IJMER
International Journal of Multidisciplinary and Educational Research

International Journal of Multidisciplinary and Educational Research

Argumentation and Discourse Analysis in Future Intelligent Systems of Essay Grading

Naima DEBBAR[1]
Badji Mokhtar-Annaba University, Annaba, Algeria.

IJMER

Volume. 8, Issue. 4

December, 2025

Abstract
Intelligent systems of essay grading constitute important tools for educational technologies. They can significantly replace the manual scoring efforts and provide instructional feedback as well. These systems typically include two main parts: a feature extractor and an automatic grading model. The latter is generally based on computational and artificially intelligent methods. In this work, we focus on the feature extraction part. More precisely, we focus on argumentation and discourse-related features, which constitute high-level features. We discuss some state-of-the-art systems and analyze how argumentation and discourse analysis are used for extracting features and providing feedback.

Keywords: Automatic essay scoring, Artificial intelligence, Natural Language Processing, E-Learning, Assessment.

1. Introduction

Automatic essay scoring (AES), also called Automated Essay Grading (AEG) or Automated Essay Evaluation (AEE), concerns grading essays using computers with minimum, or even without, human intervention. These systems constitute a valuable asset for future educational technologies. AES are becoming more and more widespread in modern educational technologies, especially for large open classes like MOOCs (Massive Open Online Courses). The main advantage of AES is preserving teachers' effort and time. Indeed, manual scoring of essays is a hard and laborious process, especially for massive classes and language courses. AES also overcomes some factors that influence manual evaluation, such as the evaluator's mental state, the biases, and the disparity between evaluators (Alqahtani and Alsaif, 2020). Furthermore, most of the current AES provide not only a holistic score but also instructional feedback to the user, which considerably helps to improve the writing quality. This is as if each student has her or his own personal tutor, whom she or he can permanently consult.

AES for English has been widely introduced, and many commercial applications are available. Project Essay Grade (PEG) (Page, 1966) was the first AES system for English. Other early works include IntelliMetric, developed by Vantage Learning in 1998 (Foltz et al. 1999); E-Rater, developed by Educational Testing Services (ETS) in 1998 (Burstein, 2003a); and Intelligent Essay Assessor (IEA) (Landauer, 2003). Recent works are based on more sophisticated AI methods, which are directly applied to texts without using hand-crafted features (Dasgupta et al. 2018; Nadeem et al. 2019; Cropley and Marrone 2022). This aims at avoiding feature extraction steps, which are time-consuming (Uto et al., 2021).

AES systems have been based on a large variety of methods, ranging from simple similarity measures to sophisticated AI methods like deep neural networks. In terms of features, AES systems have also used a large variety of features, ranging from word and sentence counts to discourse, argumentation, and coherence analysis (Ke and Ng, 2019; Hussein et al., 2019; Wang et al., 2022). In this paper, we focus on high-level features used in AES. More precisely, we study the role of argumentation and discourse analysis in modern AES systems and discuss how they can be integrated into future systems.

The rest of this paper is organized as follows: Section 2 gives some theoretical background about AES. It presents the architecture of these systems and the features used. This section also presents RST because it is extremely related to argumentation and discourse analysis in AES. Section

3 describes some state-of-the-art AES systems and discusses how they integrate argumentation and discourse concepts. Finally, discussion and conclusion are given in Section 6.

2. Theoretical Background

2.1. Architecture of AES Systems

These systems typically include two main parts: a feature extractor and a prediction model. To score an essay, it is first presented to the feature extractor, and then the obtained features are presented to the prediction model, which provides a score for the essay (Figure 1).

