Dialogue and Discourse 4(2) (2013) 34-52 doi: 10.5087/dad.2013.202

Holistic Annotation of Discourse Coherence Quality in Noisy Essay Writing

Jill Burstein JBURSTEIN@ETS.ORG
Educational Testing Service
Rosedale Road, MS 11-R
Princeton, New Jersey 08540

Joel Tetreault[1] JOEL.TETREAULT@NUANCE.COM
Nuance Communications, Inc.
1198 East Arques Avenue
Sunnyvale, CA 94085

Martin Chodorow MARTIN.CHODOROW@HUNTER.CUNY.EDU
Hunter College and the Graduate Center of CUNY
695 Park Avenue
New York, NY 10065

Editors: Stefanie Dipper, Heike Zinsmeister, Bonnie Webber

Abstract

In this paper, we describe a holistic annotation scheme for coherence quality that requires little expertise on the part of the human annotator; we also present computational systems for evaluating coherence quality in essays. A formidable challenge to reliability when annotating discourse coherence comes from differences among annotators in the inferences that they draw when reading an essay. This may reflect differences in their background knowledge or in their willingness to bridge what might otherwise seem to be disconnected portions of the text. Despite these differences, we achieved adequate reliability for a holistic binary coherence annotation of essays written for five different types of large-scale assessment. When designing computational systems to score these essays for coherence quality, we faced a number of issues not encountered in most previous work, which has focused primarily on well-formed text. Foremost among these was the need to develop models based on features that reflect the coherence quality criteria which are found in human essay scoring guides, including aspects of writing quality (e.g., the presence of grammatical errors) that might interfere with constructing the meaning of the essay. Such features are needed to produce meaningful scores and to provide the basis for instructional feedback to the student or test-taker. We present results of testing various computational models on essays using the binary discourse coherence score.

Keywords: discourse coherence annotation, essay data, automated essay scoring

[1] The work reported here was completed while the second author was employed at Educational Testing Service.

Submitted 02/12; Accepted 11/12; Published online 04/13

1. Introduction

Motivation. In the field of automated essay evaluation, we are tasked with building systems that evaluate the overall quality of essays (Attali & Burstein, 2006; Burstein, Tetreault & Madnani, 2013; and, Elliot & Klobucar, 2013). These systems must be able to handle noisy essay data. In the context of automated essay evaluation for educational instruction and assessment, system features must be consistent with human scoring criteria to satisfy measurement standards. Specifically, are the system features used to predict the quality of the essay consistent with the scoring rubric criteria (i.e., criteria that are desirable to measure essay quality)? Further, system outcomes must be transparent (easily explainable). In other words, in the same way a teacher would explain why he or she had assigned a grade (score) to a writing assignment, we must be able to explain to score recipients (typically test-takers, teachers, and students) the system features that predicted outcomes. Consistent with this, in developing educational technology, explanation of system decisions is often achieved through explicit feedback aligned with scoring rubric criteria, especially in instructional settings. For instance, e-rater®, an automated essay scoring system used in assessment and instruction (Burstein, Chodorow, & Leacock, 2004; Attali & Burstein, 2006; Burstein et al., 2013), uses a set of well-defined features that are aligned with human scoring criteria. The system assigns scores to essays in standardized assessment and instructional settings.

In the United States, the practical need for text analysis capabilities for writing tasks is now being driven by increased requirements for standardized curricula and assessments. More recently, the need for applications for text analysis has been emphasized by the Common Core State Standards Initiative (CCSSI).[2] This relatively new initiative has now been adopted by most states for use in kindergarten through 12th grade (K-12) classrooms, and it is likely to have a strong influence on teaching standards in K-12 education. CCSSI illustrates what K-12 students should be learning with regard to Reading, Writing, Speaking, Listening, Language, and Media and Technology. More specifically, it describes language structures that learners need to grasp as they progress to the higher grades in preparation for college readiness in reading and writing.[3] In terms of writing skill and discourse coherence, there is a strong focus on students' ability to "*marshal an argument*." Given the continued emphasis on the need for feedback, and a specific interest in supporting students' ability to develop argumentation in text, the development of automated capabilities that evaluate aspects of argumentation such as, *discourse coherence quality*, will be critical.

Consistent with the need to build technology that can evaluate aspects of argumentation, we have been developing a system for automated evaluation of discourse coherence quality (Burstein, Tetreault, & Andreyev, 2010). A major goal is to ensure that the system addresses scoring criteria specifically related to discourse coherence in noisy essay writing. Human readers

[2] See http://www.corestandards.org/.

[3] See http://www.corestandards.org/assets/Publishers_Criteria_for_3-12.pdf.

who score essays written for standardized writing exams are instructed to consider discourse coherence quality, assigning lower scores to essays that are less coherent. For example, some essays are likely to contain ungrammatical structures, misspellings and typographical errors, poor organizational structure (e.g., lack of transitions), and other linguistic features that contribute to the difficulty readers may have in constructing the meaning of the text. To assess text quality, it is necessary to isolate the linguistic sources of this difficulty if we are to provide meaningful essay scores and instructional feedback that addresses relevant writing features contributing to overall essay quality.

Constructing Meaning. A core challenge in the evaluation of discourse coherence quality is that it is related to a reader's construction of text meaning. From Halliday and Hasan's (1976) perspective, a text is a semantic unit, as opposed to a large grammatical unit, or super-sentence. Further, there are a number of contributors to *textual continuity –the reader's ability to construct meaning from the text.* These include overt linguistic features, such as relationships between words, well-formed sentences, and meaningful transitions. As well, interpretation of a text plays a role in the reader's perception of the text continuity, or quality of discourse coherence. Also related to text interpretation, Shriver (1989) argues that the quality of a text can be judged to some extent by the *correct* (author-intended) inferences that readers draw from the text. From a cognitive psychology perspective, Graesser, McNamara, Louwerse, and Cai's (2004) view coherence is a psychological construct, and "*…coherence relations are constructed in the mind of the reader and depend on the skills and knowledge that the reader brings to the situation.* (p.1)." As is also pointed out in Graesser et al. (2004), readers' construction of meaning may differ based on individual inferences. Van den Broek (2012) asserts that skilled readers have a "large toolbox" of reading strategies that they can access to build coherence as they read a text. These include retrieving information from previous, earlier parts of the text, bringing in prior knowledge, and accessing other sources of information, for example, via internet search. Van den Broek observes that these processes are learned over time and become increasingly more automated with practice. In reading a text, users draw on these strategies in a "landscape of activations" in which they move between concepts from sentence to sentence.

