (Carlson and Marcu, 2001). According to these guidelines, annotators segment a transcribed spoken response into Elementary Discourse Unit (EDU) spans of text (corresponding to clauses or clause-like units), and indicate rhetorical relations between non-overlapping spans which typically consist of a nucleus (the most essential information in the rhetorical relation) and a satellite (supporting or background information).

In contrast to well-formed written text, non-native spontaneous speech frequently contains ungrammatical sentences, disfluencies, fillers, hesitations, false starts, and unfinished utterances. In some cases, these spoken responses do not constitute coherent, well-formed discourse. In order to account for these differences, we created an addendum to the RST Discourse Treebank manual introducing the following additional relations:

disfluency relations (in which the disfluent span is the satellite and the corresponding fluent span is the nucleus), **awkward relations** (corresponding to portions of the response where the speaker's discourse structure is infelicitous; awkward relations are based on pre-existing relations, such as *awkward-Reason*, if the intended relation is clear but is expressed incoherently, or *awkward-Other* if there is no clear relation between the awkward EDU and the surrounding discourse), **unfinished utterance relations** (representing EDUs at the end of a response that are incomplete because the test taker ran out of time, in which the incomplete span is the satellite and the root node of the discourse tree is the nucleus), and **discourse particle relations** (such as *you know* and *right*, which are satellites of adjacent spans).

The discourse annotation tool used in the RST Discourse Treebank[3] was also adopted for this study. Using this tool, annotators incrementally build hierarchical discourse trees, in which the leaves are the EDUs and the internal nodes correspond to contiguous spans of text. When the annotators assign the rhetorical relation for a node of the tree, they provide the relation's label (drawn from the pre-defined set of relations in the annotation guidelines) and also indicate whether the spans that comprise the relation are nuclei or satellites. Figure 1 shows an example of an annotated RST tree for a response with a proficiency score of 1. This response includes three disfluencies (EDUs 3, 6, and 9), which are satellites of the corresponding repair nuclei. In addition, the response also includes an awkward Comment-Topic relation between EDU 2 and the node combining EDUs 3-11, indicated by *awkward-Comment-Topic-2*; in this multinuclear relation, the annotator judged that the second branch of the relation was awkward, which is indicated by the *2* that was appended to the relation label.

3.3 Human Annotations

Among the 600 annotations obtained in Wang et al. (2017a), 120 responses from 6 test questions (5 responses from each score level for each question) were double annotated. The standard evaluation method of F1 scores on three levels (span, nuclearity, and relation) (Marcu, 2000b) was used to evaluate the human agreement, where the F1 scores were calculated globally by comparing the two annotators' labels from all samples, i.e., a micro-averaged F1 score. The human agreement results are 86.8%, 72.2%, and 58.2%, according to the span, nuclearity, and relation levels respectively. This level of agreement is similar to the inter-annotator agreement rates on the RST Discourse Treebank, i.e., 88.3% on span, 77.3% on nuclearity, and 64.7% on relation, respectively (Joty et al., 2015; Morey et al., 2017).

The human agreement results also indicate that two annotators tend to agree better on responses from speakers with higher speaking proficiency levels, which is demonstrated by positive correlations (Pearson correlation coefficients) between the F1 agreement scores (F1 scores from each of the double annotated samples) and the human proficiency ratings, approximately 0.2 on all three levels. Meanwhile, the correlations between F1 agreement scores and the human coherence scores are even higher, reaching 0.358 on the fully labeled relation level, which means that human raters agreed better with each other on responses receiving higher coherence scores, as expected. In addition, annotators also provided feedback that this data set posed some unique challenges compared to the data set used to create the RST Discourse Treebank. While the Wall Street Journal articles are written and edited by professionals, our data set consisted of human transcriptions of nonnative spontaneous speech, which were at times unintelligible due to the lack of proficiency and transcription inaccuracy.