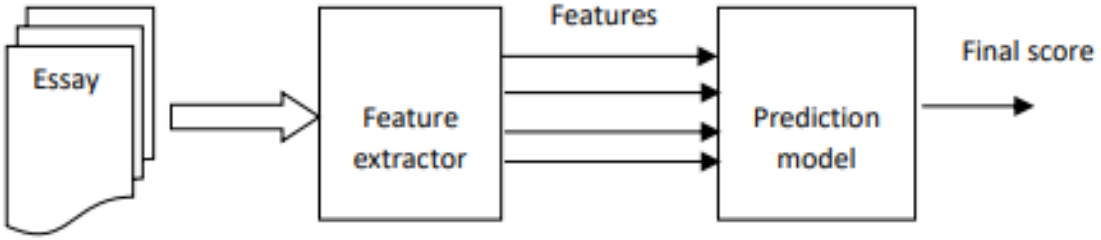

Figure 1. A typical representation of an AES system

2.2. Prediction Models

AES can be classified into two categories: handcrafted feature-based and neural-based methods. In the first category, the two steps, i.e., feature extraction and prediction, are performed independently, while in the second, a neural network performs both tasks. Since the main focus of this work is analyzing features, we give only a simplified description of some concepts of predicting methods:

Latent Semantic Analysis (LSA) is a statistical method that represents meaning in a text. The application of LSA to a corpus of texts consists of representing the texts with a term-by-document matrix, in which the columns represent documents and the rows represent terms. A term can be a word, phrase, or other unit. A document can be a sentence, a paragraph, or something else.

Artificial neural networks (ANN), or simply neural networks (NN), are computing systems inspired by the biological neural networks in animal brains. An ANN consists of an ensemble of connected units called artificial neurons. Neural networks are machine learning-based models that learn via training examples labeled with the desired output. In AES, an ANN can be trained using a corpus of scored essays. After training, it can predict scores for new essays.

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) is a new powerful language model introduced by researchers at Google. It consists of transformer encoder layers, and it converts each word into an integer code.

2.3. Features

AES systems are based on a large variety of features, ranging from simple word and sentence counts to high-level features like discourse, argumentation, and coherence analysis (Ke and Ng, 2019; Hussein, 2019; Wang, 2022). (Table 1) illustrates an example of categorization of the AES features (Ramesh and Sanampudi, 2022).

Table 1. Types of features used in AES (Ramesh and Sanampudi, 2022)

Statistical features	Style-based features	Content-based features
Essay length with respect to the number of words, Essay length with respect to sentence Average sentence length, Average word length, N-gram	Sentence structure, Part of speech (POS), Punctuation, Grammatical, Vocabulary, Logical operators	Cohesion between sentences in a document, Overlapping (prompt), Relevance of information, The semantic role of words, Correctness, Consistency, Sentence expressing key concepts

In Table 1, N-Gram is a series of N adjacent letters, syllables, or words. They can be extracted from a text or speech corpus.
Argumentation and discourse features can be classified in the third group, i.e., content-based features.

3. Rhetorical Structure Theory (RST)

Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) is one of the most important methods for discourse analysis. The central principle of RST is rhetorical relations, which connect two adjacent and non-overlapping text spans, called discourse units. These units are: nuclei (N), the most important parts, and satellites (S), the less important. The role of satellites is just to help understand the nuclei. The text is still understood without satellites.

Discourse relations in RST are "nucleus-satellite" relations or "multinuclear" relations, where both spans are important. Schemas specify how spans of text can appear, and they define the possible RST text structures. In RST, there are five kinds of schemas, represented by the five examples illustrated in Figure 2. The curves represent relations, and the straight lines represent nuclear spans. According to Mann and Thompson (1988), the CONTRAST schema always has exactly two nuclei, while both sequence and junction have indefinitely many nuclei, which they are without satellites. Rhetorical relations reflect semantic, intentional, and textual relations that hold between text spans.

