
Computers and the Humanities **37:** 455–467, 2003.

Notes and Discussion

A Machine Learning Approach for Identification of Thesis and Conclusion Statements in Student Essays

JILL BURSTEIN[1] and DANIEL MARCU[2]
[1]*Educational Testing Service, Princeton, NJ 08541, USA*
E-mail: Jburstein@ets.org
[2]*University of Southern California/Information Sciences Institute, 4676 Admiralty Way, Suite 1001, Marina del Rey, CA 90292, USA*
E-mail: marcu@isi.edu

Abstract. This study describes and evaluates two essay-based discourse analysis systems that identify thesis and conclusion statements from student essays written on six different essay topics. Essays used to train and evaluate the systems were annotated by two human judges, according to a discourse annotation protocol. Using a machine learning approach, a number of discourse-related features were automatically extracted from a set of annotated training data. Using these features, two discourse analysis models were built using C5.0 with boosting: a topic-dependent and a topic-independent model. Both systems outperformed a positional algorithm. While the topic-dependent system showed somewhat higher performance, the topic-independent system showed similar results, indicating that a system can generalize to unseen data – that is, essay responses on topics that the system has not seen in training.

Key words: discourse analysis, discourse annotation, essay evaluation, machine learning, text classification

1. Introduction: Motivation for Automated Discourse Analysis

Software for automated evaluation of student essays has become a prevalent technology over the past few years. Many colleges, universities, public school districts, and language testing organizations use automated essay scoring technologies to provide grades to student essays (Burstein, 2003; Elliott, 2003; Landauer *et al.*, 2003; Larkey and Croft, 2003; Page, 2003).

As educators became more comfortable with automated essay scoring technology, they also gained an awareness about the need for more comprehensive analyses of student writing. For example, they were interested in the evaluation of grammar error detection in essays (Leacock and Chodorow, 2003). They also had a strong interest in automated analysis of the essay-based discourse features (Burstein *et al.*, 2003; Burstein and Marcu, 2003).

The literature in the teaching of writing suggests that invention, arrangement and revision in essay writing must be developed in order to produce effective writing. Stated in practical terms, students at all levels, elementary school through post-secondary education, can benefit from practice applications that give them an opportunity to work on discourse structure in essay writing. Teacher's feedback about students' writing is often expressed in general terms which is of little help; to be useful, the feedback must be grounded and must refer to the specific text of the essay (Scardamalia and Bereiter, 1985; White, 1994). If a system can automatically identify the actual text associated with discourse elements in student essays, then feedback like that used in traditional, textbook teaching of writing can be directed toward specific text segments in students writing. These kinds of questions are often used in textbooks to encourage students to reflect on the organizational components in their writing: a) *Is the intention of my thesis statement clear*? b) *Does my thesis statement respond directly to the essay question*? c) *Are the main points in my essay clearly stated*? and d) *Does my conclusion relate to my original thesis statement*? If these questions were presented, along with specific text segments from students' essays, this would help students think about specific parts of their essay.

This study builds on previous work that reports on the identification of a single sentence associated with the thesis statement text segment, using Bayesian classification (Burstein *et al*., 2001). It relates specifically to system performance with regard to a system's recognition of the possible multiple text segments corresponding to *thesis* and *conclusion* text segments in student writing. A machine learning decision tree algorithm, C5.0 with boosting, was used for model building and labeling. The results indicate that the system can automatically identify features in student writing and can be used to identify thesis and conclusion statements in student essays.

In this article, we address the following questions: 1) Can a system be built that reliably identifies thesis and conclusion statements?, 2) Moreover, how does system performance compare to a baseline, and inter-annotator agreement between human judges?, 3) Will the system be able to generalize across genre and grade-level to some extent?, and 4) How well does the system generalize to unseen essay responses? That is, can the system identify thesis and conclusion statements on essay topics that it has not been trained on?

