Automated Detection of Local Coherence in Short Argumentative Essays Based on Centering Theory

Vasile Rus, Nobal Niraula

Department of Computer Science, The University of Memphis
Memphis, TN, 38152, USA
{vrus, nobal}@memphis.edu

Abstract. We describe in this paper an automated method for assessing local coherence in short argumentative essays. We use ideas from Centering Theory to measure local coherence of essays' paragraphs and compare it to human judgments on one analytical feature of essay quality called Continuity. Paragraphs which correspond to a discourse segment in our work and which are dominated by one prominent concept were deemed locally coherent according to Centering Theory. A dominance measure was proposed based on which local coherence was judged. Results on a corpus of 184 argumentative essays showed promising results. Our findings also suggest that focusing on nominal subject for detecting candidate concepts for a discourse segment's central concept is sufficient, which confirms previous findings. Compared to previous approaches to assessing local discourse coherence in essays, our method is fully automated.

Keywords: coherence, centering, short essay scoring.

1 Introduction

We describe in this paper an automated method for assessing local coherence in short argumentative essays. The method was inspired from an expert analysis of short essays' cohesion and coherence and relies on ideas from Centering Theory [6].

Coherence is an inherent property of text discourse, which is defined as a set of coherent sentences [8]. A related concept, cohesion refers to the explicit cues in the text that link the ideas together [4, 8, 10]. While cohesion defines the texture that keeps a text together (in the sense defined by Halliday and Hassan [7]), coherence defines the overall structure and meaning of the text, i.e. the discourse. In other words, cohesion is the fabric while coherence is the outfit. Obviously, same fabric could lead to very different outfits, some more "coherent" than others. It is important to add that according to a school of thought from Cognitive Science, the coherence, i.e. the outfit in our metaphor, is reader-dependent [4, 10]. That is, different readers could see different outfits depending on their background relative to the text they are reading. It is beyond the purpose of this paper to discuss this point in more detail. On the other hand, the Natural Language Processing (NLP) community adopts a more reader-independent view of coherence (or by default an average reader is assumed; [8]), as explained later. In this paper, we adopt the NLP-community's view of

coherence but not necessarily disagreeing with the other view of reader-dependence coherence.

Following Centering Theory, we distinguish between local and global coherence. Local coherence refers to the coherence of a discourse segment, e.g. a paragraph, while global coherence refers to the overall coherence of an entire text, i.e. an essay in our case.

We focus on short argumentative essays similar to the ones used in the Standard Achievement Test (SAT). In such essays, students are required to take a position (main thesis) relative to the topic/theme described by the essay prompt and argue for it. Arguments must be supported by evidence to be convincing.

Our work can be viewed as being part of the broader context of automatic assessment of written essays. Automatic essay scoring (AES; also known as automatic essay grading or automatic essay evaluation) is the task of automatically assessing student-written essays or, in a broader sense, any written response [2,15,18]. AES holds the hope of providing systematic, timely, and cost-effective solutions to the task of grading essays, which can be very expensive for standardized tests that are taken by hundreds of thousands of students and whose essays must be graded.

It is beyond the scope of this paper to provide an overview of previous work related to automatic scoring of essays. Rather, we focus on discourse level features that could be used in AES tools. Indeed, the discourse structure of essays is an important aspect used often in assessing the essays' quality [1]. To this end, we analyzed the relation between coherence and essay quality as measured by holistic scores.

We started with a corpus of 184 essays which were analyzed along two analytical features, Continuity and Reader Orientation, as well as overall quality. We make the claim the Continuity measures local coherence as defined in Grosz and Sidner's theory of discourse [5] and Grosz, Joshi, and Weinstein's Centering Theory [6]. This paper is a first step towards providing the evidence supporting this claim. Indeed, we show how an automated method based on ideas inspired from Centering Theory can be used to compute essays' local coherence. A centered paragraph would be a paragraph dominated by one central concept which should appear in prominent syntactic roles in the sentences of the paragraph according to Centering Theory. Discourse segments dominated by one center concept are deemed locally coherent.

