Classifying paragraph types using linguistic features: Is paragraph positioning important?

Scott A. Crossley°, Kyle Dempsey^ & Danielle S. McNamara^

°Georgia State University, GA - ^University of Memphis, TN - ^Arizona State University, AZ | USA

Abstract: This study examines the potential for computational tools and human raters to classify paragraphs based on positioning. In this study, a corpus of 182 paragraphs was collected from student, argumentative essays. The paragraphs selected were initial, middle, and final paragraphs and their positioning related to introductory, body, and concluding paragraphs. The paragraphs were analyzed by the computational tool Coh-Metrix on a variety of linguistic features with correlates to textual cohesion and lexical sophistication and then modeled using statistical techniques. The paragraphs were also classified by human raters based on paragraph positioning. The performance of the reported model was well above chance and reported an accuracy of classification that was similar to human judgments of paragraph type (66% accuracy for human versus 65% accuracy for our model). The model's accuracy increased when longer paragraphs that provided more linguistic coverage and paragraphs judged by human raters to be of higher quality were examined. The findings support the notions that paragraph types contain specific linguistic features that allow them to be distinguished from one another. The finding reported in this study should prove beneficial in classroom writing instruction and in automated writing assessment.

Keywords: paragraph structure, paragraph function, corpus linguistics, computational linguistics, cognitive modeling

journal of WRITING RESEARCH

Crossley, S.A., Dempsey, K., & McNamara, D.S. (2011). Classifying paragraph types using linguistic features: Is paragraph positioning important? *Journal of Writing Research*, 3(2), 119-143. http://dx.doi.org/10.17239/jowr-2011.03.02.3

Contact: Scott A. Crossley, Department of Applied Linguistics, Georgia State University, Atlanta, A 30303 | USA - sacrossley@gmail.com.

1. INTRODUCTION

The study of the paragraph in the United States has a long and storied history in composition studies (e.g. Christensen, 1965; Karrfalt, 1968) that likely reached its zenith with the "Symposium on the Paragraph" held at the annual College Composition and Communications conference in 1965. Most studies analyzing the paragraph have focused on the rhetorical role of the paragraph within an essay or the structure of the paragraph as a homogeneous entity.[1] Rhetorically, paragraphs follow roles as introductory paragraphs, body paragraphs, or concluding paragraphs. Internally, a paragraph's structure includes topic sentences, body sentences, coordinating and subordinating sentences, and concluding sentences (Christensen, 1965; Grady, 1971). The rhetorical roles and structures of paragraphs have been standardized over the past 40 years or so and distinctions between paragraph features and paragraph types are commonly found in composition textbooks and taught in composition classes, especially in the United States. Our focus in this study is not on redefining the paragraph or its purpose, but on further exploring the linguistic construction of different paragraph types using computational tools.

We take a novel approach in this study by combining corpus and computational linguistic approaches for text analysis and comparing them with human judgments of text type (i.e., a gold standard). Our goal is to provide a linguistic model that permits the discrimination of paragraph types based on linguistic features. Discriminating among paragraph types will allow us to better understand their underlying linguistic features and afford us the opportunity to explain how these linguistic features function within paragraph structures and how they relate to the rhetorical role of the paragraphs. Developing empirically based models of paragraph types and functions will help writers better understand the form and structure of paragraphs and, as a result, help them more efficiently organize essays as well as perceive important links between sentences, paragraphs, and compositions. Computational models of paragraph types can also be used to automatically evaluate texts in intelligent tutoring systems that teach writing skills as well as provide real-time feedback to students regarding the structural patterns and linguistic quality of an essay.

1.1 Paragraph Structure

A paragraph is generally defined as a group of sentences developing a central theme. Paragraphs have often been treated as homogeneous elements that share many similar features regardless of their location in an essay. Therefore, paragraph structure has been argued to be both definable and traceable, much like the structure of sentences (Becker, 1965; Christensen, 1965; Cohan, 1976). Paragraphs have also been described as recognizable units of linguistic structure (Warner, 1979) that can be separated based on grammatical structure (Becker, 1965) and semantic relationships (i.e., content; Christensen, 1965). Such assertions are based on studies that demonstrate readers can

reliably agree on the location of removed paragraph boundaries (Bond & Hayes, 1984; Koen, Becker, & Young, 1969; Young & Becker, 1966). These studies have shown that readers can distinguish original paragraph boundaries that have been removed, even when content words are replaced with nonsense words, based on the cohesive properties of the paragraph (i.e., pronoun reference and word repetition) and the length of sentences and the text (Bond & Hayes, 1984).