4 Automatic Parsing

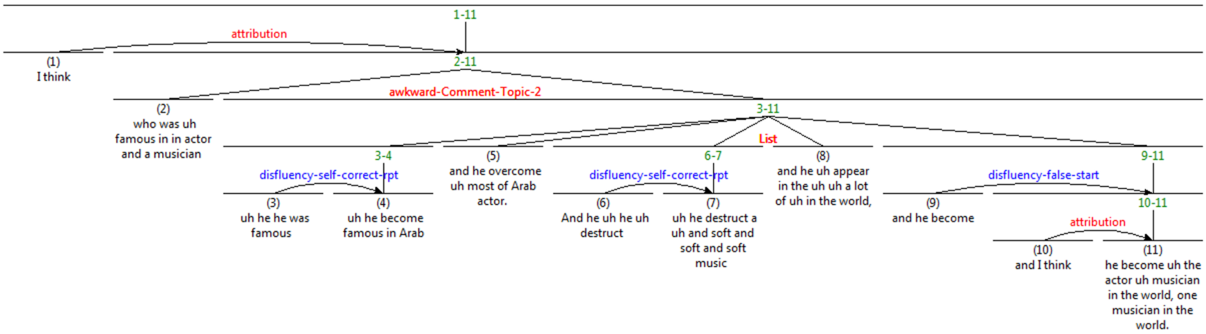
4.1 Parser Training

There has been a variety of research on document-level discourse parsing based on the RST Discourse Treebank, and multiple RST parsers are available as open source tools. In this study, since the focus of our research is not to investigate advanced techniques to improve the state-of-art in parsing, we employed a pre-existing open-source parser from Heilman and Sagae (2015)[4], which was implemented following the work of Sagae (2009) and Ji and Eisenstein (2014). It is a fast, transition-based parser and can process short documents such as news articles or essays in less

[3] Downloaded from http://www.isi.edu/licensed-sw/RSTTool/index.html

[4] Downloaded from https://github.com/EducationalTestingService/discourse-parsing

Figure 1: Example of an annotated RST tree on a response with a proficiency score of 1.


1-11
attribution
(1)
I think
2-11
(2)
who was uh famous in in actor and a musician
awkward-Comment-Topic-2
3-11
List
3-4
disfluency-self-correct-rpt
(3)
uh he he was famous
(4)
uh he become famous in Arab
(5)
and he overcome uh most of Arab actor.
6-7
disfluency-self-correct-rpt
(6)
And he uh he uh destruct
(7)
uh he destruct a uh and soft and soft and soft music
(8)
and he uh appear in the uh uh a lot of uh in the world,
9-11
disfluency-false-start
(9)
and he become
10-11
attribution
(10)
and I think
(11)
he become uh the actor uh musician in the world, one musician in the world.

than a second. Since the ultimate goal is to introduce the discourse parser into an automated speech scoring system consisting of many interdependent downstream components, reducing the amount of time required for extracting discourse features is an advantage. We first examined the performance of this selected parser by re-training and re-evaluating it on the RST Discourse Treebank with the standard data partition as in Heilman and Sagae (2015), i.e., 347 samples as the training set, 40 of them were used as the development set, and 38 samples as the test set. In this paper, all the parsers we built were evaluated with the micro F1 score. When using the gold standard syntax trees and EDU segmentations, the F1 scores on three levels of span, nuclearity, and relation can reach 84.1%, 69.6%, and 56.5% respectively, which are close to state-of-the-art accuracy, as reported in Morey et al. (2017).