2. Topic Selection and Manual Annotation

In order to answer the questions presented in the Introduction, essay test topics (test questions) were selected across two genres and two populations of students (with regard to grade level.) Human judges annotated essay samples. Annotation was done both for purposes of system training, and evaluations, so that the system performance could be compared to the agreement of two human judges.

"You can't always do what you want to do!," my mother said. She scolded me for doing what I thought was best for me. It is very difficult to do something that I do not want to do. **<Thesis>**But now that I am mature enough to take responsibility for my actions, I understand that many times in our lives we have to do what we should do. However, making important decisions, like determining your goal for the future, should be something that you want to do and enjoy doing**. </Thesis>**

I've seen many successful people who are doctors, artists, teachers, designers, etc. In my opinion they were considered successful people because they were able to find what they enjoy doing and worked hard for it**.** It is easy to determine that he/she is successful, not because it's what others think, but because he/she have succeed in what he/she wanted to do.

In Korea, where I grew up, many parents seem to push their children into being doctors, lawyers, engineer etc. Parents believe that their kids should become what they believe is right for them, but most kids have their own choice and often doesn't choose the same career as their parent's**.** I've seen a doctor who wasn't happy at all with her job because she thought that becoming doctor is what she should do. That person later had to switch her job to what she really wanted to do since she was a little girl, which was teaching.

<Conclusion> Parents might know what's best for their own children in daily base, but deciding a long term goal for them should be one's own decision of what he/she likes to do and want to do. **</Conclusion>**

Figure 1. An essay with human judge annotations for thesis and conclusion.

2.1. ABOUT THE TOPICS

In this study, we used six topics from two different writing genres: five of the topics were *persuasive*, and one was *informative*. These are defined as follows in the context of school-based writing instruction. Persuasive writing requires the writer to state an opinion on a particular topic, and to support the stated opinion to convince the reader that the perspective is valid and well-supported. An informative writing task requires the writer to state their opinion on a particular topic. This is typically more personal and descriptive writing. Four of the five sets of persuasive essay responses were written by college freshman (A, B, G, and H), and the fifth by 12th graders (N). The sixth, informative set of essay responses, was also written by 12th graders.

2.2. DESCRIPTIONS OF DISCOURSE CATEGORIES

Two human judges were trained to label several discourse elements according to a protocol designed by the authors and four writing experts. Below are descriptions of the discourse categories. Both *thesis* and *conclusion* statements can contain one or more sentences. An essay annotated by a human judge for thesis and conclusion statements is in Figure 1.

Thesis. The "Thesis" represents the text segment that contains the writer's position statement, and has a direct relationship with the essay topic. Conclusion. The "Conclusion" is the main idea that summarizes the entire argument developed by a student in an essay. The conclusion may contain new information, such as 'provocative or profound' thoughts that reflect the writer's position – in an attempt to leave the reader with something to think about.

2.2.1. *Annotator Training*

The judges were instructed to assign one label to each sentence. Pre-training of the judges was done on an initial set of essays from the three different essay questions that the judges would be annotating.[1] During this phase, the authors and the judges discussed, and labeled together approximately 50 essays, across the three topics. During the next training phase, the judges labeled an additional set of approximately 100 essays on each of the three topics. Kappa statistics were run on their independent judgements every hour, and if the kappa for any particular category fell below 0.8, then the judges were asked to review the protocol until their agreement was acceptable. In the next phase, annotation (post-training) began, and the judges did not discuss their labeling decisions. In this post-training phase, judges annotated independent data sets for three different topics. There were approximately 40 overlapping essays in each of these three data sets. Agreement between the two judges is reported in the following section based on the overlapping essays. Kappa, Precision, Recall, and F-measures are reported. Approximately 360 essays (including the 40 essays in common per topic) were annotated for these three topics (A, B, and C). For three additional topics (G, H, and N), approximately 300 essays were annotated by two judges. For these topics, each judge had a unique set of essays. Annotations from all six topics were used in the experiment described in a later section.