The theme of the paragraph is often found in an identifiable topic sentence (Christensen, 1965; Oshima & Hogue, 1997; Warner, 1979; Warriner, 1958) meant to limit the scope of the paragraph (Arnaudet & Barrett, 1990). However, research has shown that topic sentences are only produced in about half of all paragraphs (McCarthy, Renner, Duncan, Duran, Lightman, & McNamara, 2008; Popken, 1987, 1988). When a topic sentence is used, it is thought to limit the scope of the paragraph such that each paragraph has identifiable characteristics that set it apart from other, individual paragraphs.

Linguistically, sentences within paragraphs also cohere, generally through repetition of key words (Christensen, 1965) and conjunctions (Warner, 1979). From a rhetorical position, this cohesiveness is often referred to as subordination, coordination (Christensen, 1965), or addition (Karrfalt, 1968). Subordination reveals itself in either grammatical or semantic subordination. Grammatical subordination involves anaphor, transitional markers (e.g. *therefore, nevertheless, thus*), the use of word repetition (especially root words), and the use of synonyms to link similar words. Coordination between sentences links equivalent ideas generally through parallel structures such as similar syntactic structures and semantic groupings. Addition requires that an additional sentence be added at the end of a paragraph that generalizes the previous sentences. However, textual cohesion, especially as found in subordination and coordination, is argued to vary across paragraph types (Christensen, 1965).

1.2 Essay Structure

Paragraphs come to form the basic structure of an essay. The most common such structure found in writing classes and texts books in the United States is the five-paragraph theme (FPT; Albertson, 2007; Johnson, Smagorinsky, Thompson, & Fry, 2003; Kinneavy & Warriner, 1993; Nunnally, 1991). The FPT is argued to be both a beneficial and disadvantageous writing strategy. The FPT outline is beneficial because it provides a set structure from which to develop a rhetorical position. Additionally, high school students that rely on the organizational theme of the FPT in standardized testing situations are just as likely to earn high scores as those that use non-formulaic schemes (Albertson, 2007). However, the FPT is considered disadvantageous because it does not balance the rhetorical needs of the audience, the writing purpose, and the message. As a result, essays written using the FPT may not convey contextually sensitive information (Hillocks, 2005; Mabry, 1999). For high school and freshmen college students, the usefulness of the FPT is thought to outweigh its potential for limiting creativity because it is seen as a useful guide for early essay writing success that eventually may prime

developing writers to advance their writing techniques and modes of expression (Haswell, 1986). The importance of the FPT in the United States is even more consequential when one considers the reality of using standardized testing metrics such as the Scholastic Assessment Test (SAT) or the American College Test (ACT), which privilege the FPT and likely influence teacher practice (Brindley & Schneider, 2002; Hillocks, 2002).

In the United States, the focus on the FPT has contributed to clarifying the definition of essay structure at the paragraph level. This structure includes an introduction paragraph, followed by body paragraphs that support the topic of the introduction, and a conclusion paragraph that summarizes the paper (Grady, 1971). These paragraph types come to structure the FPT essay.

1.3 Paragraph Types

Recommendations for writing the various types of paragraphs found in essays abound in writing textbooks, but were likely standardized by Grady (1971). Grady's guidelines were specific for introductory, body, and concluding paragraphs.[2] He argued that introductions should begin with the main idea of the paper as found in the topic sentence and then present the supporting arguments to be discussed in general detail. These sentences should be immediately related but not interrelated. Body paragraphs (or topic paragraphs) should be a tight grouping of ideas that expand on the supporting arguments found in the introduction and relate to the central theme. However, each body paragraph should be dissimilar to other supporting body paragraphs. Body paragraphs should also be guided by a topic sentence that relates to all sentences within the paragraph. The concluding paragraph should contain a summarization of all the information presented in the essay without presenting new information. The conclusion should then present a general statement of findings and speculation. Grady did not argue that there would only be one introductory or conclusion paragraph in an essay, but rather argued that there is a connection between text length and the number of introductory and conclusion paragraphs in an essay such that longer essays may contain introductions and conclusions that comprise more than one paragraph.