In this work, the annotated data obtained as described in Section 3 was used for parser building and evaluation. Among the 1440 annotated responses, the data was split into a training set with 1271 single-annotated responses, a development set with 49 single-annotated responses, and a test set with 120 double-annotated responses. Afterwards, the 49 responses in the development set were further double annotated, which allowed us to tune the parser on annotations from both human experts. In contrast to the Wall Street Journal articles in the RST Discourse Treebank (RST_DT), the responses in the corpus of non-native spontaneous speech (RST_SS) are much shorter. Table 1 compares the RST_DT and the RST_SS data sets in terms of the means and standard deviations of the number of EDUs and word tokens. It shows that the RST_SS corpus has more samples (1271 vs. 347 in the training set), but the total numbers of EDUs and words in RST_SS are similar to the RST_DT corpus (18,171 vs. 19,443 EDUs and 156,254 vs. 184,352 words in the training set).

Table 1: Average numbers of EDUs and word tokens (and their standard deviations) appearing in the RST Discourse Treebank (RST_DT) and the annotated corpus of non-native spontaneous speech (RST_SS).

	# Samples	# EDUs Mean (std)	# Words Mean (std)
		RST_DT	
Train	347	56.0 (51.5)	531.3 (464.0)
Test	38	61.7 (63.4)	570.2 (549.0)
		RST_SS	
Train	1271	14.3 (4.7)	122.9 (36.0)
Dev	49	13.0 (4.7)	112.7 (35.7)
Test	120	14.8 (4.4)	127.4 (33.5)

In addition, Table 2 shows the most common relations that appear in the training sets of RST_SS and lists their percentages, taken according to their frequency. The percentage of these relations appearing in the RST_DT are also included for comparison. The top five most common relations overlap, but the other five relations that frequently appear in RST_SS are relatively rare in RST_DT, especially the disfluency-self-correct-rpt and disfluency-false-start relations, which is unique to the spoken responses and will not appear in the written texts. In addition, the proportions of each relation appearing in RST_SS and RST_DT are quite different.

4.2 Parser Evaluation

For comparison, we trained three different parsers on both RST_DT and RST_SS: (a) **RST_SS:** using the training set from the corpus of non-native spontaneous speech, where 49 double-annotated responses were used as the development set; (b) **RST_DT:** using the training set from the RST Dis-

Table 2: Top 10 relations appearing in the training set of the annotated corpus of spontaneous speech (RST_SS). The percentages of each relation appearing in both RST_SS and the RST Discourse Treebank (RST_DT) are listed for comparison.

	RST_SS	RST_DT
list	18.2%	13.3%
elaboration -object-attribute-e	7.8%	10.4%
same-unit	7.1%	11.1%
attribution	5.5%	11.3%
elaboration-additional	4.7%	13.2%
reason	3.3%	0.8%
disfluency -self-correct-rpt	2.8%	–
evidence	2.4%	0.7%
disfluency-false-start	2.3%	–
conclusion	2.2%	0.02%

Table 3: Discourse parsing performance in terms of F1 scores (%) on three levels of Span, Nuclearity, and Relation. Human agreements are also listed for comparison. Within each cell, two micro F1 scores according to the gold standards from each of two human annotators are both reported.

	Span	Nucleus	Relation
RST_SS	75.5 76.2	56.4 58.6	41.2 43.1
RST_DT	73.0 73.8	53.0 54.8	35.0 36.5
RST_SS + RST_DT	76.1 77.0	57.6 59.7	42.6 44.4
Human	86.8	72.2	58.2

course Treebank, where 40 samples from the training set were separated as the development set; and (c) **RST_SS + RST_DT:** using the training sets from both RST_SS and RST_DT, where the development set is the same one used in (a). These three parsers were evaluated on the same test set from RST_SS, where the gold standard EDU segmentations were used. As shown in Table 3, the parser trained on RST_SS outperformed the one trained on RST_DT, especially on the relation level, i.e., 41.2%/43.1% vs. 35.0%/36.5%. By combining both data corpora, the F1 scores can further be improved.