2.2.1.1. *Human judge agreement*

In order to build a system that can automatically identify discourse elements in student essays, we first have to be certain that humans can do this task reliably. It is critical that the annotation process yields agreement that is high enough between human judges, such that it suggests that people can agree on how to categorize the discourse elements. As is stated in the above section, during the training of the judges for this study, Kappa statistics were computed on a regular basis. Kappa between the judges for each category had to be maintained at least 0.8, since this is believed to represent strong agreement (Krippendorff, 1980). The agreement statistics shown in Table I indicate that agreement between human judges was high for both the Thesis and Conclusions discourse categories. The results are based on approximately 40 essays for three topics. These 40 essays were annotated independently by both judges, without discussion.

3. Essay-Feature Discourse Analyzer

The model built to assign thesis and conclusion labels to sentences in essays is based on a training sample of approximately 1200 essays: 200 essays from each topic. All responses were manually annotated for thesis and conclusion statements. For model building, discourse-relevant features in an essay were extracted from each sentence. Each model is built using these features as input to C5.0 with

Table I. Inter-annotator agreement between 2 Human Judges for Thesis and Conclusion Statements

Topic	Discourse elements							
	Thesis				Conclusion			
	K	P	R	F	K	P	R	F
A	0.92	0.97	0.89	0.93	1.00	1.00	1.00	1.00
B	0.77	0.82	0.78	0.80	0.90	0.91	0.92	0.91
C	0.94	0.92	0.99	0.96	0.77	0.82	0.78	0.80

boosting.[2] The following features were used for model building, and subsequent label assignment on unseen essay data in the test sets.

3.1. SENTENCE AND PARAGRAPH POSITION

Four features relevant to sentence and paragraph position were used. Three were continuous attributes, and the fourth was a discrete attribute: a) the sentence number within the essay, b) the sentence number within its paragraph, c) the paragraph number in which the sentence occurs, and d) the relative position of the paragraph in which the sentence occurs (i.e., first paragraph, body paragraph, and final paragraph).

3.2. RST RHETORICAL RELATIONS AND STATUS

RST rhetorical relations and status are assigned to sentences from an existing discourse parser (Marcu, 2000). According to RST (Mann and Thompson, 1988), one can associate a rhetorical structure tree to any text. The leaves of the tree correspond to elementary discourse units and the internal nodes correspond to contiguous text spans. Each node in a tree is characterized by a *status* (nucleus or satellite) and *a rhetorical relation*, which is a relation that holds between two non-overlapping text spans. The distinction between nuclei and satellites comes from the empirical observation that the nucleus expresses what is more essential to the writer's intention than the satellite; and that the nucleus of a rhetorical relation is comprehensible independent of the satellite, but not vice versa. When spans are equally important, the relation is multinuclear. Rhetorical relations reflect semantic, intentional, and textual relations that hold between text spans as is illustrated in Figure 2. For example, one text span may elaborate on another text span; the information in two text spans may be in contrast; and the information in one text span may provide background for the information presented in another text span. Figure 2 displays in the style of Mann and Thompson (1988) the rhetorical structure tree of a text

fragment. In Figure 2, nuclei are represented using straight lines; satellites using arcs. Internal nodes are labeled with rhetorical relation names.

We built RST trees automatically for each essay using the cue-phrase-based discourse parser of Marcu (2000). We then associated with each sentence in an essay a feature that reflected the status of its parent node (nucleus or satellite), and another feature that reflected its rhetorical relation. For example, for the last sentence in Figure 2 we associated the status satellite and the relation elaboration because that sentence is the satellite of an elaboration relation. For sentence 1, we associated the status nucleus and the relation elaboration because that sentence is the nucleus of an elaboration relation.