2. METHODS

Our purpose in this study is to test the notion that a paragraph's structure is definable and traceable (Christensen, 1965) by using computational, corpus, and machine learning techniques to assess differences in paragraph types based on their position in an essay. Our approach is very much rooted in the work of past analyses that have examined the linguistic features of texts to better understand their rhetorical functions (i.e., Swales, 1990; Biber, 1989). For the purposes of our study, we collected a corpus of argumentative essays generated by college freshman composition students. We divided the essays into paragraphs based on positioning: the initial paragraph, middle paragraphs, and the final paragraph. We argue that these positions relate to

introductory paragraphs, body paragraphs, and conclusion paragraphs. We then used the computational tool Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004) to analyze the linguistic features in the paragraphs and used the reported findings as the foundation for a statistical analysis. We also used human judgments of paragraph type from the same corpus as a baseline from which to compare the efficacy of the statistical model. Using corpus, computational, and statistical approaches, we examined whether linguistic features can be used to distinguish essay paragraph types and to what degree (as compared to human judgments).

2.1 Corpus Selection

The corpus used for this study was collected from undergraduate students at a large university in the United States enrolled in a freshman composition class. The students were not instructed to use a five-paragraph themed essay; however, the median number of paragraphs for the collected essays was five (Mean = 5.358; Minimum = 1; Maximum = 10), potentially reflecting the dominance of the five-paragraph theme in the United States. The corpus was designed to consider learner variables such as age (university students in their 20s), proficiency (average writers not in remedial composition courses or in advanced composition courses), and mother tongue (all native speakers of English). The corpus was also designed to consider task variables such as medium (writing), genre (argumentative essays), prompts (three different prompts, see Table 1 for more detail), and essay length (between 500 and 1,000 words). The corpus was not corrected for spelling or grammar errors. The corpus has proven reliable in past studies examining the linguistic constructs of essay writing (Crossley & McNamara, 2009; Crossley & McNamara, 2010; McNamara, Crossley, & McCarthy, 2010). To develop the paragraph corpus used in this study, we randomly selected 182 paragraphs from the corpus from essays that contained at least 3 paragraphs. These 182 paragraphs came from 76 essays. Of these 182 paragraphs, roughly 20% were initial paragraphs, 60% were middle paragraphs, and 20% were final paragraphs. Descriptive statistics for this corpus are provided in Table 2.

Table 1: Essay topics in corpus

Topics	Number of paragraphs
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?	71
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.	51
In his novel 'Animal Farm', George Orwell wrote "All men are equal: but some are more equal than others". How true is this today?	60

Table 2: Descriptive statistics for corpus

Paragraph	Mean number of words	Standard deviation	Paragraphs in training set	Paragraphs in test set	Paragraphs in total corpus
Initial	98.22	43.28	19	13	32
Middle	141.07	65.07	78	36	114
Final	103.33	59.74	24	12	36

2.2 Human Judgments

We collected human judgments of paragraph types (i.e., introduction, body, and conclusion paragraphs) to assess human ability to distinguish among these types of paragraphs. To collect these judgments, two human raters were extensively trained on a paragraph type coding scheme and completed ratings on a training set. The raters were trained to consider multiple characteristics of paragraph types such as introduction types, theses, arguments, topic sentences, evidential sentences, conclusion summaries, and conclusion types. After training, the expert raters scored all paragraphs in the corpus. The raters were not given specific linguistic information about potential differences between paragraph types. Using the coding scheme, the expert-raters classified each paragraph on a binary (yes/no) level as to whether a paragraph was an initial, middle, or final paragraph. The presentation of the paragraph was randomized and raters were not told the percentage of paragraphs in each classification. The two raters had an inter-rater reliability of $r = .72$. If rater differences existed, they were resolved through discussion until agreement was met.