Furthermore, besides using gold standard EDU segmentations, we also applied the automatic EDU segmenter within the parser to generate segmentations and then build the RST trees upon them. The evaluation results showed that F1 scores of all three parsers were greatly reduced through this transition. For example, they were decreased to 53.0%/53.6% on span, 40.4%/41.9% on nuclearity, and 29.3%/31.1% on relation for parser (a) trained on **RST_SS**. Therefore, the improvement of EDU segmentations is also a research focus of our future work. In the following section on discourse modeling for spontaneous speech, parser (a), which was trained on RST_SS and using automatic EDU segmentations, was employed for discourse modeling.

5 Discouse Features

The ultimate goal of this line of research is to investigate which features are effective for automatically assessing discourse structure in non-native spontaneous speech. We previously used RST trees for this purpose and proposed several features based on the distribution of relations and the structure of trees (Wang et al., 2017a), including the number of EDUs (n_edu), the number of relations (n_rel), the number of awkward relations (n_awk_rel), the number of rhetorical relations, i.e., relations that were neither classified as awkward nor as disfluencies (n_rhe_rel), the number of different types of rhetorical relations (n_rhe_rel_types), the percentage of rhetorical relations (perc_rhe_rel) out of all relations, the depth of the RST trees (tree_depth), and the ratio between n_edu and tree_depth (ratio_nedu_depth).

In this work, we first examined these eight features on the 1271 single-annotated responses, i.e., the RST_SS training set used to build the automatic parser as described in Section 4.1. Features were extracted from the manually annotated trees, and then the Pearson correlation coefficients of these features with both the holistic proficiency scores as well as the discourse coherence scores are reported in Table 4, which demonstrates the effectiveness of these features. The n_rhe_rel feature achieves the highest correlation with the holistic proficiency scores at 0.719, and the normalized feature perc_rhe_rel achieves the highest correlation with the coherence scores at 0.609. There are six features that receive higher correlations with the proficiency scores, whereas the other two features (n_awk_rel and perc_rhe_rel) receive higher absolute correlations with the coherence scores. This is consistent with our previous obser-

Table 4: Pearson correlation coefficients (r) of discourse features with both the holistic proficiency scores as well as the discourse coherence scores.

Features	Proficiency	Coherence
n_edu	0.612	0.366
n_rel	0.624	0.391
n_awk_rel	-0.425	-0.533
n_rhe_rel	0.719	0.536
n_rhe_relTypes	0.675	0.547
perc_rhe_rel	0.586	0.609
tree_depth	0.402	0.249
ratio_nedu_depth	0.536	0.308

vations, where RST-based discourse features generally have higher correlations with the holistic speaking proficiency scores than with the more specific discourse coherence scores (Wang et al., 2017a). One potential explanation could be the difference in score range: 1-3 for the discourse scores vs. 1-4 for the more fine-grained holistic proficiency scores.

6 Automated Scoring

Besides examining the discourse features based on the manually annotated trees as above, this study also conducted an experiment to examine them on automatically generated trees to measure the discourse structure of non-native spontaneous speech, and then further employ them in an automated spoken English assessment system, SpeechRater^{TM} (Zechner et al., 2007, 2009).

6.1 Experimental Setup

The task is to build effective classification models, referred to as "scoring models", which can automatically predict the holistic proficiency scores by measuring the different aspects of non-native speaking proficiency, including pronunciation, prosody, fluency, vocabulary, grammar, and, in particular, discourse in spontaneous speech. In order to obtain credible evaluation results, this study collected a large data set from the operational TOEFL iBT assessment to conduct this experiment, which includes 17,194 speakers who responded to all the six test questions as described in Section 3.1. The holistic proficiency scores were provided during the operational test, but more specific discourse coherence scores were not available for this large data set. The whole data set was partitioned into two sets: one containing 12,194 speakers (73,164 responses) as the training set to build the scoring models, and the other one containing 5,000 speakers (30,000 responses) to test the model performance.