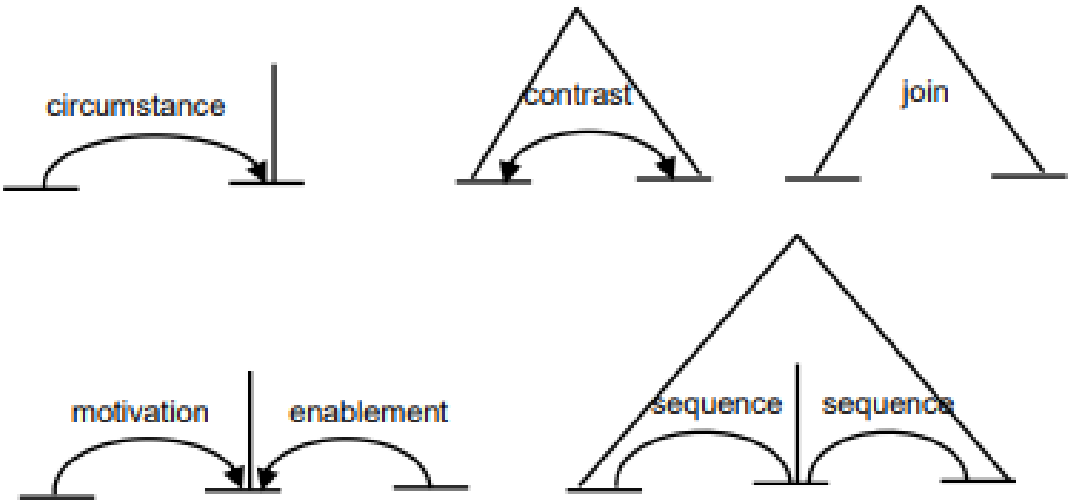

Figure 2. Examples of the five schemas types in RST. The horizontal lines represent text spans and the vertical and diagonal lines represent of spans (Mann and Thompson 1988).

(Figure 3) illustrates the relations defined in Mann and Thompson (1988). They are grouped according to a specific kind of resemblance. Each group includes relations that share some characteristics but differ in one or two particular attributes.

Circumstance
Solutionhood
Elaboration
Background
Enablement and Motivation
- Enablement
- Motivation

Evidence and Justify
- Evidence
- Justify

Relations of Cause
- Volitional Cause
- Non-Volitional Cause
- Volitional Result
- Non-Volitional Result
- Purpose

Antithesis and Concession
- Antithesis
- Concession

Condition and Otherwise
- Condition
- Otherwise

Interpretation and Evaluation
- Interpretation
- Evaluation

Restatement and Summary
- Restatement
- Summary

Other Relations
- Sequence
- Contrast

Figure 3. The rhetorical relations defined in (Mann and Thompson, 1988)

In RST, a discourse structure can be represented by a hierarchical tree in which nodes are linked with rhetorical relations. Nodes are either nuclei or satellites. (Figure 4) illustrates an example of a text's RST tree taken from Mann and Thompson (1988). The nuclei are represented by straight lines, and the satellites by arcs.

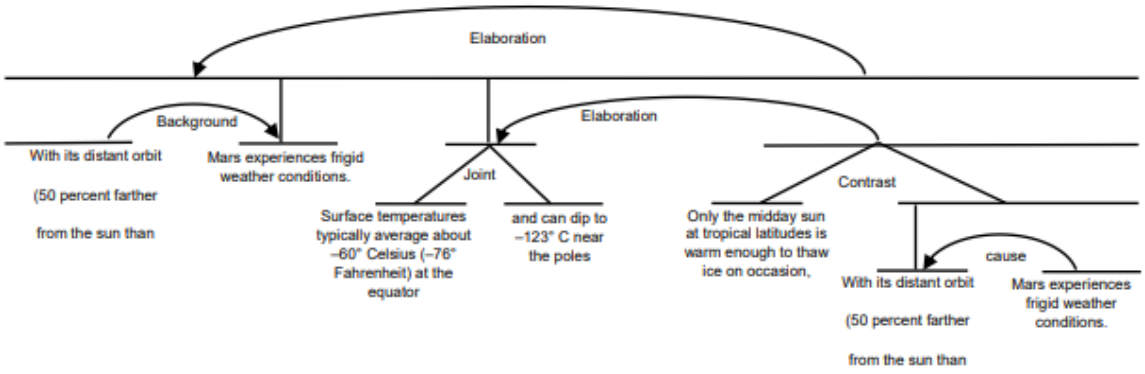

Figure 4. Example of rhetorical structure tree (Mann and Thompson, 1988)