2.3 Statistical Analysis

To examine the hypothesis that there are linguistic features that differentiate various paragraph types, we conducted a discriminant function analysis using computational indices provided by Coh-Metrix. All the selected Coh-Metrix indices have been validated in past writing studies (Crossley & McNamara, 2009; Crossley, McNamara, Weston, & McClain Sullivan, 2011; McNamara et al., 2010) and have strong overlap to paragraph development, writing theory, and text readability. However, the indices we selected are specific to the tool we chose (Coh-Metrix) and thus may not reflect all possible linguistic features that may differentiate paragraph types. The selected measures evaluated lexical coreferentiality, semantic coreferentiality (Latent Semantic Analysis), word frequency measures, word information measures from the MRC Psycholinguistic database, hypernymy and polysemy values, causality, and syntactic complexity from which to select individual indices. We also included a measure of text length. These measures are briefly discussed below. More detailed information can be found in Graesser et al. (2004) and McNamara, Louwerse, McCarthy, & Graesser (2010).

Coreferentiality

Coh-Metrix considers four forms of lexical co-reference between sentences: noun overlap between sentences, argument overlap between sentences, stem overlap between sentences, and content word overlap between sentences. Noun overlap measures how often a common noun is shared between two sentences. Argument overlap measures how often two sentences share nouns with common stems, while stem overlap measures how often a noun in one sentence shares a common stem with other word types in another sentence. Content word overlap measures how often sentences share content words. Word overlap is an important indicator of text cohesion and greater word overlap helps to construct larger units of meaning in a text (Just & Carpenter, 1987; Rayner & Pollatsek, 1994). Word overlap is also an important indicator of paragraph boundaries (Bond & Hayes, 1984).

Latent Semantic Analysis (LSA)

Coh-Metrix measures semantic coreferentiality using LSA, which is a mathematical and statistical technique for representing deeper world knowledge based on large corpora of texts. LSA uses a general form of factor analysis to condense a very large corpus of texts to 300-500 dimensions. These dimensions represent how often a word occurs within a document (defined at the sentence level, the paragraph level, or in larger sections of texts) and each word, sentence, or text is calculated as a weighted vector (Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998). The relationships between the vectors form the basis for representing semantic similarity between words. However, unlike lexical markers of coreferentiality, LSA connects words that are semantically similar, but may not be related morphologically. For instance, the word *mouse* has a higher LSA score when compared to *cat* than to either *dog* or *house*.

In addition, Coh-Metrix also analyzes *givenness* through LSA by measuring the proportion of new information each sentence provides according to LSA. Given information is recoverable from the preceding discourse (Halliday, 1967) and does not require activation (Chafe, 1975). Given information is thus less taxing on the reader's cognitive load. To compute the LSA givenness index, each sentence in the input text is represented by an LSA vector. The amount of new information a sentence provides is computed from the component of the corresponding sentence vector that is perpendicular to the space spanned by the previous sentence vectors. Similarly, the amount of given information of a sentence is the parallel component of the sentence vector to the span of the previous sentence vectors (Hempelmann, Dufty, McCarthy, Graesser, Cai, & McNamara, 2005).

Word Frequency

Word frequency in Coh-Metrix refers to metrics of how frequently particular words occur in the English language. The primary frequency count in Coh-Metrix is provided by CELEX (Baayen, Piepenbrock, & Gulikers, 1995), the database from the Centre for

Lexical Information, which consists of word frequencies reported in the 1991 version of the COBUILD corpus, a 17.9 million-word corpus. Measuring word frequency is important because more frequent words allow for quicker text decoding (Perfetti, 1985; Rayner & Pollatsek, 1994). Automatic decoding reduces demands on a reader's working memory. In contrast, when readers have difficulty decoding a text, more processing resources are dedicated to decoding rather than comprehension, which, in turn, negatively affects readers' ability to recall the text (Field, 2004). Studies have also demonstrated that frequent words are also processed and comprehended more quickly than infrequent words (Haberlandt & Graesser, 1985; Just & Carpenter, 1980).