3.3. CUE TERM DISCOURSE FEATURES

A discourse analysis module identifies cue words, terms, and syntactic structures that function as discourse markers. Earlier research has indicated that these discourse features have been shown to predict the human-reader-assigned essay grades, and can be associated with *organization of ideas* in an essay (see Burstein *et al*., 1998a, b; Burstein, 2002). Accordingly, the labeled units of discourse that are output from this module also appear to be related to particular discourse elements in essays, such as Thesis and Conclusion statements.

The module contains a lexicon that is based on the conceptual framework of conjunctive relations from Quirk *et al*. (1985). For instance, in this framework, cue terms, such as *In summary*, and *In conclusion*, are classified as conjuncts that are associated with the discourse function of "summarizing" an argument. The conjunct classifiers may contain information about whether or not the item is a kind of discourse development term. For example, the word *because* further develops the idea in the writer's initial statement that "people should travel to new places" in the sentence, "*I think that people should travel to new places because it enhances their perspective*." Alternatively, a classifier may indicate that a cue word is used to initiate an argument. For example, for the word *first* to be considered as a discourse marker (indicating a parallel relation) it *must not* be a nominal modifier, as in the sentence, "*The first time I went to Europe was in 1982*," in which *first* modifies the noun "*time*." Instead, *first* must occur as an adverbial conjunct to be considered a discourse marker, as in the sentence, "*First, I think that people should travel to new places*." Syntactic structures, such as infinitive clauses, are also used to identify the beginning of a new argument based on the position of the clause within a sentence, along with the position of the sentence within a paragraph. For instance, infinitive clauses that begin sentences, and are also toward the beginning of a paragraph are more often indicators of the beginning of a new argument.

While other discourse analyzers indicate hierarchical discourse relationships in text (Marcu, 2000), the output of this module does not. The discourse analysis module produces a flat, linear sequence of labeled units. For instance, in the essay

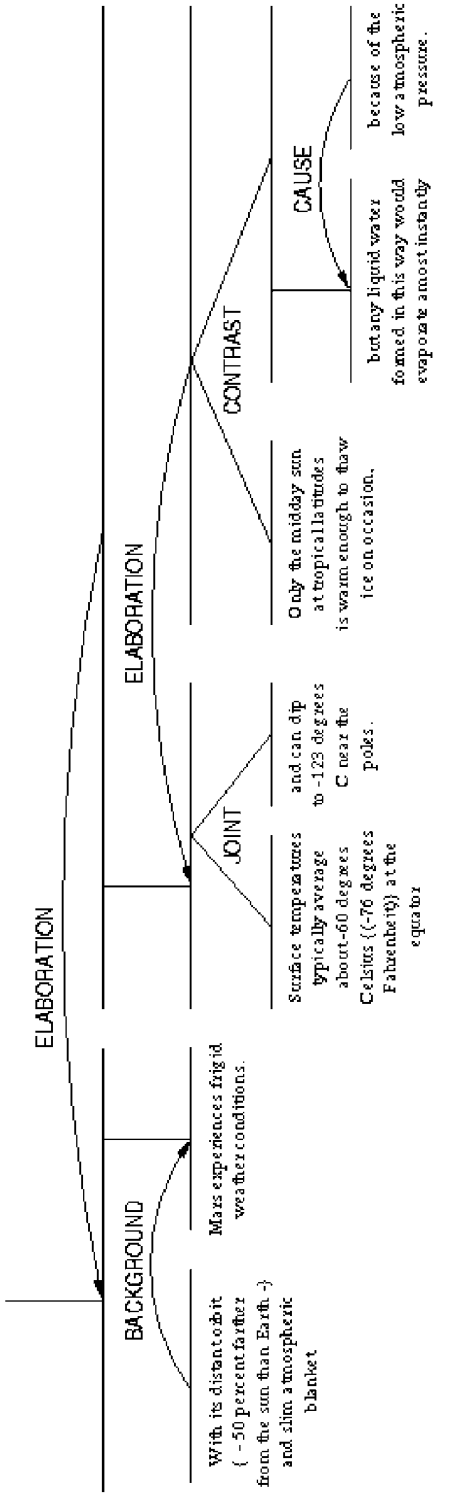

Figure 2. Example of an RST tree.

text, the word *however* may indicate that a *contrast* relationship exists; though, it does not show the related contrasting text segments.