Evaluation of Discourse Coherence Quality. In scoring writing assessments, how do human raters tasked with evaluating discourse coherence in a text determine what renders higher and lower discourse coherence? By design, conventional holistic scoring guides developed for standardized writing assessments are intended to offer human raters guidelines in terms of how to assign an overall score that takes into account many aspects of writing (Coward, 1950; Huddleston, 1952; and, Godshalk, Swineford, & Coffman, 1966). These include word choice, syntactic variety, grammaticality, and organizational quality. With regard to discourse coherence, while scoring guides provide a general idea of features that might play a role in evaluating discourse coherence quality, they do not prescribe to the rater the linguistic features that contribute specifically to good and bad qualities of coherence. In addition, no specific examples of essays illustrating variation in quality specifically related to coherence are provided. Rubric criteria in scoring guides that describe the relative quality of discourse coherence might read something like this: "*displays unity, progression, and coherence, though it may contain occasional redundancy, digression, or unclear connections*"; "*is not clearly organized; some* *parts may be clear while others are disjointed or confused... errors in grammar interfere with reader understanding"*. Such criteria are not clear-cut, and leave the necessary wiggle room for individual differences inherent in judgments about discourse coherence as suggested by Halliday and Hasan (1976), Shriver (1989), and Graesser et al. (2004).

We use a holistic approach in our annotation scheme that focuses on *reading to evaluate* versus *reading to comprehend* (Shriver, 1989). The research described in this paper shows promise that we can capture annotators' "impressions" of discourse coherence quality and provide a meaningful holistic evaluation score for discourse coherence in noisy essay data. By meaningful, what we mean is that we extract linguistically-grounded features from the annotated data set that can be mapped to scoring criteria related to discourse coherence and can also be translated into feedback for students and test-takers in instructional settings (i.e., classroom instruction or test preparation). Features related to discourse coherence are used to build and evaluate system models of discourse coherence quality. This paper focuses on a discussion of issues for annotation of discourse coherence quality and related annotation challenges, especially those related to the inferences that readers draw. We also consider the need for consistency between the feature set used for building coherence models and the scoring rubric criteria, in order for coherence scores to be meaningful and valid.[4] Further, to demonstrate the effectiveness of a holistic annotation scheme, we present our discourse coherence system development and evaluation.

The remainder of this paper is organized as follows. In Section 2, we discuss computational approaches to the evaluation of discourse coherence and related work on annotation methods used to build and evaluate systems for predicting text coherence and readability. In Section 3 we describe a holistic annotation scheme, related challenges, and the development and evaluation of discourse coherence models for five data sets. Section 4 provides further discussion and conclusions.

2. Related Work

Early approaches to discourse analysis of texts have, in general, been limited to lexical cohesion, specifically, repetition of vocabulary in a text (Hearst, 1997; Foltz, Kintsch, & Landauer, 1998). TextTiling, for instance, identifies lexical chains (essentially, repetition of vocabulary) across adjacent sentences following the notion of identifying subtopic structure in text (Hearst, 1997). Similarly, Foltz, et al.. (1998) describe systems that use latent semantic analysis (LSA) to measure lexical relatedness between text segments by using vector-based similarity between adjacent sentences.

The focus of this section is a taxonomy of annotation schemes related to discourse coherence in well-formed texts and noisy essays. The annotations are then used in supervised approaches for building systems to predict discourse coherence quality. Table 1 illustrates various aspects of annotation schemes which relate to the time and expertise required for each task, the task goals, and the text types to which the annotation scheme has been applied. In the context of

[4] See Attali and Burstein (2006) for a discussion of feature validity and human scoring criteria in the context of automated essay scoring.

the set of annotation schemes illustrated in Table 1, the holistic scheme that we apply to essay data and that we use for building coherence quality models is relatively low cost in that it requires less expertise and less time. In Section 3, we discuss how this low cost scheme was successfully used to assign accurate coherence quality scores to noisy essay data using linguistic features that can be mapped back to criteria in human scoring guides.

2.1 Annotation Schemes and Supervised Computational Approaches

Miltsakaki and Kukich (2000) were the first to complete an annotation study that focused on discourse coherence in essay data. They implemented a genre-independent annotation task that requires significant expertise related to Centering Theory (Grosz et al. 1995). The goal of this work was to use Centering Theory to identify topic shift in essays. In Miltsakaki and Kukich, 100 essays were manually annotated to label specific features that identified where Centering information occurred in the text. The data were used to show that essays with a higher proportion of *topic shifts* (i.e., *Rough Shifts* in Centering Theory) tended to have lower scores (based on a standard 6-point holistic scoring scale, where 6 indicates the best writing and lower scores indicate poorer writing skill.) More recently, Wang, Harrington, and White (2012) conducted a study that builds on Miltsakaki and Kukich in which annotators identified specific coherence breakdown points in essays. As part of this work, breakdown points in essays were provided as feedback to non-native English speaker writers. As an outcome essay revisions showed improvements in writing related to coherence features. Higgins, Burstein, Marcu, and Gentile (2004) implemented a genre-dependent annotation method and system to predict discourse coherence quality in essays. Their annotation scheme required expertise about essay-based discourse structure, specifically, the ability to rate the coherence between specific essay-based discourse elements, such as an essay's *thesis statement* and each of its *main points*. This method showed some success, but it was reliant on organizational structures that are most consistent with responses to expository and persuasive writing. Other discourse coherence schemes for well-formed text, such as Wolf and Gibson (2005), have required annotators to label text segments with particular discourse coherence relations, for instance, *cause-effect*, *example*, and *elaboration*. This work is similar to Miltsakaki and Kukich, and to Higgins et al.'s feature-specific annotation task, both of which require significant expertise.

Barzilay and Lapata (2005, 2008) also implemented a supervised method to predict discourse coherence quality in well-formed texts (also see Rus & Niraula, 2012). Their method takes into account the distributional and syntactic properties of entities (nouns and pronouns) in a text and is theoretically aligned with Centering Theory (Grosz, Joshi, and Weinstein, 1995). The theory asserts that the discourse in a text contains a set of textual segments that contain discourse entities. Centering Theory ranks these entities by their importance. The theory can be used to track the entities as well as topic clusters and shifts. Barzilay and Lapata's entity-grid algorithm keeps track of the distribution of entity transitions between adjacent sentences and computes a value for all transition types based on their proportion of occurrence in a text. The algorithm has been evaluated with three tasks using well-formed newspaper corpora: *text ordering*, *summary coherence evaluation*, and *readability assessment*. For *summary evaluations*, the entity-grid approach was successfully used to predict *coherence quality* of summaries written by humans versus machine-generated summaries. For *readability assessment*, Barzilay and Lapata (2008) used their algorithm to determine if a well-formed text was an article from an elementary-level version of Encyclopedia Britannica or the adult-level version of the same article. They found increased performance of a system that predicted readability (higher or lower grade level) when standard readability syntactic measures were added to the model that used the entity grid alone (also see Graesser et al., 2004; Sheehan, Kostin, Futagi, & Flor, 2010; and Graesser, McNamara, and Kulkowich, 2011 for evaluations that discuss readability assessment and coherence measures). Barzilay and Lapata's measures used parse tree information (Charniak, 2000) and included sentence length, average parse tree height, average number of NPs, average number of VPs, and average number of subordinate clauses (Schwarm & Ostendorf, 2005). Like other evaluations of readability assessment, features addressing technical quality were not included since this work analyzed only well-formed texts.