The baseline scoring model was built with approximately 130 automatic features extracted from the SpeechRater system, which can measure the pronunciation, prosody, fluency, rhythm, vocabulary, and grammar of spontaneous speech. All SpeechRater features were extracted either directly from the speech signal or from the output of a Kaldi-based automatic speech recognizer (Qian et al., 2016) with a word error rate of 20.9% on an independent evaluation set with non-native spontaneous speech from the TOEFL iBT speaking test.

Based on the automatic speech recognition output (without punctuations and capitalization) generated by SpeechRater, the automatic parsers developed in section 4.1 were applied to extract RST trees. Afterwards, the RST-based features were automatically obtained. Therefore, in this process, no manual transcriptions or manual annotations were involved. Furthermore, the RST-based discourse features can be combined with the baseline features to extend the ability of SpeechRater to assess the discourse structure of non-native spontaneous speech.

6.2 Results and Discussion

The automatically generated discourse features were first examined on the scoring model training partition, where Pearson correlation coefficients between automatic features and proficiency scores were calculated. There were two sets of features extracted and examined, based on two different parsers: one was trained with RST_SS and the other one was trained with both RST_SS and RST_DT as shown in Section 4.1.

Table 5 presents the Pearson correlation coefficients of these two sets of features with the proficiency scores. For the five features n_edu, n_rel, n_awk_rel, n_rhe_rel, and tree_depth, the difference is limited, i.e., smaller than 0.004. In contrast, the other three features, n_rhe_rel_types, perc_rhe_rel, and ratio_nedu_depth, achieve better correlations with features based on the RST_SS parser. This indicates the effectiveness of our annotations in capturing discourse in spoken language. Therefore, in the following experiments on scoring models, the features were obtained using the parser

Table 5: Pearson correlation coefficients (r) of discourse features with both the holistic proficiency scores. RST_SS indicates using the parser trained with the annotations on speech data during the feature generation, and RST_SS + RST_DT indicates using the parser trained with both the annotations on speech data and the RST Discourse Treebank.

Features	RST_SS	RST_SS + RST_DT
n_edu	0.424	0.427
n_rel	0.401	0.405
n_awk_rel	-0.096	-0.096
n_rhe_rel	0.418	0.42
n_rhe_rel_types	0.314	0.308
perc_rhe_rel	0.225	0.211
tree_depth	0.329	0.328
ratio_nedu_depth	0.316	0.289

trained with the RST_SS data. Even though all these features were extracted with the automatic speech recognition output and with the automatic parser, they can still achieve moderate correlations with the proficiency scores in a range of 0.2-0.5, except for the feature based on count of awkward relations. The absolute correlation of n_awk_rel feature is less than 0.1, which was caused by the failure of the automatic parser to identify awkward relations.

Furthermore, scoring models were built with SpeechRater features and RST-based discourse features to automatically predict the holistic proficiency scores using the machine learning tool of scikit-learn[5] (Pedregosa et al., 2011). For this experiment, we used the Random Forest classification method to build the scoring models.

Table 6 shows that the baseline system with 131 SpeechRater features can reach an accuracy of 65.3%. By introducing the eight RST-based features, there is a very slight improvement on the accuracy to 65.4% and no improvement in terms of the Pearson correlation coefficient between the automatic and human scores. A scoring system only using eight RST-based features can achieve an accuracy of 55.9%. These results indicate that the proposed features can be used to measure the discourse coherence of non-native spontaneous spoken responses. Due to the fact that these 131 SpeechRater features are powerful in measuring various aspects of non-native spontaneous speech, the improvement by introducing discourse features to predict the holistic proficiency scores is limited. But on the other hand, by employing the proposed discourse-level features, the validity of an automatic system for English language proficiency assessment can be improved, because it enables the measurement of an important aspect of speech that appears in the human scoring rubrics.

Table 6: Performance of the automatic scoring models to predict holistic proficiency scores. The baseline system was built with 131 SpeechRater features, and the automatically generated 8 RST-based features were appended to measure the discourse structure.