In our work, a discourse segment corresponds to one paragraph. In other words, we make the assumption that each paragraph contains only one discourse segment. While this is not always the case because a paragraph may contain several discourse segments, each with its own center, our analysis of argumentative essays indicated that students tend to have only one discourse segment per paragraph. That is, there is a tendency to develop only one supporting argument in short essays' body paragraphs. Another argument spawns from the difficulty of detecting discourse segments automatically [19], existing algorithms assuming coherence when detecting the segments which is not a valid assumption in student-written essays.

To decide whether the paragraph is locally coherence we automatically compute a dominance percentage for each concept appearing in subject positions in essays' paragraphs. An analysis of variance (ANOVA) with the dominance percentage of the most dominant concept in a paragraph as the factor was then performed to reveal any significant differences between low and high cohesion.

The rest of the paper is organized as in the followings. The next section provides an overview of related work on coherence and automated methods to capture coherence with a focus on methods developed in the context of automated essay scoring (AES). Then, we provide the conceptual framework behind our basic idea of using Centering Theory to capture the local coherence of short essays. The Experiments and Results section describes our experimental setup and the results obtained. We conclude with Conclusions and Future Work.

Literature Review

According to McNamara and colleagues [10], coherence is the understanding the reader derives from the text while cohesion refers to the explicit cues in text that allow the reader to connect the ideas in it. Coherence is therefore a reader-dependent feature of text influenced by factors such as prior knowledge and reading skills [10].

Researchers in the field of Natural Language Processing (NLP; [9]) define coherence in a more reader-independent way as the goal is to develop automated tools to process discourse, e.g. discourse parsers, without a specific user model in mind or with an implied assumption of an average reader. Typically, the development of automated discourse processing tools is based on journalistic texts such as news articles, which are targeted primarily to average readers. We adopt a similar reader-independent view in this article. For instance, Jurafsky and Martin [9] refer to coherence as the meaning relations between textual units while cohesion is the way in which textual units are linked together. The meaning of the entire discourse can be understood by following the coherence relations among its textual units. Based on this view, NLP researchers defined a set of coherence relations, such as explanation, which are used in the development of automated discourse parsers to discover the structure of discourse [11, 12]. The output of discourse parsers is usually a discourse tree implying a hierarchically structure but more complex structures are possible such as graphs in which a coherent or discourse relation can be observed between any two textual units [14]. It should be noted that linear discourse structures are also used [2].

An important issue in discourse parsing is choosing the set of coherence relations to use. Hierarchical models of discourse distinguish among global coherence, which is revealed through the coherence relations among larger discourse segments shown at higher levels in discourse parse trees, and local coherence, which is observed among utterances within a discourse segment. A discourse segment is more or less coherent depending to the cognitive load it puts on the reader. Discourse segments with a clear focus, i.e. center of attention, are more coherent according to Centering Theory [6].

As mentioned before, our work is conducted in the context of automatic essay scoring. The discourse structure of essays is an important aspect used often in assessing the essays' quality [1]. The complex interactions among coherence, cohesion, and text organization are essential to accurately infer the discourse plan and structure.

It is beyond the scope of this work to analyze in-depth the advantages and disadvantages of AES systems. We rather focus on discourse analysis components used in AES systems. More details regarding automated analysis of essay's discourse

structure are available for E-rater, the AES system developed by Educational Testing Services (ETS; [1, 15]). Shermis and Burstein [15] described in detail E-raters' method for analyzing essays' discourse structure. They assumed a linear discourse structure for essays which are regarded as a sequence of discourse spans each serving one essay-specific communicative goal such as thesis statement, main idea, or supporting idea. It should be noted that E-rater was developed for assisting with scoring the Analytical Writing Assessments of the Graduate Management Admission Test (GMAT). GMAT essay questions are of two types: analysis of an issue and analysis of an argument. E-rater's discourse processing module was developed with the aim of handling both types of essays. Shermis and Burstein [15] did not report about the usefulness of their discourse processing method on scoring essays or providing feedback to individual students, in case the system is used in instructional settings.