Using well-formed texts, Pitler and Nenkova (2008) show that a text coherence detection system yields the best performance when it includes features using the Barzilay and Lapata (2008) entity grids, syntactic features, discourse relations from the Penn Discourse Treebank (Prasad, Dinesh, Lee, Miltsakaki, Robaldo, Joshi, and Webber, 2008), and vocabulary and length features. Evaluations were based on 30 Wall Street Journal articles that were annotated for coherence by at least three college students. The coherence annotation scheme, similar to our holistic scheme (described in Section 3), asked annotators to provide a rating on a scale of 1 – 5 in response to each of these four questions related to the article's coherence: (1) *How well-written is this article?*; (2) *How well does the text fit together?*; (3) *How easy was it to understand?*; and (4) *How interesting is this article*?

Crossley and McNamara (2011) also developed an annotation scheme for discourse coherence in essays. This is the only other work, to our knowledge, that includes a technical quality measure (i.e., an annotation measure that assesses Standard English conventions) for evaluating a system for automated detection of coherence quality. In Crossley and McNamara, raters assigned a holistic score (from 1 to 6) to multiple traits of an essay which were believed to contribute to essay coherence. These addressed the following categories: (a) quality of the Introductory Paragraph, including connections between thesis and essay discourse; (b) connections between topics across paragraphs in the essay; and (c) grammar, syntax, and mechanics. As Crossley and McNamara discuss, their annotators were experts so this was a higher cost annotation task with regard to *expertise*. Their findings differ from Burstein et al.'s (2010) findings and those reported later in this paper inasmuch as their results did not show technical quality (e.g., grammatical errors) to be a strong predictor of coherence quality in the context of test-takers' essays. This may be related to the nature of their essay data, which consisted only of essays written by native speakers during an undergraduate composition course. This is possibly a population of more proficient writers who are less likely to make grammatical and spelling errors in their writing.

Pitler and Nenkova's (2008) annotation scheme probably most closely reflects the annotation approach that we use in our work (Burstein et al., 2010), which is described in Section 3. Like Pitler and Nenkova, our annotators use a holistic annotation scheme to rate the overall coherence of a text. The task is completed at the document level and requires no specific linguistic or essay-scoring expertise. The major difference between our work and Pitler and

Nenkova's is the data itself. They annotate well-formed text, and we annotate noisy essay data. That said, it is not surprising that our feature set for predicting coherence includes a technical quality feature, and theirs does not.

Reference	Scheme Class	Text Type	Scheme Goal(s)	Annotation Level	Expertise Level
Miltsakaki and Kukich (2000)	FS	Essay Responses	Capture Centering Theory "Rough Shifts"	Word	High
Higgins et al. (2004)	FS	Essay Responses	Identify Lexical Relationships between Discourse Segments in Essays (e.g., Thesis, Main Points, Supporting Ideas, Conclusion)	Discourse Segment	Medium
Wolf and Gibson (2005)	FS	*Wall Street Journal & AP Newswire* articles	Identify rhetorical relations related to coherence, e.g., *cause-effec*t	Discourse Segment	High
Pitler and Nenkova (2008)	H	*Wall Street Journal* articles	Determine 'ease' of reading for a text	Document	Low
Barzilay and Lapata (2008)	H	1. Human and machine-generated summaries 2. Elementary- and adult-level *Encyclopedia Brittanica* articles	1. Determine Discourse Coherence Quality of Summaries 2. Predict text level (elementary or adult)	Document	Low
Burstein et al. (2010)	H	Essay Responses	Determine low/high discourse coherence quality	Document	Low
Crossley and McNamara (2011)	H/MT	Essay Responses	Predict overall essay score	• Discourse Segments • Intra-Text Segments • Document	High
Wang, Harrington, & White	FS	Essay Responses	Identifies breakdown points in essays	• Intra-Text Segments	High

Table 1. Taxonomy of Discourse Coherence Annotation Schemes. FS is feature-specific, H is holistic, and H/MT is Holistic for Multiple Coherence Traits (e.g., topic sentences, paragraph transitions, organization).

3. Study: Annotation, Feature Development, and System Evaluation

As illustrated in Table 1, our holistic annotation scheme attempts to be less time-consuming, is performed at the document level, and requires no specific linguistic expertise. It is similar to the document-level annotation approaches used in Barzilay and Lapata (2008) in which raters assigned a coherence rating on a 7-point scale with regard to relative coherence quality of multidocument summaries generated by humans and an automatic summarization system, and to the Pitler and Nenkova (2008) work in which annotators responded to a set of questions about text quality.

The annotation scheme elicits holistic scores from annotators about essays as opposed to asking for labels related to specific linguistic structure as has been done in previous work with discourse coherence in essay data (Miltsakaki and Kukich, 2000; Higgins et al., 2004). It is also more closely aligned with the holistic scoring process, in which readers assign an overall score to an essay based on a general impression, given fuzzy criteria. In our work, annotators essentially assign a holistic (impressionistic) score specifically for discourse coherence quality based on a few features that may contribute to an essay's coherence (e.g., *displays unity, progression and coherence*). It is critical that systems designed to evaluate linguistic aspects of test-taker or student writing address valid measurement criteria specified in scoring guides. In theory, these criteria are likely to be related to text features that readers use to construct text meaning. From a practical standpoint, these criteria need to be made available to users for score explanation. Users include test-takers who might be preparing for an assessment or who might question a score reported from a standardized test, as well as teachers or students who might inquire about an evaluation in an instructional environment.

In the remainder of this section, we will describe our data, the annotation scheme, the training and data annotation process, and the subsequent use of the annotated data to build coherence quality models that predict low- and high-coherence in essays.

3.1 Data and Annotation

Data. Essay data sub-corpora included writing samples from native and non-native English speakers, ranging from 6th to 12th grade, and across undergraduate and graduate-level writing assessments. Writing task types were varied and included expository and persuasive writing, subject-based writing, and summary writing. The total number of essays was 1555. There were five different task types (essay data sub-corpora). Data sample sizes ranged from approximately 250 to 400 essays.

Annotation Scheme. The annotation scheme is composed of a 3-point scale as follows: Score Point 1 indicates that an essay is incoherent (low coherence; no meaning can be constructed); Score Point 2 indicates that an essay is mostly coherent (essentially coherent; text meaning can basically be constructed, but one or two identifiable points were confusing); and, Score Point 3 indicates that there are no problems with coherence (high coherence; text meaning can easily be constructed). Annotators labeled each essay with one of the three score points.

For Score Point 2, annotators had to do an additional task. Score Point 2 essays represent those essays in which most of the meaning of the text can be easily understood, but there are one or two *identifiable points* where coherence breakdown has occurred. For these essays, annotators had to label the "awkward sentence(s)" where the coherence breakdown occurred. Annotators could also add comments to describe the confusion points in the essays where it may have taken a few tries to construct meaning or where meaning construction might have been unsuccessful. The annotation protocol includes specific examples of essays at each of the three score points. For

Score Point 2, the protocol also contains example awkward sentences that illustrate where sources of coherence breakdown appear in an essay, as in the essay excerpt in Figure 1 below. While there may be different interpretations of this text, we have indicated in ***italicized bold*** font our judgments about which sentences were awkward. In (1), the *awkward sentence* [1] reference to "these companies" does not have a natural antecedent and is confusing. In *awkward sentence* [2], it is not clear who or what "posers" are. In *awkward sentence* [3], there is a similar issue, and "products" does not have a natural antecedent. These points of confusion can be resolved using inference so that we can essentially construct meaning for this text. It is therefore assigned to Score Point 2.