3.4. Lexical items for general essay & category-specific language

Through empirical analysis of the training data from this study, and previous work (see Burstein *et al*., 2001; Burstein and Marcu, 2002), we found that there were particular words and terms that were characteristic of a general essay sublanguage, and a sublanguage related to some essay-based discourse categories. For example, lexical items related to general language in essays included words such as, *should*, *might*, *agree*, *disagree*, and *I*. Lexical items such as *opinion* and *feel* can be associated with the Thesis statement, while the term *In conclusion* is clearly associated with the essay Conclusion category. Words and terms associated with the general essay and category-specific language were used as features.

4. Experiment

The results reported in this study are based on seven data sets. The system was trained using manually annotated data from these seven sets. In one of the seven data sets, essay responses from all six topics were included (ALL) (topic-dependent system). Results are reported for each topic-dependent test set. For the remaining six sets, only five topics were included in training, and the 6th topic was held out for testing (topic-independent system). These six additional runs were PIA, PIB, PIC, PIG, PIH, and PIN, where the final letter indicates the test topic, and data from all other topics was used for training. The same test set was used in both the topic-dependent and topic-independent systems. Both systems were built so that generalizability of essay-based discourse labeling could be evaluated. In other words, if the topic-independent system performance outperformed the positional baseline, and at least approximated the performance of the topic-dependent system, this would suggest that a topic-independent system could be used.

All seven of the training sets contained approximately 1200 essays. For the ALL data set, the test set contained a sample from each of the six topics. The test set for ALL contained approximately 300 essays: 50 essays per topic. These same topic subsets were used to evaluate the topic-independent system.

4.1. Positional algorithm

Essay length is highly correlated with human or machine scores (i.e., the longer the essay, the higher the score). Similarly, the position of the text in an essay is highly related to particular discourse elements. Therefore, we computed a positional label for the thesis and conclusion discourse categories. The method outlined in Table II was used for computing baselines reported in a later section.

Table II. Method for computing positional baselines

Number of paragraphs (P) in essay	Discourse label	
	Thesis	Conclusion
3 or more	All text in P 1, excluding the 1st sentence.	All text in final P.
2 or more	Select all text in the first P.	Select all text in final P.
1	Select nothing.	Select nothing.

4.2. RESULTS

Table III shows average results for all three systems: positional, topic-dependent and topic-independent. For the ALL data, the training and test sets contain sample responses from all six topics. For the topic-independent data, Table III shows the average agreement across all runs illustrated explicitly in Table IV, below. The training sets from the topic-independent system did not contain essay responses from the test set. Agreement with a single human judge shows that for all measures of agreement, both discourse-based systems outperform the positional algorithm for all cases of thesis and conclusion identification. The systems' performance is in between baseline system, and human inter-annotator agreement (see Table I). As might be expected, the topic-dependent system outperforms the topic-independent system in the majority of cases. The results of the topic-independent system approximate the topic-dependent system. Results are more comparable between the topic-dependent and topic-independent systems for conclusion statement assignments, than for thesis statements.

In Table IV, we see that for the topic-specific results, both discourse-based systems outperform the positional baselines, with the exception of the topic-independent run, PIC, for thesis statement assignment. Topic-dependent and topic-independent results are generally comparable for across topic-specific runs. With regard to precision, the topic-independent system ranges from a 1 to 5 percent decrease in assignment of thesis statement labels, and a 1 to 3 percent drop in conclusion statement labeling.