The most related work to ours is by Miltsakaki and Kukich [11] who assessed local discourse coherence of essays using a measure of topic continuity, called rough-shift transitions. As opposed to their work, our method is fully automated, e.g. they manually solved the coreferring expressions while we automatically detect them. Furthermore, we propose a different measure of topic continuity called dominance. In addition, we worked with a larger set of essays (184 versus 100).

Our goal here is different from designing a fully-automated method for parsing essays' discourse. Rather, we analyze the relationship between coherence and essay quality as measured by holistic scores and relate the observed patterns of this relationship to elements from Centering Theory.

Motivation

Our work started by analyzing the relation between cohesion, coherence and essay quality as measured by holistic scores. We focus only on coherence and essay quality in this paper. We worked with a corpus of 184 essays which were analyzed manually by two experts along two analytical features, Continuity and Reader Orientation, as well as overall quality. The details of how the corpus was collected and annotated are provided in [3]. Continuity assesses an essay's exhibited strength of connections of ideas and themes within and between the essays' paragraphs. Continuity was meant to measure cohesion according to [3] although they found out that computational indices of cohesion do not correlate well with overall essay quality. On the other hand, Continuity does correlate with overall essay quality (r=0.646). We show here that Continuity measures local coherence.

Reader Orientation represents the essay's overall coherence and ease of understanding. The holistic scale and all of the analytic features had a minimum score of 1 and a maximum score of 6. Table 1 provides an overview of the essay dataset with some statistics.

A closer analysis of the relation between Continuity and Reader Orientation on one hand and overall essay quality, i.e. SAT score, on the other hand, revealed several patterns.

Table 1. The set of essays and their distribution by essay prompt together with some statistics.

Prompt Label	***Prompt***	***Total***	***Number of Words (SD)***	***Number of Types (SD)***	***Number of Paragraphs (SD)***
Creativity	*Some people say that in our modern world, dominated by science, technology, and industrialization, there is no longer a place for dreaming and imagination. What is your opinion?*	59	693.36 (139.47)	239.07 (48.70)	5.37 (1.19)
Religion and Television	*Marx once said that religion was the opium of the masses. If he was alive at the end of the 20th century, he would replace religion with television.*	60	708.37 (140.63)	254.5 (45.93)	5.28 (1.64)
Unequal Men	*In his novel 'Animal Farm', George Orwell wrote "All men are equal: but some are more equal than others". How true is this today?*	65	758.57 (112.96)	258.06 (44.29)	5.35 (1.37)

The two measures of Continuity (C; average=3.57, stdev=1.24) and Reader Orientation (RO; average=3.96, stdev=0.90) correlate strongly with the holistic SAT scores (average=3.53, stdev=1.07), the latter correlating stronger (r=.786 for RO; r=0.646 for C). RO tends to be higher than SAT scores for low and medium quality essays (as judged by holistic SAT scores). Continuity scores tend to be closer to SAT holistic scores with a higher degree of variation for medium quality essays. For extremely high-quality essays (i.e., those essays scored 5.5 or 6) the three measures (C, RO, and SAT) tend to converge.

Probably the most interesting emerging pattern is the fact that while RO scores tend to follow the SAT scores closely, the C scores fluctuate around the SAT scores. Indeed, our data shows that essays can have low or high Continuity scores while still being overall high or low quality essays. That is, at almost all levels of SAT scores, except the very extremes (SAT scores of 1, 1.5 and 5.5, 6) Continuity scores group in two very different categories, low and high scores, forming a binary pattern. This pattern can be seen in Figure 1 where we plotted the scores for the three measures with the essays on X-axis being ordered based on the holistic SAT scores.