> *Media in all physical or visual forms from magazines, to movies, to billboards, and television display images of beauty from both male and female. Not a day goes by when a person is not exposed to these campaigns.* ***These companies are trying to promote their product or service in the most elegant way, but the qualities seen on the images are compared with themselves who may not match up by far.*** [1] ***When look at these images, it's easy to forget that the posers depend on their careers to look the way they do, but the observers do not.*** [2] *Thus, a desire to have the same look continuously builds and results in taking action. It is the state of mind that they are dissatisfied with their appearance and they will take any steps to gain that appearance. Psychological problems may occur and lead to serious disorders such as anorexia.* ***Furthermore, the products or others related to that kind, are sold leading to obsessive and unnecessary spending.*** [3]

Figure 1. Excerpt from an essay labeled as Score Point 2. Italicized bold font indicates sentences that caused confusion for the reader (annotator). Different readers may find different aspects confusing.

In the protocol, annotators were instructed not to consider grammatical or spelling errors as lowering coherence scores, unless those errors interfered with the annotators' ability to construct meaning. This is aligned with Shriver's (1989) assertion that such error types are not notable and we tend to ignore them while reading, unless they slow down our reading and make us re-read. That said, the interference of grammatical and spelling errors should appear in essays that annotators assign Score Point 1 or 2. In the case of Score Point 2 writing, these would be cases where such errors caused the annotator (reader) to have to re-read a part of the text because grammatical or spelling errors interfered with understanding of a particular section of the essay.

Annotator Training. During the initial training and development of the annotation scheme (Burstein et al., 2010), two annotators worked with two of the authors. Both annotators were research assistants who had experience doing varied kinds of linguistic annotation. However, neither annotator had previous experience with annotating discourse coherence. The initial three data sets, including the 6th-12th grade data, and college undergraduate and graduate level writing assessments, were labeled by two annotators as described in Burstein et al. (2010). The annotation scheme was developed and refined over a period of about one month in discussion with the first two authors and two annotators from Burstein et al. (2010). Training began at the point at which the authors and the annotators agreed on the annotation protocol instructions and the examples of essays that illustrated coherence at each of the 3 score points. Seventy-five essays were annotated using a training set that contained essays from different task types and grade levels for three of the five data sets reported in this paper. Annotations took approximately one week. Annotators had less than acceptable agreement for Score Point 2, but weighted kappa of 0.68 was achieved overall for the 3-point scale. *Beyond training*, annotators continued to label essays using the 3-point scale, but in the end, Score Point 2 inter-annotator agreement remained a challenge. The annotators could not reach acceptable levels of agreement for this middle score. Therefore, the three score points were collapsed to a 2-point scale. *Score point 1* remained the classification for essays with *low coherence*, and *Score Points 2 and 3* were combined into a single class to represent essays with *high coherence*. Based on the two-point scale, kappa for this final set of data was 0.67 for approximately 750 essays (see Burstein et al., 2010 for details). These annotations were then successfully used to build a discourse coherence system that assigned *low* and *high* coherence scores to essays across different writing tasks and test-taker and student populations.

Correlations with holistic essay scores. Part of the decision to include a new feature in an automated essay scoring system (e-rater® in this case; see Attali & Burstein, 2006) is determined by examining the correlations between the new feature and the human rater holistic essay score. It is important to determine that the new feature is accounting for additional variance and is not redundant with existing features used to predict the holistic essay score, and that it offers more fine-grained information related to the *ease of reading* (i.e., the ease with which a reader constructs meaning). Therefore, for the first three data sets from the 2010 study, correlations were calculated to evaluate the relationships among the 2-point coherence quality scores, existing features, and the human rater holistic essay scores. These holistic essay scores range from 1 to 5 or from 1 to 6, depending on the assessment. The score of 1 indicates a lower quality essay, and the higher scores of 5 and 6 indicate higher quality essays. Pearson correlations between human discourse coherence scores and human (overall) holistic essay scores were between 0.46 and 0.58. While these are moderate-to-strong correlations, they still suggest that the discourse coherence quality scores capture characteristics of essay writing that are not fully explained by the holistic essay scores. For example, it is possible for an essay with a low holistic score to have high coherence. That essay might have received a low score for reasons independent of coherence, such as not offering a sufficient amount of supporting evidence, or not responding to the essay prompt. We address this point further in the Evaluation section in Tables 2 and 3, where in end-to-end comparisons, the discourse coherence systems we tested consistently outperformed e-rater in prediction of discourse coherence quality scores. This outcome is consistent with the correlations between the human coherence and holistic essay scores.

Annotation Challenge: Score Point 2. As mentioned above, Score Point 2 remained a challenge, since inter-annotator agreement was unacceptably low, resulting in a kappa below 0.60, the minimum value which is generally considered “substantial” agreement (Landis & Koch, 1977). Without a substantial rate of agreement, system building based on annotations is often difficult and unreliable. As a result, we could not build a system with acceptable agreement across our 3-point scale, requiring us to instead use the binary classifications of *high* and *low* coherence. A possible explanation for low agreement for Score Point 2 may be related to an individual reader’s inferencing decisions. Recall that Score Point 1 essays are those that are incoherent to the extent that the breakdown points could not be identified; a Score Point 3 essay was one that could be read easily without noticing any points of coherence breakdown. By contrast, Score Point 2 labels indicated that a few *identifiable* coherence breakdown points could be located. From the annotation task, however, it seemed that annotators could not easily agree on Score Point 2 labels. As discussed earlier, Graesser et al. (2004) and van den Broek (2012) assert that readers *construct* the meaning of a text. Van den Broek suggests that to build coherence while reading a text, readers consult a set of strategies, including integrating sections of the text, but also accessing external sources of information, including prior knowledge (see van den Broek (1993)'s discussion of "types of inferences"). It would not be unreasonable to assume that our annotators' prior knowledge or perspectives on an essay topic influence how they build up meaning and how they ultimately assign a coherence score. More concretely, though, reading through several cases in our data where the annotators differed, with one assigning a Score Point 2 and the other a Score Point 3, the annotator who assigned the Score Point 2 typically indicated in a "comments section" that the *awkward* sentences she identified as causing coherence breakdown seemed "out of place" or were "missing a transition". The lack of transitional elements seemed to inhibit the annotators' ability to construct meaning in a segment of an essay, leading to a Score Point of 2. The level of disagreement among Score Point 2 essays would indicate that readers (annotators) were constructing meaning in texts differently in these cases: one annotator seemed able to make the transitions while another could not. Figure 2 illustrates an example of an essay that one annotator assigned Score Point 3 and the other Score Point 2. The annotator who assigned Score Point 2 indicated that transitions were missing and certain sentences seemed out of place. On the other hand, the annotator who labeled the essay as Score Point 3 was able to construct meaning and figure out the relationships between these sentences and other sentences either in the text or, perhaps, related to some prior knowledge about the topic.