5. Discussion and Conclusions

The study shows that a machine learning approach outperforms a positional algorithm for automatically identifying *thesis* and *conclusion* statements in student essays. Since the planned use for this kind of system is for classroom writing instruction, results are discussed in terms of precision. In instructional applications, higher precision is preferable, since this means that the student will be presented with the most reliable feedback. Discussion of performance, therefore, refers to precision values.

Table III. Average agreement between 1 human judge and systems: Precision, Recall, and F-measures

System	Thesis			Conclusion		
	P	R	F	P	R	F
Positional	0.43	0.63	0.51	0.51	0.71	0.60
Topic-Dependent	0.56	0.69	0.62	0.75	0.88	0.81
Topic-Independent	0.52	0.58	0.54	0.74	0.83	0.80

Table IV. Agreement between 1 human judge and systems (positional = POS; Topic-Dependent = TD; Topic-Independent = TI), evaluating individual topic subsets: Precision, Recall, and F-measures

System	Thesis			Conclusion		
	P	R	F	P	R	F
POS						
A	0.58	0.69	0.63	0.62	0.69	0.65
B	0.43	0.61	0.50	0.49	0.75	0.59
C	0.54	0.71	0.61	0.83	0.75	0.79
G	0.37	0.68	0.48	0.34	0.70	0.46
H	0.26	0.64	0.37	0.34	0.67	0.45
N	0.25	0.29	0.27	0.33	0.62	0.43
TD						
A	0.68	0.72	0.70	0.76	0.91	0.83
B	0.52	0.62	0.57	0.63	0.95	0.76
C	0.60	0.77	0.68	0.92	0.94	0.93
G	0.46	0.63	0.53	0.67	0.79	0.73
H	0.43	0.56	0.48	0.80	0.76	0.78
N	0.56	0.71	0.63	0.65	0.71	0.68
TI						
PIA	0.63	0.59	0.61	0.77	0.91	0.83
PIB	0.51	0.51	0.51	0.60	0.92	0.83
PIC	0.54	0.53	0.53	0.91	0.90	0.91
PIG	0.49	0.63	0.53	0.66	0.81	0.79
PIH	0.43	0.59	0.49	0.79	0.76	0.78
PIN	0.51	0.65	0.57	0.68	0.68	0.68

The results presented in Tables III and IV indicate that performance of both discourse-based systems exceeds that of the positional algorithm, with the exception of the topic-independent system, PIC, for identification of thesis statements. One possible explanation might be the following. Topic C is the only *informative* topic. It may be the case that the non-positional features being used for thesis statements look somewhat different for the informative genre. Certainly, more topics from this genre would need to be evaluated to confirm this.

For identification of conclusion statements, the topic-dependent and topic-independent systems have overall higher agreement than for thesis statements, across all measures. As well, there is greater comparability between the two systems for system assignment of conclusion statements. The agreement for the positional algorithm is fairly comparable for thesis and conclusion statements. Therefore, this would suggest that the features used to automatically assign discourse labels are making a stronger contribution toward the recognition of conclusion statements. Additionally, a characteristic of conclusion statements is that they are by definition in the final paragraph. Accurately finding conclusions is more a matter of identifying which sentences (if any) in the final paragraph are actually part of the conclusion statement. On the other hand, thesis statements, although typically found in the first paragraph of an essay, may occur in later paragraphs in an essay. Thesis statements are somewhat more difficult to model as is apparent when we compare system performance for thesis and conclusion statements.

Overall, the results in this study indicate that it is worth continuing research using machine learning approaches for this task, since they clearly outperform the positional baseline algorithm. This assumption has already been borne out in subsequent systems, where we have extended the number of discourse categories to include background information, main ideas and supporting ideas, using enhanced machine learning methods. Details of an essay-based discourse analysis system that is deployed as a commercial application may be found in Burstein *et al*. (2003) and Burstein *et al*. (forthcoming).