Word Information (MRC Psycholinguistic Database)

Coh-Metrix calculates information at the lexical level on five psycholinguistic matrices: familiarity, concreteness, imagability, meaningfulness, and age of acquisition. All of these measures are reported by the MRC Psycholinguistic Database (Coltheart, 1981) and are based on the works of Paivio (1965), Toglia and Battig (1978) and Gilhooly and Logie (1980), who used human subjects to rate large collections of words for the targeted psychological properties. Because most MRC measures are based on psycholinguistic experiments, the coverage of words differs among the measures (e.g., the database contains 9,240 words with imagery ratings and 9,392 with familiarity ratings). Many of these indices are important for measuring the strength of word associations and general lexical difficulty. For example, the MRC word meaningfulness score relates to how strongly words associate with other words, and how likely words are to prime or activate other words. In relation to lexical difficulty, MRC word familiarity, concreteness, imagability, and age of acquisition scores measure lexical constructs such as word exposure (familiarity), word abstractness (concreteness), the evocation of mental and sensory images (imagability), and intuited order of lexical acquisition (age of acquisition). For a full review of these indices as reported by Coh-Metrix, refer to Salsbury, Crossley, and McNamara (2010).

Hypernymy and Polysemy Indices

Coh-Metrix tracks the relative ambiguity of a text by calculating its lexical, polysemy value, which refers to the number of meanings or senses within a word. Coh-Metrix tracks the relative specificity of a text by calculating its lexical, hypernymy value, which refers to the number of levels a word has in a conceptual, taxonomic hierarchy. The number of meanings and the number of levels attributed to a word are measured in Coh-Metrix using WordNet (Fellbaum, 1998; Miller, Beckwith, Fellbaum, Gross, & Miller, 1990). For instance, on a hypernymic scale, *animal* would be less specific than *dog*. According to the polysemy value, *class* (*Her clothes have class. The class was short. I am class of 99.*) would be considered more ambiguous than *lentil*, which has only one sense.

Causal Cohesion

Causal Cohesion is measured in Coh-Metrix by calculating the ratio of causal verbs to causal particles (Dufty, Hempelmann, et al., 2005). The incidence of causal verbs and causal particles in a text relates to the conveyance of causal content and causal cohesion. The causal verb count is based on the number of main causal verbs identified through WordNet (Fellbaum, 1998; Miller et al., 1990). These include verbs such as *kill, throw*, and *pour*. The causal particle count is based on a defined set of causal particles such as *because*, *consequence of*, and *as a result*. Causality is an important indicator of relations between events and actions (i.e., stories with an action plot or science texts with causal mechanisms). Causality can also show causal relationships between simple clauses at the sentential level (Pearson, 1974-1975).

Syntactic Complexity

Syntactic complexity is measured by *Coh-Metrix* in four major ways. First, there is an index that calculates the mean number of words before the main verb with the assumption that more words before the main verb lead to more complex syntactic structure. Second, there is an index that measures the mean number of high level constituents (sentences and embedded sentence constituents) per word with the understanding that more higher-level constituents per word leads to a more complex syntactic structure. Coh-Metrix also calculates an index that assesses syntactic similarity by measuring the uniformity and consistency of the syntactic constructions in the text. This index not only looks at syntactic similarity at the phrasal level, but also takes account of the parts of speech involved. The uniformity of syntax across sentences results in overlap between sentences, leading to text that is easier to process. In contrast, sentences that demonstrate complex syntactic constructions such as embedded constituents are more difficult to process and comprehend (Perfetti, Landi, & Oakhill, 2005).

Connectives and Logical Operators

Coh-Metrix examines the density of connectives on two dimensions. The first dimension contrasts positive versus negative connectives, whereas the second dimension is associated with particular classes of cohesion identified by Halliday and Hasan (1976) and Louwerse (2001). These connectives are associated with positive additive (*also, moreover*), negative additive (*however, but*), positive temporal (*after, before*), negative temporal (*until*), positive logical (*and, also, then, in sum, next*) and causal (*because, so*) measures. Connectives help link ideas and clauses and lead to more cohesive texts (Crismore, Markkanen, & Steffensen, 1993; Longo, 1994) that contain more text organizational clues (van de Kopple, 1985). The logical operators measured in Coh-Metrix include variants of *or*, *and*, *not*, and *if-then* combinations. Logical connectors relate to a text's density and abstractness and correlate with working memory demands (Costerman & Fayol, 1997).