In order to characterize in more depth this difference in which Continuity (C) and Reader Orientation (RO) scores follow SAT scores, we analyzed essays that had extreme values for C and RO. In particular, we looked at four possible combinations of C and RO scores: (low-C, low-RO), (low-C, high-RO), (high-C, low-RO), and (high-C, high-RO).

The selected essays were analyzed manually in terms of local coherence at paragraph level. We show in Figure 2 two paragraphs from two different essays that have low and high Continuity scores, respectively, while both having high Reader Orientation scores. The two essays are of high quality as their SAT scores are 5.5 and 4.5, respectively. These examples, which differ significantly in their Continuity scores, display different interrelations among the main theme and the supportive arguments. The high Continuity paragraph at the top of Figure 2 has a “thin” main theme, *dreaming and imagination*, which mentioned only in the opening and closing sentences of the paragraph. The paragraph focuses then on the supporting argument that is being developed and forms the central concept of the paragraph.

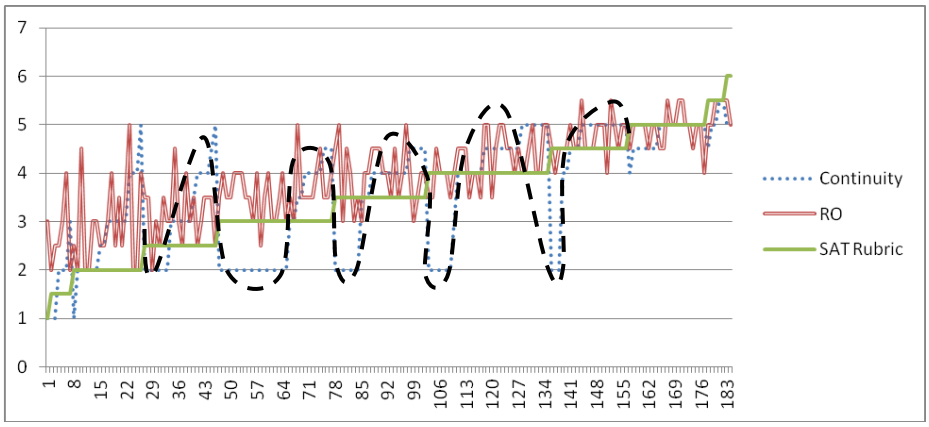

Fig. 1. Continuity, Reader Orientation, and SAT scores for our MSU set. Essays on the X-axis are ordered based on their holistic SAT score. The binary pattern for Continuity is shown with the bold black line. The bold, flat line indicates a floor-effect for Continuity-scores for medium quality essays.

The modern workers are separated from the land and their tools and their skills of craftsmanship are rendered useless.
They have become wage slaves as their working hours have constantly grown longer.
They are forced to accept whatever conditions their employer may impose on them and are subjected to physical and intellectual examinations everyday.
…
All the while the employers are writing themselves multimillion dollar checks and only handing out two and three percent raises to the workers when the economy is good.
…
Unlike worker of the past who dreamed of working their own land, the modern worker continues to believe they are indispensable to their employer and often times work so hard and receive so little in return.

By definition science is a branch of knowledge involving systematized observation and experimentation.
Without dreams and imagination, the ideas for these systematic observations and experimentations would not have a bases to take root from.
Scientists are humans, and human nature is to dream and imagine things.
…
One of these ideas was from the imagination of a scientist who found the medicine penicillin.
Dreams and imagination are the foundation on which the reality of science is actually based upon.

Fig. 2. Examples of two essay paragraphs one with good local coherence (top; Continuity score of 5.5) and one with poor local coherence (Continuity score of 2).

On the other hand, in the low Continuity paragraph, shown at the bottom of Figure 2, the main theme is referred to throughout the entire paragraph, being present even in the middle of the essay's body paragraphs. The student writer comes back to the main

theme, *dreaming and imagination*, every other sentence, in this example paragraph. The paragraphs seems to lack focus, that is, is not centered on a single concepts and therefore having low local coherence according to Centering Theory.