It may be difficult to build systems that identify this kind of fine-grained distinction when humans cannot agree. However, what we can do is to capture text segments in essays, such as discourse connectors, or repeated concepts that serve as a continuous thread in the text. Capturing this information can lead to the development of useful feedback in instructional settings, which can help users understand if they are missing transitions that could result in a lack of clarity in their writing. Systems could also identify text segments or concepts that did not appear to be related to other parts of the text (lexically) and could flag these as *potentially* unconnected ideas. Features capturing discourse relations and repeated concepts are used in our system and are described below. In theory, these features could be used to identify weak relationships between text segments as well.

3.2 System Features and Evaluation

Using n-fold cross-validation with C5.0, a decision-tree classifier (Quinlan, 1993), systems were built to classify essays in each of the five data sets, which represented five different writing tasks and populations. Across the five tasks, there was a mixture of native and non-native speaker data. Task types varied and included essays that call for summarization and for expository and persuasive writing. Figures 3 and 4 illustrate relevant discourse features that were predictive of coherence quality.

Throughout the history of time there have been many great leaders, along with many poor ones. These leaders exemplify traits above all others that help them become what we see them as today.

Gandhi was a powerful man who used civil disobedience as his weapon against guns, to help Hindu's and Muslims unite in peace. Though his acts started off small they gradually grew to be what we consider the greatest non-violent movements of all time. Gandhi's first movement was when he burnt his pass card - which restricted Muslims and Hindus to travel without them. for this he was beaten by British soldiers and taken to prison. **Later on Mahatma Gandhi started such movements as the homespun movement and the salt march. [1]**

The homespun movement was when Gandhi refused to wear any British cloth therefore made his one clothes that were merely whites pieces of cloth worn in a toga style. **From there he made speeches all across India about this movement. [2]** Hindu's and Muslim's responded with great support to this movement and also participated.

…

Gandhi's life came to an end when he was shot walking around at a non- violent movement by a Muslim who disagreed with his beliefs. Once Gandhi died the fighting between this Muslim's and the Hindu's went down hill and only got worse until the British freed India to become its own country. So although Gandhi worked for many years to bring peace to a land of war, his dream did not come true until many years after his death.

Figure 2. Excerpt from an essay labeled as Score Point 2 by one annotator and Score Point 3 by a second annotator. The Score Point 2 annotator indicated that **awkward sentence [1]** "needed a transition" and **awkward sentence [2]** "did not fit into the paragraph".

3.2.1 Feature Set. The set of features used in each model is intended to capture holistic rubric criteria from scoring guides that correspond to discourse coherence in high quality essays: "*displays unity, progression, and coherence, though it may contain occasional redundancy, digression, or unclear connections*"; and in low quality essays: "*is not clearly organized; some parts may be clear while others are disjointed or confused... errors in grammar interfere with reader understanding"*. These features include the following: proportion of grammatical errors using error features from e-rater (Attali & Burstein, 2006; Burstein et al., 2013), entity-grid transition probabilities (Barzilay and Lapata, 2005; Barzilay& Lapata, 2008), rhetorical structure theory (RST) features (Mann and Thompson, 1988) derived from an RST parser (Marcu, 2000), and a type-token feature that computes the entity type/token ratios from specific words and terms recovered from the entity grid. High-level descriptions of the features are given below. (For more detailed descriptions of the entity-grid approach, refer to Barzilay & Lapata, 2008); for more detail about the rhetorical structure tree parser, see Marcu, 2000). Below, we describe each feature type and its corresponding scoring rubric criterion.

Entity-grid transition probabilities (Entity Grids). Entity-grid transition probabilities are intended to address *unity*, *progression* and *coherence* by tracking nouns and pronouns in text. The entity grids characterize occurrences of words across a text in syntactic roles. An entity grid is constructed in which all entities (nouns and pronouns) are represented by their roles in a sentence (i.e., Subject**,** Object**,** Other**).** *Entity grid transitions* track how the same word appears in a syntactic role across adjacent sentences. Examples of transition types are Subject-Object, Object-Object, and Object-Other. *Entity transition probabilities* represent the proportions of different entity transition types in a text. These probability values are used as features to build coherence models.

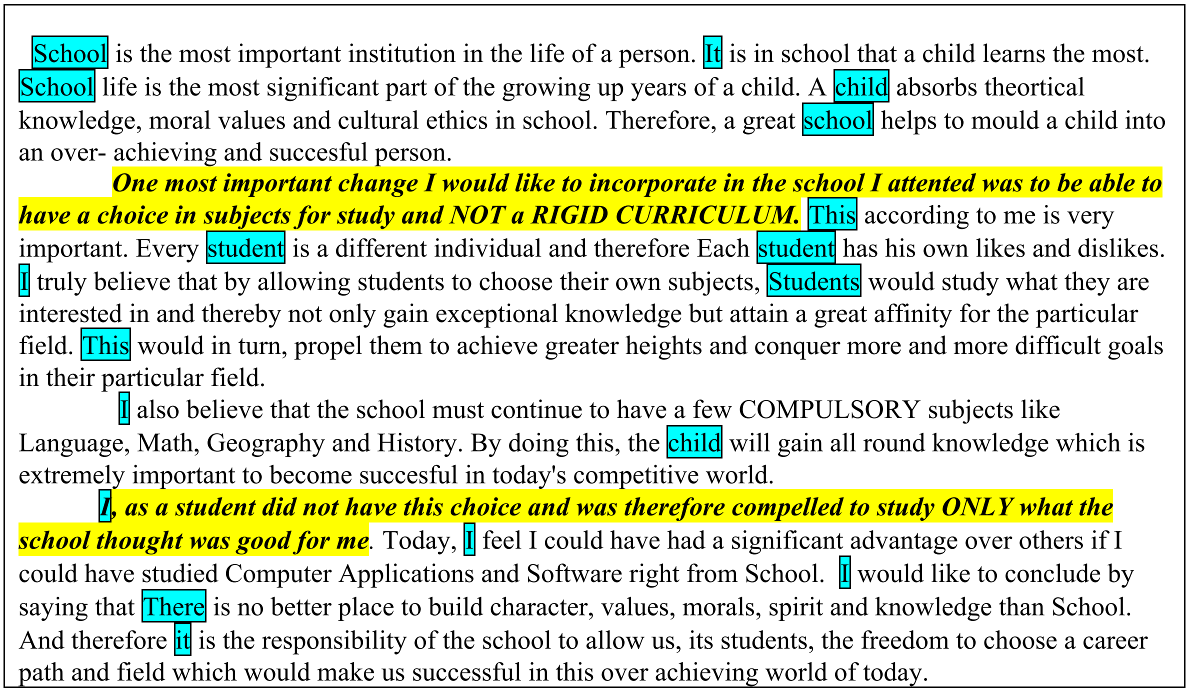
E-rater grammatical error features (ERGramError). These features address errors in grammar that could interfere with a reader's ability to construct meaning. For example, in a sentence such as, *It make student to have a good rest than staying in the school for whole term*, a reader may have to pause and consider the possible intentions of the writer. E-rater identifies more than 30 kinds of errors in grammar, such as subject-verb number disagreement, in word usage, such as missing articles, and in spelling (Attali & Burstein, 2006; Burstein et al., 2013). Aggregate counts of these individual errors are used as features in e-rater, and these same features were used in our model for predicting discourse coherence quality.