	Accuracy (%)	r
RST	55.9	0.371
SpeechRater	65.3	0.587
SpeechRater + RST	65.4	0.587

7 Conclusion

The goal of this research effort is to model discourse structure in non-native spontaneous speech to facilitate the automatic assessment of English language proficiency. In order to achieve this goal, we first obtained an annotated corpus of 1440 spoken responses produced by non-native speakers of English in the context of an English speaking proficiency assessment using Rhetorical Structure Theory and then trained automatic discourse parsers based on the human annotations. Subsequently, discourse features were extracted from the speech signal using automatic speech recognition output and automatically parsed RST trees; these features mostly achieved moderate correlations with human holistic proficiency scores ranging between 0.2 and 0.5. Finally, a scoring model trained using the eight proposed discourse features can predict the proficiency scores with an accuracy of 55.9%, and by introducing them into an automatic speech scoring system, the validity of the system can be improved.

References

Chloé Braud, Maximin Coavoux, and Anders Søgaard. 2017. Cross-lingual RST discourse parsing. In *Proceedings of the EACL conference*.

Jill Burstein, Daniel Marcu, and Kevin Knight. 2003. Finding the WRITE stuff: Automatic identification

[5] SKLL, a python tool making the running of scikit-learn experiments simpler, was used. Downloaded from https://github.com/EducationalTestingService/skll.

of discourse structure in student essays. *IEEE Intelligent Systems*, 18(1):32–39.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. Technical Report ISI-TR-545, ISI Technical Report.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurows. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue*, pages 1–10.

Lei Chen, Klaus Zechner, and Xiaoming Xi. 2009. Improved pronunciation features for construct-driven assessment of non-native spontaneous speech. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 442–449.

Miao Chen and Klaus Zechner. 2011. 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: Human Language Technologies*, pages 722–731, Portland, Oregon, USA.

Jian Cheng. 2011. Automatic assessment of prosody in high-stakes English tests. In *Proceedings of Interspeech*, pages 27–31.

Catia Cucchiarini, Helmer Strik, and Lou Boves. 2002. Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America*, 111(6):2862–2873.

Isin Demirsahin and Deniz Zeyrek. 2014. Annotating discourse connectives in spoken Turkish. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 105–109.

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

Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–521.

Vanessa Wei Feng, Ziheng Lin, and Graeme Hirst. 2014. The impact of deep hierarchical discourse structures in the evaluation of text coherence. In *Proceedings of COLING 2014, The 25th International Conference on Computational Linguistics: Technical Papers*, pages 940–949.

Michael Heilman and Kenji Sagae. 2015. Fast rhetorical structure theory discourse parsing. *CoRR*, abs/1505.02425.

Hugo Hernault, Helmut Prendinger, David A. duVerle, and Mitsuru Ishizuka. 2010. HILDA: A discourse parser using support vector machine classification. *Dialogue & Discourse*, 1(3):1–33.

Derrick Higgins, Xiaoming Xi, Klaus Zechner, and David Williamson. 2011. A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech and Language*, 25:282–306.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13–24.

Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Yashar Mehdad. 2013. Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 486–496, Sofia, Bulgaria.

Shafiq Joty, Giuseppe Carenini, and Raymond T Ng. 2015. Codra: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435.

Qi Li, Tianshi Li, and Baobao Chang. 2016. Discourse parsing with attention-based hierarchical neural networks. In *Proceedings of the EMNLP conference*, pages 362–371.

Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. 2014. Text-level discourse dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25–35.

Yang Liu and Mirella Lapata. 2017. Learning contextually informed representations for linear-time discourse parsing. In *Proceedings of the EMNLP conference*, pages 1289–1298.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse (Text)*, 8(3):243–281.

Daniel Marcu. 2000a. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26:395–448.

Daniel Marcu. 2000b. *The Theory and Practice of Discourse Parsing and Summarization.* MIT Press.