Our conclusion from the qualitative analysis of paragraphs in essays with extreme values for Continuity and Reader Orientation is that the essay body paragraphs' discourse flow must focus on the supporting arguments with as few references to the main theme as possible. To test this hypothesis we will propose a method based on Centering Theory. The idea is to automatically detect the centers of paragraphs in essays. Paragraphs that are not focused (i.e. have more than one center or focus) will lack local coherence according to Centering Theory and expected to have corresponding low Continuity scores assigned by human judges.

Centering-based Approach To Detecting Local Coherence

As already mentioned, our basic approach to capturing local coherence in essays is to use ideas from Grosz and Sidner's theory of discourse [5] and Grosz, Joshi, and Weinstein's Centering Theory [6]. Grosz and Sidner view discourse as the result of three interrelated components: linguistic structure, intentional structure, and attentional state. The linguistic structure refers to the discourse segments into which utterances naturally group. The intentional structure captures speaker's intentions expressed as discourse purposes and their relationships. The intentions provide the basic rationale for the discourse. The attentional state models the discourse participants' focus of attention at any given moment. Centering Theory links the attentional state and perceived coherence of discourse segments, i.e. local coherence. Discourse is more coherent if it limits the number of inferences necessary to understand its utterances. More locally, discourse segments are perceived more coherent if the focus of attention within the segment is carefully maintained by the use of appropriate linguistic devices, e.g. placing the entity that is the discourse focus in prominent syntactic positions, e.g. subject, in utterances. In our work, we make the assumption that a discourse segment corresponds to one paragraph. While this is not always the case because a paragraph may contain several discourse segments in some cases, each with its own center, our analysis of essays indicated that students tend to have only one discourse segment per paragraph.

To illustrate the relation between centering and local coherence we use the two paragraphs in Figure 3 (from [6]). The top paragraph is coherent because it has one center, i.e. *John*, appearing in subject positions of every sentence in the paragraph. The bottom paragraph is less coherent as it lacks centering on one concept – both John and the store appear in subject positions of paragraph's sentences.

In our case, paragraphs with high Continuity are characterized by high local coherence while low Continuity paragraphs have poor local coherence (i.e. their Continuity is broken by too many references to the main theme of the essay at body-paragraph level). From a Centering Theory point of view, the former paragraphs lack a clear focus at discourse segment level as the center of attention switches back and forth between the discourse segment purpose (DSP, i.e. the supporting argument being developed in the segment) and the overall discourse purpose (DP, i.e. the main

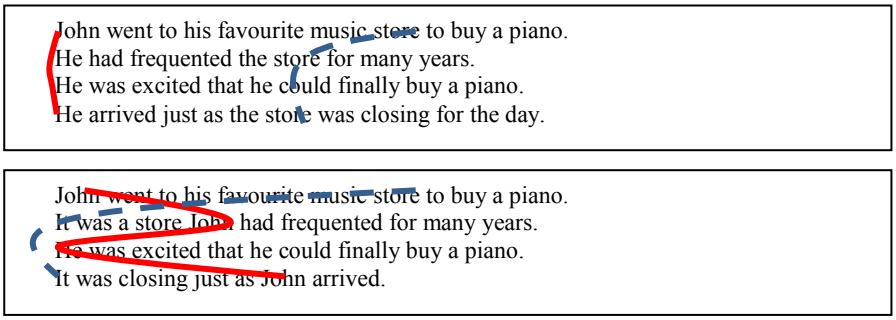

Fig. 3. Examples of two discourse segments presenting the same information in different ways (from [6]). The top paragraph is coherent while the bottom one is less coherent according to Centering Theory as it leads to higher inference loads.

theme of the essay). Examples of paragraphs with high and low Continuity scores are shown in Figure 2. The top paragraph in the figure has a Continuity score of 5.5 (on a 1-6 scale) while the bottom one has a Continuity score of 2. A centering analysis of the two paragraphs (a center is considered a word appearing in the main subject position of each sentence) reveals that the high Continuity paragraph at the top of the figure has a clear center, i.e. *workers*. The low Continuity paragraph switches between two centers *science-ideas-scientists* to *dreams and imagination*.