Maximum LSA Value for Distant Sentence Pairs (maxLSA). In the course of our research, we observed that some high coherence essays tended to reintroduce earlier topics later in the essay. This reintroduction provided a sort of "thread" that maintained coherent connections promoting *discourse unity* within the text. To represent this, we used Latent Semantic Analysis (LSA), a statistical method used to determine semantic similarity between text units (i.e., do these text units use similar vocabulary) (Foltz et al., 1998). Previously, Higgins, et al.(2004) had used a similar technique, random indexing (Kanerva, Kristoferson, & Holst, 2000; Sahlgren, 2001), to measure similarity between discourse unit text segments within an essay (e.g., similarity between the thesis statement and the conclusion). With LSA, we computed similarity values for all sentence pairs in the essay and found that the maximum LSA value associated with pairs that were separated by more than five intervening sentences was highly correlated with the human coherence annotations. This feature, which measures similarity at a distance and should, therefore, be sensitive to reintroduction of topics, was added to our set of features. Re-introducing a concept later in a text is consistent with the backward inference strategy discussed in van den Broek (1993). The positive correlations between the maxLSA value and discourse coherence scores suggest that clear re-introduction of material presented earlier in a text does contribute to higher coherence. Table 4 indicates that it was predictive of the holistic coherence quality in two of the five data sets we tested.

RST-derived features (RST). Rhetorical relations tell us about *discourse connections* with regard to how clauses and sentences in a text are rhetorically linked. RST relations include, for example, *Antithesis, Cause, Comparison*, *Elaboration* and *Example*. We used rhetorical relations and derived features to evaluate if and how certain rhetorical relations, combinations of rhetorical relations, or rhetorical relation tree structures might contribute to discourse coherence quality. These included the following: (a) relative frequencies of n-gram rhetorical relations in the context of the tree structure (unigrams, or occurrences of a single relation (e.g., *ThemeShift*); bigrams (e.g., "*ThemeShift -> Elaboration*"); and trigrams (e.g., "*ThemeShift -> Elaboration ->*

Circumstance"); (b) relative proportions of leaf-parent relation types in rhetorical structure trees; and (c) counts of root node relation types in rhetorical structure trees. See Figure 4 below for an example of an RST parse tree (Marcu, 2000).

Type/Token Ratios for Entities (Type/Token). Type/token ratios can be used to track redundancy in essays. As discussed above, the entity grid transition probabilities are used as features in the coherence models, but they measure only local transition patterns in adjacent sentences rather than more global reuse of words. For instance, if there is a high probability for the "Subject-Subject" transition to occur in an essay, this indicates that the writer is repeating an entity in Subject position in adjacent sentences, but we do not know if the same word is being repeated throughout the text or if a variety of words are. The type/token feature can distinguish between these two cases. For instance, if there are 10 Subject-Subject transitions, and 5 different word types appear in these transitions, the type/token ratio would be 5/10 (0.50); if, on the other hand, there is only 1 word type (e.g., "*I*"), the type/token ratio would be 1/10 (0.10). Thus, higher ratios indicate that more concepts are being introduced in a given syntactic role, and lower ratios indicate fewer concepts.



School is the most important institution in the life of a person. It is in school that a child learns the most. School life is the most significant part of the growing up years of a child. A child absorbs theortical knowledge, moral values and cultural ethics in school. Therefore, a great school helps to mould a child into an over- achieving and succesful person.

One most important change I would like to incorporate in the school I attented was to be able to have a choice in subjects for study and NOT a RIGID CURRICULUM. This according to me is very important. Every student is a different individual and therefore Each student has his own likes and dislikes. I truly believe that by allowing students to choose their own subjects, Students would study what they are interested in and thereby not only gain exceptional knowledge but attain a great affinity for the particular field. This would in turn, propel them to achieve greater heights and conquer more and more difficult goals in their particular field.

I also believe that the school must continue to have a few COMPULSORY subjects like Language, Math, Geography and History. By doing this, the child will gain all round knowledge which is extremely important to become succesful in today's competitive world.

I, as a student did not have this choice and was therefore compelled to study ONLY what the school thought was good for me. Today, I feel I could have had a significant advantage over others if I could have studied Computer Applications and Software right from School. I would like to conclude by saying that There is no better place to build character, values, morals, spirit and knowledge than School. And therefore it is the responsibility of the school to allow us, its students, the freedom to choose a career path and field which would make us successful in this over achieving world of today.

Figure 3. Illustration of example discourse features automatically detected and used to generate a discourse coherence quality score for this essay written by a non-native speaker, from the EP/NNS data set described in Tables 2 and 3. ***Italicized bold*** indicates the long distance sentence pair with the **maximum LSA value**, illustrating the re-introduction of a similar concept, promoting *discourse unity* in the essay; Turquoise highlighted words indicate entity types used in the **type/token feature** that represents the repetition of *entities* extracted from the *entity grid*. In this case, there are reasonably strong relationships among a small and related variety of entities, e.g., *school*, *child*, and *student(s)*, which support lexical cohesiveness, or *discourse unity*. Other entities are referential or personal pronouns associated with statement of opinion.

3.2.2 Evaluation. Developing a system that assigns discourse coherence on a 3-point scale has been a challenge. As discussed earlier, it appears that for the Score Point 2 category, annotators had lower agreement. Therefore, the Score Points 2 and 3 were collapsed to represent the "high" coherence class, and all Score Point 1 annotations are assigned to the "low" coherence class. Using these two classes, we have built systems that assign high and low coherence scores and compared them to three baselines. Baseline systems included: (1) assigning the *majority class* (i.e., assignment of the more frequent category, "high coherence", to every essay); (2) using the *e-rater* automated essay scoring system for assigning coherence scores; and (3) using only *e-rater's grammatical error features* to assign coherence scores.

Performance on high and low coherence essays is measured in terms of *precision*, *recall* and *F-score*. A system's precision on high coherence essays, for example, is the number of essays that both the system and the human annotator agreed are "high", divided by the number of essays that the system labeled "high". Recall is the number of essays that both the system and the human annotator agreed are "high", divided by the number of essays that the human labeled "high". The F-score is the harmonic mean of precision and recall. Precision, Recall and F-score can be calculated for "low" coherence essays in a similar manner. Tables 2 and 3 below show separate performance measures for high and low predictions, respectively. These measures are more informative about system performance than the overall values where low and high predictions are collapsed. The best system outperformed the F-score baselines for four of the five data sets for prediction of "high" coherence, and across all five data sets in the prediction of "low" coherence. For the Professional Proficiency exam, it should be noted that the data for this exam contained a much lower proportion of Score Point 1 essays than the other data sets, and the system had a more difficult time predicting the low coherence essays in this set. In Tables 2 and 3, the best system (**boldface**) for each data set was built using a combination of the features described above in section 3.2.1. Table 4 shows the features that contributed to the best system for each data set.