4. Argumentation and Discourse Analysis in AES Systems

4.1. Discourse Structures and Modes in AES

Discourse structure is an important factor for evaluating essay cohesion. Accordingly, defining discourse structures have been introduced in several AES systems. Most approaches to defining discourse are based on RST. The following paragraphs describe some works that propose AES systems based on discourse analysis and RST:

In Burgstein (2003b), the authors proposed a system that automatically identifies discourse structure in essays. They assumed that essays can be divided into sequences of discourse parts and that each part is related to a global communicative goal. They encoded the communicative goals using labels that are commonly used in teaching writing, like thesis statements, main ideas, and conclusions. The proposed system is based on constructing a rhetorical structure tree using an automatic discourse parser. It gives RST-based rhetorical relations to the essay sentences. Then, two features are associated with each sentence. The first feature represents if the parent node is a nucleus or satellite, while the second feature represents its rhetorical relation. In addition to rhetorical features, the authors used other features, such as syntactic structure and lexical elements. For example, they consider some lexical items like "opinion" and "feel" as terms linked to thesis statements and the term "in

conclusion" linked to conclusions. These terms help the system define discourse structure. To train and evaluate the proposed system, the authors introduced a dataset that contains annotated essays. The adopted discourse elements are: title, introductory material, main idea, supporting idea, conclusion, and irrelevant segments. The latter indicates that it is not fitting for the other discourse categories.

Song et al. (2017) proposed a system that automatically identifies discourse mode in essays. These modes are: narration, exposition, description, argument, and emotion expression. They introduced a corpus of narrative Chinese student essays that were manually annotated with discourse modes at the sentence level. In addition, they introduced two discourse-mode-based features for automatic essay scoring. These two features are: (i) the proportion of each discourse mode, which is calculated as the ratio of the number of corresponding sentences to the total number of sentences; (ii) the bag of N-grams of discourse modes, which is based on the mode's sequences of sentences in the essay.

In Azmi et al. (2019), the authors checked the cohesion of an essay by applying RST. Indeed, the structures in RST are hierarchic, and they can be represented using an RST-tree (rhetorical structure tree). The authors examine the coherence relations in a given essay by constructing an RST tree. Accordingly, if the entire essay can be transformed into an RST tree, then it can be considered coherent. In addition, they determine the quality of the cohesion by the number of levels in the generated RST tree. (Feng et al., 2014) studied the effect of deep discourse structures on evaluating texts. They compared a model with a full hierarchical discourse structure based on RST against two models based on shallow discourse relations. They concluded that deep discourse structures provide a better evaluation of text coherence.

Discourse structure has also been evaluated in spontaneous speech. For example, Wang et al. (2017) used an RST-annotated corpus for evaluating discourse structure in non-native speech. This corpus consists of 600 speeches. They examined eight extracted features, which are: the number of EDUs (Elementary Discourse Unit); the number of relations; the number of awkward relations; the number of rhetorical relations; the number of different types of rhetorical relations; the percentage of rhetorical relations out of all relations; the depth of the RST trees; and the ratio between the number of EDUs and the tree depth. They concluded that the RST annotation provides similar inter-annotator agreement rates. They had high correlations with holistic speaking proficiency and discourse coherence scores. In addition, they concluded that the percentage of rhetorical relations is the most influential feature. The same authors (Wang et al., 2019) introduced a larger corpus, which contains 1440 non-native speeches annotated using RST. They proposed an automatic parser trained on this corpus. Then, some features extracted from the parsed RST trees are used for predicting holistic proficiency scores.