Expert raters seem to be generous when grading essays with respect to local coherence as long as the intentions of the speaker are clear (good global coherence) and the essay are contents rich (enough information expressed with a sizeable vocabulary). Such expert raters seem to be able to untangle the intertwined ideas inside discourse segments with poor local coherence without much effort and therefore do not see the need to penalize the student-writers significantly.

To distinguish the two types of discourse structures, we advance the idea of automatically identifying the degree of centering in essay's paragraphs. Centers are concepts appearing in prominent syntactic roles, e.g. subjects, of sentences. We considered several instances of this basic idea. In one instance, we consider concepts occurring only in subject positions of the main clause. In another instance, we consider concepts occurring in subject roles in any clause. Furthermore, another parameter that leads to two other instances is whether to consider the opening and closing sentences of paragraphs which in essay's body paragraphs usually refer to the main theme of the essay and not necessarily to the supporting argument which is the center of the paragraph.

Besides the mentioned benefit to AES system developers, the proposed method to assess local coherence of essays could inform the development of objective criteria to be incorporated in scoring rubrics which in turn can be used to train human raters.

Furthermore, the proposed ideas could be used in instructional settings to teach students to write quality essays that are locally and globally coherent.

3 Experimental Setup and Results

This section presents the experiments we conducted to study the relationship between local coherence, as defined by Centering Theory, and human measurements of essay quality such as Continuity. The basic idea is to see if paragraphs that are centered, i.e. have one central concept, do correspond to high scores of Continuity and vice versa. We present results for several instances of this basic idea as explain next.

First, we will vary the way we consider candidate centers. Given a paragraph, centers can be considered either words in the subject position of the main verb of the main clause of a sentence or words that are subjects in any clause of a sentence. Second, instead of selecting centers as individual words we can select centers as sets of related words. For instance, if the words *couple* and *parents* appear in subject positions in two consecutive sentences they more or less refer to the same center (for space reasons we do not show the actual paragraph from which these example words were chosen). In such cases, we propose to use word-to-word similarity measures to decide which candidate centers should be collated together.

To illustrate the details of our basic approach, we will use the example paragraph at the top of Figure 3. Given such a paragraph, we detect first the words that are nominal subjects, i.e. the subjects of the main verbs of each sentence, using a dependency parser and then generate a center-matrix – see Table 2 – in which there is one column for each nominal subject and one row for each sentence in the paragraph.

Table 2. Example of center-matrix for the top paragraph in Figure 3.

	Workers	*They*	*Worker*	*Employers*
Sentence 1	1	0	0	0
Sentence 2	0	1	0	0
Sentence 3	0	1	0	0
Sentence 4	0	0	0	1
Sentence 5	0	0	1	0

An ideal center-matrix corresponding to an ideally centered paragraph would have a column filled with 1s or something close to that. In our example, the presence of pronouns or related words referring to the same center (coreferrents) complicates the derivation of the matrix calling for a more general approach. Such an approach would link a center to its pronominal and also nominal referents. The more general approach would collate together different columns in the matrix that refer to the same center. The center-matrix obtained initially from the example paragraph should be transformed into a simplified matrix where columns referring to the dominant center, *workers/they/worker* in our case, are collated together in a single column (see Table 3). We initially used the BART coreference resolution system [20] to collate columns together. The results were not close to our expectations which determined us to use semantic similarity between words. That is, for each two columns in the center-matrix we compute a similarity score using Latent Semantic Analysis (LSA; [9]). If the similarity is above a certain threshold (.30 – determined empirically on a subset of the essay corpus), the corresponding columns are collated.