Mathieu Morey, Philippe Muller, and Nicholas Asher. 2017. How much progress have we made on RST discourse parsing? a replication study of recent results on the RST-DT. In *Proceedings of the EMNLP conference*, pages 1319–1324.

A Paulino, Gerardo Sierra, Laura Hernandez-Dominguez, Iria da Cunha, and Gemma Bel-Enguix. 2018. Rhetorical relations in the speech of alzheimer's patients and healthy elderly subjects: An approach from the rst. *Computacion y Sistemas*, 22:895–905.

Anayeli Paulino and Gerardo Sierra. 2017. Applying the rhetorical structure theory in alzheimer patients' speech. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, page 34–38.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, and Livio Robaldo. 2008. The Penn Discourse TreeBank 2.0. In *The 6th International Conference on Language Resources and Evaluation (LREC)*, pages 2961–2968.

Yao Qian, Xinhao Wang, Keelan Evanini, and David Suendermann-Oeft. 2016. Self-adaptive DNN for improving spoken language proficiency assessment. In *Proceedings of Interspeech 2016*, pages 3122–3126.

Ines Rehbein, Merel Scholman, and Vera Demberg. 2016. Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks. In *The Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1039–1046.

Kenji Sagae. 2009. Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing. In *Proceedings of the 11th International Conference on Parsing Technologies*, pages 81–84.

Ted J. M. Sanders, Wilbert P. M. Spooren, and Leo G. M. Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse Processes*, 15(1):1–35.

Svetlana Stoyanchev and Srinivas Bangalore. 2015. Discourse in customer care dialogues. Poster presented at the Workshop of Identification and Annotation of Discourse Relations in Spoken Language. Saarbrücken, Germany.

Maite Taboada and William C. Mann. 2006a. Applications of Rhetorical Structure Theory. *Discourse Studies*, 8(4):567–588.

Maite Taboada and William C. Mann. 2006b. Rhetorical Structure Theory: Looking back and moving ahead. *Discourse Studies*, 8(3):423–459.

Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad, and Aravind Joshi. 2010. Annotation of discourse relations for conversational spoken dialogs. In *The Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2084–2090.

Xinhao Wang, James Bruno, Hillary Molloy, Keelan Evanini, and Klaus Zechner. 2017a. Discourse annotation of non-native spontaneous spoken responses using the rhetorical structure theory framework. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 263–268.

Xinhao Wang, Keelan Evanini, and Klaus Zechner. 2013. Coherence modeling for the automated assessment of spontaneous spoken responses. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings*, pages 814–819, Atlanta, Georgia.

Xinhao Wang, Keelan Evanini, Klaus Zechner, and Matthew Mulholland. 2017b. Modeling discourse coherence for the automated scoring of spontaneous spoken responses. In *Proceedings of the Seventh ISCA workshop on Speech and Language Technology in Education 2017, SLaTE, August 25–26, Djurö, Stockholm, Sweden.*

Yizhong Wang, Sujian Li, and Houfeng Wang. 2017c. A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 184–188.

Shasha Xie, Keelan Evanini, and Klaus Zechner. 2012. Exploring content features for automated speech scoring. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 103–111.

Su-Youn Yoon and Suma Bhat. 2012. Assessment of ESL learners' syntactic competence based on similarity measures. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 600–608.

Su-Youn Yoon, Suma Bhat, and Klaus Zechner. 2012. Vocabulary profile as a measure of vocabulary sophistication. In *Proceedings of the 7th Workshop on the Innovative Use of NLP for Building Educational Applications*, pages 180–189.

Klaus Zechner, Derrick Higgins, and Xiaoming Xi. 2007. Speechrater^{SM}: A construct-driven approach to scoring spontaneous non-native speech. In *Proceedings of the International Speech Communication Association Special Interest Group on Speech and Language Technology in Education*, pages 128–131.

Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M. Williamson. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, 51(10):883–895.