System	**EP/ NNC** n=196	**Summary/NNC** n=304	**EP/GL** n=210	**PPE** n=355	**EP/ 6-12** n =220
	P/R/F	**P/R/F**	**P/R/F**	**P/R/F**	**P/R/F**
Majority Class	77/100/87	76/100/76	82/100/90	91/100/95	87/99/92
E-rater	87/86/86	81/90/86	86/97/91	91/97/94	89/91/90
ERGramError	90/85/88	78/93/85	86/97/91	91/100/95	90/94/92
Best System	**93/91/92**	**85/94/89**	**91/94/93**	93/96/95	**94/99/96**

Table 2. P/R/F indicates system Precision, Recall, and F-scores (x 100) for "**High**" Discourse Coherence essays for 5 Essay Data Sets. EP/NNC is expository and persuasive, non-native, college level assessment writing; Summary/NNC is summary, non-native, college level assessment writing; EP/GL is expository and persuasive, graduate level assessment writing; PPE is the Professional Proficiency Exam; and EP/6-12 is elementary, middle, and high school writing. E-rater = e-rater system features. ERGramError = E-rater Grammatical Error features.

4 Discussion and Conclusions

In this paper, we have described and compared different discourse coherence annotation schemes and related studies. Most work in this area has been evaluated for building systems to handle well-formed texts. There has also been considerably more work in the area of coherence and its relationship to readability (*reading for comprehension*) as opposed to coherence quality in noisy data (in this case, essays) (*reading for evaluation*). We have shown that a holistic annotation scheme that requires no linguistic expertise can be successfully used to build discourse coherence systems that classify low- and high-coherence quality in 1555 essays from 5 different data

System	**EP/ NNC** n=58	**Summary/NNC** n=94	**EP/GL** n=47	**PPE** n=37	**EP/ 6-12** n =33
	P/R/F	**P/R/F**	**P/R/F**	**P/R/F**	**P/R/F**
Majority Class	0/0/0	0/0/0	0/0/0	0/0/0	0/0/0
E-rater	54/57/55	56/33/40	68/32/43	14/5/8	32/27/30
ERGramError	58/69/63	39/14/20	68/28/39	0/0/0	41/27/33
Best System	**72/75/74**	**71/47/57**	**69/57/63**	**48/32/39**	**84/52/64**

Table 3. P/R/F indicates system Precision, Recall, and F-scores (x 100) for "**Low**" Discourse Coherence essays for 5 Essay Data Sets. EP/NNC is expository and persuasive, non-native, college level assessment writing; Summary/NNC is summary, non-native, college level assessment writing; EP/GL is expository and persuasive, graduate level assessment writing; PPE is the Professional Proficiency Exam; and EP/6-12 is elementary, middle, and high school writing. E-rater = e-rater system features. ERGramError = E-rater Grammatical Error features.

Data Set	**Best System Feature Set**
EP/ NNC	EntityGrid + Type/Token + RST + maxLSA
Summary/NNC	ERGramError + RST
EP/GL	EntityGrid + ERGramError + Type/Token + RST
PPE	EntityGrid + ERGramError + Type/Token + RST
EP/ 6-12	EntityGrid + ERGramError + Type/Token + RST+ maxLSA

Table 4. Feature sets used in the best system for each data set. EP/NNC is expository and persuasive, non-native, college level assessment writing; Summary/NNC is summary, non-native, college level assessment writing; EP/GL is expository and persuasive, graduate level assessment writing; PPE is the Professional Proficiency Exam; and EP/6-12 is elementary, middle, and high school writing.

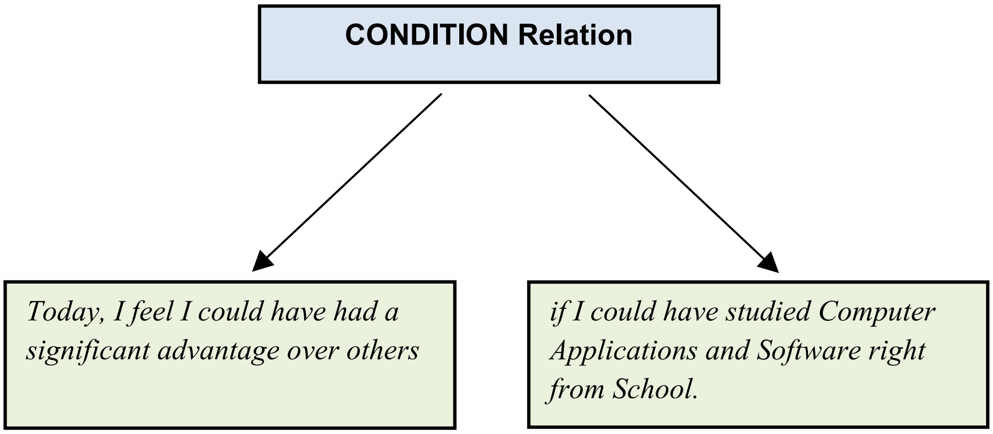

CONDITION Relation
Today, I feel I could have had a significant advantage over others
if I could have studied Computer Applications and Software right from School.

Figure 4. Illustration of an RST representation for a sentence from the essay in Figure 3. The RST relations contribute to discourse relationships and conceptual connectedness within the essay that can support readers in building meaning and coherence. In this illustration, the RST parser identified a CONDITION Relation within the sentence "*Today, I feel*

I could have had a significant advantage over others if I could have studied Computer Applications and Software right from School."

samples, including native and non-native speaker populations, 6th through 12th grade, and college- and graduate-level populations, and from numerous topic domains across the multiple essay question topics in our essay sub-corpora. The features that are used to build models can be mapped back to scoring guide criteria in the following way. Entity-grid transition probabilities offer information that can be related to the organization of ideas and how they are distributed in a text; the type-token feature based on the entities from the grid offers information related to the degree of repetition of ideas in the text; rhetorical structure features provide a sense of the organizational and development units of discourse in the text; and the grammatical error features provide a sense of the technical quality of the text. These features could be used to provide explanations of coherence scores and feedback to students, test-takers, teachers, and other score recipients.

As discussed throughout the paper, the most challenging aspect in this work was annotation of Score Point 2, "*mostly coherent"* essays – specifically, those essays where the reader experiences one or two identifiable points of confusion (coherence breakdown). For system model building and score prediction, we collapsed the Score Point 2 and Score Point 3 essays into a single "high" class. Essentially, when essays had very low or very high coherence, annotators could agree. A possible explanation for the assignment of Score Point 2 may be related to differences in how individual readers construct meaning or employ inference during reading. This is consistent with Halliday and Hasan's (1976) assertion that a text is a semantic unit and that the text continuity is related to how an individual constructs its meaning. Especially with regard to employing inference strategies, this might also be explained by Graesser et al.'s 2004 assertion that coherence is a psychological construct where individuals may vary in their use of inference, and similarly, by van den Broek's (1993 and 2012) discussions of readers' reading strategies that take into account not only the text itself, but also external sources, such as internet search and prior knowledge.