Table 3. Example of collated center-matrix for the top paragraph in Figure 3. The columns corresponding to *workers*, *worker*, and *they* in Table 2 were collated together.

	Workers	*Employers*
Sentence 1	1	0
Sentence 2	1	0
Sentence 3	1	0
Sentence 4	0	1
Sentence 5	1	0

Given the collated matrix, for each new center, i.e. column, we compute a dominance score which is the percentage of sentences in the paragraph in which it occurs as subject. To compute the percentages, we normalize by the largest paragraph in our collection to avoid bias towards short paragraphs. The bias consists of short paragraphs, say of two sentences, resulting in high dominance percentages even though the center occurs only in one sentence (out of two).

We present results with several variations of this basic idea. We use nominal subjects versus all subjects and no-collation versus collation of columns.

Results. To evaluate our method to automatically detect local coherence in short essays, we compared the output of the proposed method with human judgments of Continuity. We selected a subset of our original essay corpus to test our ideas. This was necessary as essay that are extremely poor may contain essays with a single very long paragraph or many very short paragraphs (<=2 sentences). Because our focus is on body paragraphs, which means paragraphs in the main body of the essay besides the introductory and conclusion paragraphs, and paragraphs which have more than the introductory and concluding sentences, i.e. having more than two sentences, we dropped in our analysis those essays that do not meet these criteria. This way, we were left with 171 essays out of the 184 original essays. Each of the remaining essays was split into individual paragraphs and for each paragraph several center-matrices were generated, one for each of the variants of the basic idea. From the matrices, we derived dominance percentages for each of the candidate centers, i.e. columns.

Once the dominance percentage for each paragraph has been obtained, we conducted an one-way analysis of variance (ANOVA) with the local coherence as the factor and human judgments of Continuity as the grouping variable. Continuity scores were used to map essay into three groups: group A included essays with Continuity scores 1-2 which correspond to low local coherence, group B included essays with Continuity scores 3-4 (intermediate local coherence), and group C included essays with Continuity scores 5-6 (high local coherence). The one-way analysis of variance would tell us whether dominance percentages, which capture centering of paragraphs, are statistically different among the three levels of local coherence (low, intermediate, and high). This in turn would inform us whether dominance percentages can be used as a predictor of local coherence. Table 4 provides a summary of the results of the analysis for the four instances of our basic method. Across all methods, differences between groups were noticed primarily for essays levels with average overall essay scores (SAT scores of 3, 3.5 and 4 which included 87 essays in total). From the table, we can see that the statistically significant (at $p<.05$ level) method is the one using

Table 4. Overall results of the automated method for local coherence and its variants.

Method	*F(85, 2)*	*Significance (p-value)*
NominalSubjects	2.096	0.129
NominalSubjects+Collation	3.126	0.014
AllSubjects	1.877	0.227
AllSubjects+Collation	2.347	0.149

only nominal subjects with collation based on semantic similarity between subject words. This result in favor of nominal subject provides support for the current view that an utterance in Centering Theory corresponds to a sentence as opposed to certain types of clauses (see [12]). The ANOVA only tells us that at least two of the groups are significantly different but not which ones. A post-hoc analysis showed that the only significantly different groups were the low and high local cohesion groups (i.e. groups A and C in our notation above). That is, the ANOVA and post-hoc analysis revealed that centering could differentiate between average essays with low and high local coherence for the body paragraphs. We have not used this finding further in a prediction model for local coherence as the subset of 87 essays for which the differences were observed to be significant was not large enough for a training-test split. We plan to address this issue in the future by collecting more essays. Furthermore, we plan to improve the way we collate centers together by addressing issues such as pronoun resolution which word-to-word similarity based collation of centers does not address. An alternative would be to use a high-quality coreference resolution engine in which case there is no need to rely on word-to-word similarity.