Acknowledgements

We owe considerable thanks to Slava Andreyev for discussions during the development of the systems, and for data preparation and system implementation. We would like to thank Marisa Farnum and Hilary Persky for their significant contributions to the annotation protocol, and Jennifer Geoghan and Jessica Miller for doing all of the annotation work. We are grateful to Richard Swartz for continuous support of this research. We thank the anonymous reviewers for their helpful comments. This work was completed at ETS Technologies, Inc. Any opinions expressed in this paper are those of the authors and not necessarily of Educational Testing Service.

Notes

[1] Our initial funding resources included annotation for the three data sets used in the pre-training. After training was completed, we received additional funding, At this point, we believed that the annotator agreement was highly reliable, and decided that it was more important to have them annotate a larger data set, then to spend additional time training.

[2] C5.0 machine learning software was licensed from RuleQuest Research. More information about the software can be found at: http://www.rulequest.com/.

References

Burstein J., Leacock C., Chodorow M. (forthcoming) Criterion On-line Essay Evaluation: An Application for Automated Evaluation of Student Essays. To appear in *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence*, Acapulco, Mexico, August, 2003.

Burstein J., Marcu D., Knight K. (2003) Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. In Harabagiu S. and Ciravegna F. (eds.), *Special Issue on Advances in Natural Language Processing, IEEE Intelligent Systems*, Vol. 18, No. 1, pp. 32–39.

Burstein J. (2003) The *E-rater*® Scoring Engine: Automated Essay Scoring With Natural Language Processing. In Shermis M.D. and Burstein J. (eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective*, Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, pp. 113–121.

Burstein J., Marcu D. (2003) Automated Evaluation of Discourse Structure in Student Essays. In Shermis M.D. and Burstein J. (eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective*, Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, pp. 209–229.

Burstein J., Marcu D., Andreyev S., Chodorow M. (2001) Towards Automatic Classification of Discourse Elements in Essays. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, July, 2001, 15–21.

Burstein J., Kukich K., Wolff S., Lu C., Chodorow M. (1998a) Enriching Automated Scoring using Discourse Marking. In *Proceedings of the Workshop on Discourse Relations & Discourse Marking, Annual Meeting of the Association of Computational Linguistics*, August, 1998. Montreal, Canada, pp. 90–97.

Burstein J., Wolff Kukich K., Lu S., Chodorow C., Braden-Harder L.M., Harris M.D. (1998b) Automated Scoring Using A Hybrid Feature Identification Technique. *Proceedings of ACL*, pp. 206–210.

Elliott S. (2003) *Intellimetric*™: From Here to Validity. In Shermis M.D. and Burstein J. (eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective*, Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, pp. 71–86.

Krippendorff K. (1980) *Content Analysis: An Introduction to Its Methodology*. Sage Publishers, Thousand Oaks, CA.

Landauer T., Laham D., Foltz P. (2003) Automated Scoring and Annotation of Essays with the Intelligent Essay Assessor. In Shermis M.D. and Burstein J. (eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective*, Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, pp. 87–112.

Larkey L., Croft W.B. (2003) A Text Categorization Approach to Automated Essay Scoring. In Shermis M.D. and Burstein J. (eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective*, Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, pp. 55–70.

Leacock C., Chodorow M. (2003) Automated Grammatical Error Detection. In Shermis M.D. and Burstein J. (eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective*, Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, pp. 195–207.

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

Marcu D. (2000) *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press.
Page E.B. (2003) Project Essay Grade: PEG. In Shermis M.D. and Burstein J. (eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective*, Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, pp. 43–54.
Quirk R., Greenbaum S., Leech S., Svartik J. (1985) *A Comprehensive Grammar of the English Language*. Longman, New York.
Scardamalia M., Bereiter C. (1985). Development of Dialectical Processes in Composition. In Olson D.R., Torrance N. and Hildyard A. (eds.), *Literacy, Language, and Learning: The Nature of Consequences of Reading and Writing*. Cambridge University Press.
White E.M. (1994) *Teaching and Assessing Writing*. Jossey-Bass Publishers, pp. 103–108.