4.2. Argumentation in AES

Argumentative discourse structures constitute a hard task because of two properties "(Stab and Gurevych, 2014): First, argumentative relations are generally implicit. Second, in contrast to RST, argumentative relations also hold between non-adjacent sentences or clauses. Recently, argument evaluation has attracted a lot of attention in the AES community. For example, Stab and Gurevych (2014) proposed an approach for finding argumentative discourse structures even with missing discourse markers. For this purpose, they introduced four feature sets: (i) Structural features, such as the location and punctuation of the argument component and its covering sentence; (ii) Lexical features, such as verbs, adverbs, and modals; (iii) Syntactic features extracted from parse trees, which are the number of sub-clauses included in the covering sentence and the depth of the parse tree; (iv) Contextual features extracted from the sentence preceding and following the covering sentence, such as the number of punctuations, the number of tokens, the number of sub-clauses, and the presence or not of modal verbs. Persing and Ng (2015) proposed to consider argument strength as a distinct dimension for scoring essays, and they introduced a corpus and a predicting model for this task. The introduced corpus contains 1000 essays annotated with argument strength; each sentence of the essays is labeled with an argument label. They considered the five following labels: Opposes, Supports, Claim, Hypothesis, or None. They used seven sentence-labeling rules. For example, the first rule is: "Sentences that begin with a comparison discourse connective or contain any string prefixes from conflict or oppose are tagged Opposes." Subsequently, they converted the scoring into features

to train the prediction model. The resulting three features define: (i) if the essay comprises at least one sentence tagged Hypothesis; (ii) if the essay comprises at least one sentence tagged Opposes; (iii) the sum of sentences tagged Claim or Support divided by the total number of paragraphs. According to the authors, these features are meaningful because, for example, an essay with lots of supporting sentences provides stronger arguments. The authors added other features like prompt agreement, which specifies the prompt statement as: agree strongly, agree somewhat, neutral, disagree somewhat, disagree strongly, not address, or no opinion.

Evaluating argumentation has also been used in the business domain. For example, Williams et al. (2021) proposed a conversational tool called "ArgueTutor," which aims at helping students write more convincing texts with adaptive argumentation feedback. They used a corpus comprising 1000 business model essays. The texts are annotated for their argumentative components, which are: claim, premise, and relations. The authors trained a BERT-based model to identify and classify argument components used for evaluating writing skills, thus providing adaptive recommendations to assist students in improving their argumentation. (Wambsganss and Niklaus, 2022) proposed a scheme for annotating argumentation, and they introduced an annotated corpus of persuasive student-written business model pitches. This corpus consists of 200 German models with 3,207 sentences annotated for argument components, their relations, and six persuasiveness scores on different levels. Then, they trained a model on this corpus and integrated it into an argumentation writing tool to support students with specific argumentation feedback and recommendations.

5. Discussion and Conclusion

AES systems aim not only to score essays but also to provide valuable feedback and recommendations. The given feedback includes several rubrics like spelling, grammar, or style. This constitutes an important tool in large-scale classes like MOOCs. Recently, the researchers have introduced more sophisticated feedback, like argumentation. This allows the users to enhance their argumentation skills. AES systems will certainly perform three important tasks: assessment, adaptive training, and personal tutoring. Indeed, tutoring tools have begun to emerge, like “ArgueTutor (Wambsganss et al., 2021), which help students enhance their argumentative capacities in the business domain. In the future, several other domains will certainly benefit from these tools. For example, law students can benefit from specialized tutoring tools that help them write pleadings and evaluate their convincing abilities.

In terms of data, the availability of datasets for a specific problem allows this problem to be widely processed, especially with learning machine-based methods. The public databases permit comparing and discussing the performance of different methods. For example, the RST Discourse Treebank (Carlson et al., 2001) has facilitated RST-based discourse analysis, and consequently, a lot of parsers are now available (Wang et al., 2019). Currently, as mentioned in Section 3, many corpuses have been introduced, including: discourse elements, discourse modes, discourse structure, argumentation discourse structure, argumentation strength, and argumentative components. Most of these corpora constitute initial work for specific tasks. In the future, these corpora will certainly be increased and enhanced. Another issue concerning corpora is the language. Indeed, the argumentation structure isn’t generally the same in different languages. In addition, students with different original languages who are studying a specific language use different argumentation structures. Some corporations dealing with this issue have already been introduced. For example, Putra et al. (2021) introduced a corpus containing essays written by English learners from many Asian countries.