Total Number of Words in the Text

While not exactly a linguistic feature, text length plays an important role in writing quality because more proficient writing generally contains more words (Crossley & McNamara, in press). The number of words in a text also relates to the density of text propositions (Kintsch & Keenan, 1973) and is an important indicator of paragraph boundaries (Bond & Hayes, 1984).

2.4 Variable Selection

To select the individual indices from the chosen Coh-Metrix measures, we used training and test sets (Witten & Frank, 2005). We divided the corpus into two sets: a training set (n = 121) and a test set (n = 61) based on a 67/33 split. The purpose of the training set was to identify which of the variables contained within the chosen Coh-Metrix banks of indices best distinguished the paragraph types based on positioning in the essays. These selected variables were later used to predict the paragraph types in the training set using a discriminant function analysis (DFA; described below). We then used the DFA model to analyze the paragraphs in the test set. Testing the explanatory power of the DFA model on the test set allowed us to accurately predict the performance of our model on an independent corpus (Witten & Frank, 2005). The results of the DFA were later compared with the results from the human analysis to examine similarities between human and machine ratings.

A discriminant function analysis is a common approach used in many previous studies that have distinguished text types (e.g. Biber 1993; Crossley & McNamara, 2009). Considering that the training set contained 121 paragraphs and using a moderate estimate of one predictor per 15 variables, we determined that 8 indices would be an appropriate number of predictors for the discriminant analysis that would not create problems of overfitting. Such a ratio is standard for analyses of this kind (Field, 2005). Avoiding over-fitting a model is crucial because if too many variables are used the model fits not just the signal of the predictors but also the unwanted noise. When a training model is overfitted, the model is predictive of the data in the training set, but will likely be inaccurate in predicting the data in the test set because the noise will not be the same from data set to data set.

3. RESULTS

3.1 Human classification accuracy

The human raters correctly allocated 138 of the 162 paragraphs in the training set (df=4, n=182) $\chi 2$= 92.976, p < .001) for an accuracy of 76%. The results for each classification are located in Table 3.

Table 3: Predicted paragraph type versus actual paragraph type for human raters

	Predicted paragraph type by human raters		
Actual paragraph type in corpus	Initial	Middle	Final
Initial	**15**	10	7
Middle	4	**104**	6
Final	4	13	**19**

We also report results in terms of precision and recall. Precision scores are computed by tallying the number of hits over the number of hits + misses. Recall is the number of correct predictions divided by the sum of the number of correct predictions and false positives. These are important because an algorithm could predict everything to be a member of a single group and score 100% in terms of recall. However, it could only do so by claiming members of the other group. If this were the case, though, the algorithm would score low in terms of precision. By reporting both values, we can better understand the accuracy of the model. The accuracy of humans for predicting paragraphs types is provided in Table 4. The combined accuracy for precision and recall scores (*F*1) for the full set was .66.

Table 4: Precision and recall for human raters' classifications of paragraphs (full set)

Paragraph set	Recall	Precision	F1
Initial	0.652	0.469	0.545
Middle	0.819	0.912	0.863
Final	0.594	0.528	0.559

3.2 Variable Selection

In order to select the variables for the DFA, one-way ANOVAs were conducted using the selected Coh-Metrix measures as the dependent variables and the paragraph types from the training set as the independent variables. We selected the variable with the largest effect size as the representative variable for that measure. We selected more than one variable from the MRC Database because the indices measured different linguistic features. Descriptive statistics for the selected variables are presented in Table 5. All measures except syntactic complexity, word frequency, and causal cohesion reported at least one index that demonstrated significant differences among paragraph types.

Table 5: Means (standard deviations), F values, and effect sizes for paragraph types

	Initial	Middle	Final	*F(2,118)*	η_p^2.
Number of words in the text	94.84	143.59	102.13	7.262	0.110
	(42.25)	*(64.8)*	*(65.69)*		
LSA given/new	0.21	0.250	0.21	5.506	0.085
	(0.06)	*(0.07)*	*(0.06)*		
Word hypernymy	1.86	1.75	1.56	4.955	0.077
	(0.36)	*(0.31)*	*(0.33)*		
Content word overlap	0.09	0.13	0.15	4.297	0.068
	(0.06)	*(0.06)*	*(0.08)*		
Word familiarity content words	576.18	576.59	582.11	4.008	0.064
	(9.24)	*(8.92)*	*(7.42)*		
Positive logical connectives	19.21	33.76	35.45	3.621	0.058
	(11.95)	*(22.74)*	*(27.01)*		
Word meaningfulness content words	440.57	426.77	425.89	3.464	0.055
	(31.24)	*(17.54)*	*(23.32)*		
Word imagability every word	328.48	323.43	318.56	3.190	0.051
	(20.13)	*(11.53)*	*(9.01)*		