4 Conclusions

We have explored in this paper the relation between local coherence and human judgments of local coherence called Continuity. We designed an automated method to characterize essays with low, medium, and high Continuity. An analysis of variance analysis revealed that indeed our proposed measure of local coherence is statistically different across the three levels of local coherence with the the differences being significant only for average essays (holistic SAT scores of 3, 3.5, and 4).

Our proposed methods for measuring automatically the local coherence of essays' discourse offers a new instrument which could be integrated in automated essay scoring (AES) systems to better score and provide feedback during writing assessment and instruction. We plan to automate the proposed thread-based analysis, validate it on large corpus of essays, and integrate it in our writing strategy training tutoring system.

Acknowledgments. This research was supported in part by Institute for Education Sciences under awards R305A100875, and R305A080589. Any opinions, findings, and conclusions or recommendations expressed in this material are solely the authors' and do not necessarily reflect the views of the sponsoring agencies.

References

1. Burstein, J., Kukich, K., Wolff, S., Lu, C, Chodorow, M., Braden-Harder, L., and Harris, M.D. (1998). Automated Scoring Using A Hybrid Feature Identification Technique, in Proceedings of ACL, 1998, 206-210.
2. Burstein, J., Marcu, D., and Knight, K. (2003). Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. IEEE Intelligent Systems, pp. 32-39, Jan/Feb, 2003.
3. Crossley, S. A., & McNamara, D. S. (2010). Cohesion, Coherence, and Expert Evaluations of Writing Proficiency. *Proceedings of the 32nd annual conference of the Cognitive Science Society.*
4. Graesser, A.C., McNamara,D.S.,& Louwerse,M.M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text. In A.P. Sweet and C.E. Snow (Eds.), Rethinking reading comprehension (pp. 82–98). New York: Guilford Publications.
5. Grosz, Barbara J. and Candace L. Sidner. (1986). Attention**,** intentions, and the structure of discourse**.** Computational Linguistics 12(3): 175-204.
6. Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):202-225.
7. Halliday, M.A.K. and Hasan, R. 1976. Cohesion in English London: Longman.
8. Jurafsky, D. & Martin, J.H. (2009). Speech and Language Processing, Prentice Hall, ISBN: 0131873210.
9. Landauer, T., McNamara, D., Dennis, S., and Kintsch, W. (eds). (2004). Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum, 1998.
10. McNamara, D. S., Kintsch, E., Butler-Songer, N., and Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. Cognition and Instruction , 14 (1), 1-43.
11. Miltsakaki, E. and Kukich, K. (2004). Evaluation of text coherence for electronic essay scoring systems, In: *Natural Language Engineering* 10:1, 2004.
12. Miltsakaki, E. (1999). Dissociating discourse salience from information structure: Evidence from a centering study in Modern Greek and Japanese. In *Computational Linguistics in the Netherelands, CLIN '99*.
13. Mann, W.C. and Thompson, S. (1988). Rhetorical Structure Theory: Toward a Functional Theory of Text Organization, Text 8(3). Pp. 243-281.
14. Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). The Penn Discourse Treebank 2.0, Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)
15. Shermis, M. D. & Burstein, J. (2003). Automated Essay Scoring: A Cross Disciplinary Perspective. Mahwah, NJ: Lawrence Erlbaum Associates.
16. Wolf, F. and Gibson, E. (2005). Representing discourse coherence: a corpus-based study. *Computational Linguistics* 31, pp. 249–287.
17. Jill Burstein, Daniel Marcu, and Kevin Knight (2003). Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. IEEE Intelligent Systems, pp. 32-39, Jan/Feb, 2003.
18. Page, E. B. (1966). The imminence of grading essays by computer. Phi Delta Kappan 48, 238–243.
19. Passonneau, R. and Litman, D. (1997). Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139.
20. Versley, Y., Ponzetto, S.P., Poesio, M., Eidelman, V., Jern, A., Smith, J., Yang, X., Moschitti, A. (2008) BART: A Modular Toolkit for Coreference Resolution. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.