The devolvement in natural language processing, big data, artificial intelligence, and the other research domains involved in developing AES systems is certainly not sufficient to gain teachers’ confidence. Contrary to popular belief, involving teachers in the process of elaborating AES systems reinforces their confidence and increases the dialogue between developers and teachers. Indeed, teachers and experts have already been involved in designing some AES systems by asking them about the main features for grading essays. The development of future AES needs more involvement from teachers, especially in designing feedback and recommendations. On the other hand, machines can’t really understand the language or the art of writing. Therefore, integrating high-level features like discourse and argumentative analysis enhances, to some extent, their judgment of the essays’ beauty. This will never be achieved without the aid of teachers and education experts.

References

Azmi, A. M., Al-Jouie, M. F., & Hussain, M. (2019). AAEE–Automated evaluation of students' essays in Arabic language. Information Processing and Management, 56(5), 1736-1752.

Burstein, J. (2003a). The E-rater® scoring engine: Automated essay scoring with natural language processing.

Burstein, J., Marcu, D., & Knight, K. (2003b). Finding the WRITE stuff: Automatic identification of discourse structure in student essays. IEEE Intelligent Systems, 18(1), 32-39.

Cropley, D. H., & Marrone, R. L. (2022). Automated scoring of figural creativity using a convolutional neural network. Psychology of Aesthetics, Creativity, and the Arts.

Dasgupta, T., Naskar, A., Dey, L., & Saha, R. (2018, July). Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications (pp. 93-102).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

Feng, V. W., Lin, Z., & Hirst, G. (2014, August). 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 (pp. 940-949).

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1(2), 939-944.

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. Peer J. Computer Science, 5, e208.

Ke, Z., & Ng, V. (2019, August). Automated Essay Scoring: A Survey of the State of the Art. In IJCAI (Vol. 19, pp. 6300-6308).

Landauer, T. K. (2003). Automatic essay assessment. Assessment in education: Principles, policy and practice, 10(3), 295-308.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. Text-interdisciplinary Journal for the Study of Discourse, 8(3), 243-281.

Nadeem, F., Nguyen, H., Liu, Y., & Ostendorf, M. (2019, August). Automated essay scoring with discourse-aware neural models. In Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications (pp. 484-493).

Page, E. B. (1966). The imminence of... grading essays by computer. The Phi Delta Kappan, 47(5), 238-243. Persing, I., & Ng, V. (2015, July). Modeling argument strength in student essays. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 543-552).

Putra, J. W. G., Teufel, S., & Tokunaga, T. (2022). Annotating argumentative structure in English-as-a-Foreign-Language learner essays. Natural Language Engineering, 28(6), 797-823.

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring system: a systematic literature review. Artificial Intelligence Review, 55(3), 2495-2527.

Song, W., Wang, D., Fu, R., Liu, L., Liu, T., & Hu, G. (2017, July). Discourse mode identification in essays. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 112-122).

Stab, C., & Gurevych, I. (2014, October). Identifying argumentative discourse structures in persuasive essays. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 46-56).

Uto, M. (2021). A review of deep-neural automated essay scoring models. Behaviormetrika, 48(2), 459-484. Wambsganss, T., & Niklaus, C. (2022, May). Modeling persuasive discourse to adaptively support students' argumentative writing. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 8748-8760).

Wambsganss, T., Kueng, T., Soellner, M., & Leimeister, J. M. (2021, May). ArgueTutor: An adaptive dialog-based learning system for argumentation skills. In Proceedings of the 2021 CHI conference on human factors in computing systems (pp. 1-13).

Wang, X., Lee, Y., & Park, J. (2022). Automated Evaluation for Student Argumentative Writing: A Survey. arXiv preprint arXiv:2205.04083.

Wang, X., Bruno, J., Molloy, H., Evanini, K., & Zechner, K. (2017, July). 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) (pp. 263-268).

Wang, X., Gyawali, B., Bruno, J. V., Molloy, H. R., Evanini, K., & Zechner, K. (2019, June). Using Rhetorical Structure Theory to assess discourse coherence for non-native spontaneous speech. In Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019 (pp. 153-162).