Note: Standard deviations are in parentheses

To avoid the risk of collinearity between variables, we ensured that no index pair was correlated above r => .70. The potential redundancy of variables correlated above .70 renders interpretation of the results difficult because it is not possible to determine which variables contribute to the model (Brace, Kemp, & Snelgar, 2006; Tabachnick & Fidell, 2001). However, none of the variables used in this analysis were highly correlated.

3.3 Pairwise Comparisons

A series of pairwise comparisons was conducted to examine the differences reported in

Table 6: Summary of pairwise comparisons among paragraph types

	Initial paragraph	Middle paragraph	Final paragraph
Number of words	--	+	-
LSA given/new	-	+	-
Word hypernymy	+	+	-
Content word overlap	-	+	+
Word familiarity	-	-	+
Connectives	-	+	+
Word meaningfulness	+	-	-
Word imagability	+	ND	-

Note: ND signifies no differences; + indicates that there is significantly more; – indicates that there is significantly less respective to the other paragraph types; -- indicates that the index is significantly lower than -.

Table 4 between initial, middle, and final paragraphs for each selected Coh-Metrix index. The findings from the pairwise comparisons are summarized in Table 6, and described below.

Number of words in the paragraph. Middle paragraphs were longer than both initial ($p < .01$) and final ($p < .01$) paragraphs. Initial paragraphs were also significantly shorter than final paragraphs ($p < .01$).
LSA Given/New Information. Middle paragraphs contained more given information than initial paragraphs ($p < .01$) and final paragraphs ($p < .01$). No significant differences were found between initial and final paragraphs.
Word Hypernymy. Final paragraphs were less specific than initial paragraphs ($p < .01$) and middle paragraphs ($p < .01$). The latter two demonstrated no significant differences in hypernymy values.
Content Word Overlap. Initial paragraphs contained less content word overla*p* than both middle and final paragraphs ($p < .01$). No significant differences were found in content word overlap between middle and final paragraphs.
Word Familiarity. Final paragraphs contained significantly more familiar words than initial and middle paragraphs ($p < .05$). No differences were found between initial and middle paragraphs.
Connectives. Initial paragraphs contained significantly fewer positive logical connectives (e.g. *and, also, then, in sum, next*) than middle and final paragraphs ($p < .01$). No significant differences were found for positive logical connectives between middle and final paragraphs.
Word Meaningfulness. Initial paragraphs contained significantly more meaningful words than middle and final paragraphs ($p < .01$). No differences in word meaningfulness scores were found between middle and final paragraphs.
Word Imagability. Initial paragraphs contained significantly more imagable words than final paragraphs ($p < .01$), but not middle paragraphs. No differences were found between middle and final paragraphs.

3.4 Accuracy of Model

We conducted a discriminant function analysis (DFA) to test the accuracy of these indices in distinguishing paragraph types. A DFA predicts group membership (in this case paragraph type) using a series of independent variables (in this case the selected Coh-Metrix variables). The training set is used to generate a discriminant function, which acts as the algorithm that predicts group membership. This discriminant function is later used to predict group membership of the essays in the test set. We describe the DFA findings by reporting an estimation of the accuracy of the analysis. This estimation is made by plotting the correspondence between the actual paragraph types in the test and training sets and the predictions made by the discriminant analysis (see Table 7).