In a study that specifically investigated readers' construction of text meaning, Kwong (2010) conducted a discourse coherence annotation task using Aesop's fables. In the study, for sentences in each story, three annotators specified key words, main information, inferred information, and the relation to the moral of the story. With regard to keywords and main information, agreement could be reasonably measured by word overlap; however, for inference and relation to the moral of the story, annotator responses varied and seemed to rely on world knowledge. Since assignment of Score Point 2 indicates that the reader had some confusion, it is possible that while one reader was confused, another reader might not be confused because he or she made an inference that held the text together. An interesting direction for future work would be to investigate the interaction among inference, prior knowledge, and the distribution of content within a text by replicating Kwong's study, but using essays instead of fables.

Acknowledgements

We would like to thank Daniel Blanchard and Slava Andreyev for engineering support in system development and evaluation. We would also like to thank Sarah Ohls, Waverly Van Winkle, and Jonathan Schmidgall for their valuable input during the development of the annotation protocol and for their subsequent annotation of the essays used in this research.

References

Attali, Y. & Burstein, J. (2006). Automated Essay Scoring with *e-rater* v.2.0.. *Journal of Technology, Learning, and Assessment*, 4(3).

Barzilay, R. & Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics,* 34(1), 1–34.

Barzilay, R. & Lapata, M. (2005). Modeling Local Coherence: An Entity-Based Approach. In *Proceedings of the 43rd Annual Meeting of the Association of Computational Linguistics*, 141–148, Ann Arbor, MI.

Burstein, J. Tetreault, J., & Andreyev, S. (2010). Using Entity-Based Features to Model Coherence in Student Essays. In *Proceedings of 10th Annual Meeting of the Human Language Technology and North American Association for Computation Linguistics,* 681–684, Los Angeles, CA.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The E-rater® Automated Essay Scoring System. In Shermis, M.D., & Burstein, J. (Eds.), *Handbook for Automated Essay Scoring*. New York: Routledge.

Charniak, E. (2000). A maximum-entropy-inspired parser. In *Proceedings of the 1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, 132–139, Seattle, WA.

Coward, A. F. (1950). The Method of reading the Foreign Service Examination in English Composition. ETS RB-50-57, Princeton, NJ: Educational Testing Service.

Crossley, S. A., & McNamara, D. S. (2011). Text coherence and judgments of essay quality: Models of quality and coherence. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society.* 1236–1241. Austin, TX: Cognitive Science Society.

Elliot, N. & Klobucar, A. (2013). Automated Essay Evaluation and the Teaching of Writing. In Shermis, M.D. & Burstein, J. (Eds.), *Handbook for Automated Essay Scoring*. New York: Routledge.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). Textual coherence using latent semantic analysis. *Discourse Processes*, 25(2&3): 285–307.

Godshalk, F. I., Swineford, F. & Coffman, W.E. (1966). The Measurement of Writing Ability. New York, NY. College Entrance Exam Board.

Graesser, A.C., McNamara, D.S., & Kulikowich, J. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*. 40: 223–234.

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

Grosz, B., Joshi, A., & Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational* Linguistics, 21(2): 203–226.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hearst, M. (1997). TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages, *Computational Linguistics*, 23 (1), 33–64.

Higgins, D., Burstein, J., Marcu, D., &. Gentile, C. (2004). Evaluating Multiple Aspects of Coherence in Student Essays. In *Proceedings of 4th Annual Meeting of the Human Language Technology and North American Association for Computation Linguistics,* 185–192, Boston, MA.

Huddleston, E. M. (1952). Measurement of Writing Ability at the College-Entrance Level: Objective vs. Subjective Testing Techniques. ETS RB-52-57.

Kanerva, P., Kristoferson, J., & Holst, A. (2000). Random indexing of text samples for Latent Semantic Analysis. In L. R. Gleitman & A. K. Josh (Eds.), In Proceedings of the 22nd Annual Conference of the Cognitive Science Society, 1036, Mahwah, NJ: Erlbaum.

Kwong, O. Y. (2010). Constructing an Annotated Story Corpus: Some Observations and Issues, In Proceedings of the Language Resources and Evaluation Conference, 2062–2067. Malta.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*. 33:159–174.

Mann ,W.C. & Thompson, S. (1988) Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text*, 8(3), 243–281.

Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization.* Cambridge, MA: The MIT Press.

Miltsakaki, E. & Kukich, K. (2000). Automated evaluation of coherence in student essays. In *Proceedings of the Language Resources and Evaluation Conference*, Athens, Greece.

Pitler, E. & Nenkova, A. (2008). Revisiting Readability: A Unified Framework for Predicting Text Quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 186–195, Edinburgh, Scotland.

Prasad, R., Dinesh, N., A., Lee, Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). The Penn Discourse Treebank 2.0. In *Proceedings of the Language Resources and Evaluation Conference*.

Quinlan. J. R., (1993). C4.5: Programs for machine learning. Morgan Kaufmann Publishers.

Rus, V., & Niraula, N. B. (2012). Automated Detection of Local Coherence in Short Essays Based on Centering Theory, In *Proceedings of CICLING 2012*, IIT Delhi, India.

Schwarm, S. E. & Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 523–530, Ann Arbor, MI.

Sahlgren, M. (2001). Vector based semantic analysis: Representing word meanings based on random labels. In *Proceedings of the ESSLLI 2001 Workshop on Semantic Knowledge Acquisition and Categorisation*. Helsinki, Finland.

Shriver, K. A. (1989). Evaluating text quality: the continuum from text-focused to reader-focused methods. *IEEE Transactions in Professional Communication*, 32(4): 238–255.

Sheehan, K. M., Kostin, I., Futagi, Y. & Flor, M. (2010). Generating Automated Text Complexity Classifications That Are Aligned With Targeted Text Complexity Standards, ETS RR-10-28, Educational Testing Service: Princeton, NJ.

Van den Broek, P. W. (2012). Individual and developmental differences in reading comprehension: Assessing cognitive processes and outcomes. In: Sabatini, J.P., Albro, E.R., O'Reilly, T. (Eds.), *Measuring up: Advances in how we assess reading ability,* 39–58. Lanham: Rowman & Littlefield Education.

Van den Broek, P., Fletcher, C. R., & Risden, K. (1993). Investigations of inferential processes in reading: A theoretical and methodological integration. *Discourse Processes*, 16, 169–180.

Wang, Y., Harrington, M, and White, P. (2012). Detecting Breakdowns in Local Coherence in the Writing of Chinese English Speakers. *The Journal of Computer Assisted Learning.* 28, 396–410.

Wolf, F., & Gibson, E. (2005). Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–288.