Table 7: *Predicted paragraph type versus actual paragraph type results from both training set and test set*

Actual paragraph type in corpus	**Predicted paragraph type by DFA**		
Training set	Initial	Middle	Final
Initial	**10**	4	5
Middle	21	**39**	18
Final	6	4	**14**
Test set	Initial	Middle	Final
Initial	**11**	0	2
Middle	5	**26**	5
Final	2	3	**7**

The results show that the discriminant analysis, using the eight variables, correctly allocated 63 of the 121 essays in the training set (df=4, n=121) χ^2= 18.290, $p < .001$). For the test set, the discriminant analysis correctly allocated 44 of the 61 essays (df=4, n=61) $\chi 2$= 37.526, $p < .001$). For the test set, the model provides 72% accuracy. The accuracy for the entire data set is 59%. The estimate of success based on chance alone for this data set is 33%. The accuracy of the model for predicting paragraph types can be found in Table 8. The combined accuracy of the model for both precision and recall ($F1$) in the training set was .56. The accuracy for the test set was .65.

Table 8: Precision and recall finding (training and test set)

Training set: Paragraph set	Recall	Precision	F1
Initial	0.568	0.526	0.546
Middle	0.830	0.503	0.626
Final	0.486	0.500	0.493
Test set: Paragraph set	Recall	Precision	F1
Initial	0.579	0.846	0.687
Middle	0.839	0.722	0.776
Final	0.412	0.583	0.483

3.5 Human classification judgments compared to machine classification

To test similarities between human classification judgments and those made by the model, we compared the human categorizations to the classification yielded by the discriminant function analysis model. Human classification judgments and the classifications made by the model were significantly similar (df=4, n=182) $\chi 2$= 31.523, $p < .001$) and were in agreement 54% of the time. The reported Kappa for the agreement between humans and the model was .253 demonstrating a fair agreement. The cross-tabulation between human raters and the DFA model is located in Table 9.

Table 9: Agreement between predictions made by human raters and the DFA model

	Predicted paragraph type by DFA		
Predicted paragraph type by raters	Initial	Middle	Final
Initial	**11**	32	12
Middle	5	**70**	2
Final	7	26	**17**

4. POST HOC ANALYSES

The following analysis explores what factors outside of the linguistic features of the paragraphs explain the classification statistics reported for both the DFA model and the human raters. We hypothesized that paragraph length and writing quality might affect the classification results. Our hypothesis was premised on the notion that longer paragraphs would provide both machines and humans with more linguistic features from which to make accurate judgments and that paragraphs judged to be of higher quality by human raters would be easier to classify because they would adhere to the rhetorical expectations of the paragraph type.

We first asked human raters to evaluate the characteristics of introductory paragraphs, body paragraphs, and conclusions paragraphs using a survey instrument. The survey instrument used in this analysis was designed to parallel the survey instrument used initially by Breetvelt, van den Bergh, and Rijlaarsdam (1996) and later adapted with a focus on structure and argumentation by Sanders and Schilperoord (2006). Three experts in language processing with Ph.D.s in either linguistics or cognitive psychology developed the instrument. It was then subjected to usability tests by expert raters with at least three years experience in essay scoring. The instrument prompted raters to provide analytic judgments on the quality of specific writing features related to introductory, body, and conclusion paragraphs using a 1 to 6 Likert interval scale. Raters evaluated introductory paragraphs based on an essay's strength of introduction and the strength of the thesis. The instrument asked raters to evaluate the

quality of body paragraphs based on strength of the topic sentences and evidential sentences. Lastly, the survey prompted raters to evaluate the quality of concluding paragraphs based on the strength of the conclusion type, how well the conclusion summarized the main ideas of the essay, and the strength of the closing of the essay.

Two expert raters with at least three year's experience teaching composition classes at a large university were selected to score the 182 paragraphs. Each feature of the paragraph was scored from a 1 (lowest) to a 6 (highest). The raters were informed that the distance between each score was equal. Accordingly, a score of 5 was as far above a score of 4 as a score of 2 was above a score of 1. The raters were first trained to use the survey instrument with 20 paragraphs. A Pearson correlation for each evaluation was conducted between all possible pairs of raters' responses. If the correlations between all raters did not exceed $r = .70$ on the items, the ratings were reexamined until scores reached the $r = .70$ threshold. After the raters had reached an inter-rater reliability of at least $r = .70$, each rater then evaluated the 182 paragraph that comprised the corpus used in this study.

4.1 Human quality scores and paragraph classification

T-tests were conducted to examine if the human quality scores of the paragraphs correctly classified by the DFA model as compared to the human quality scores of those classified incorrectly. The paragraphs that were misclassified by the DFA had a mean human quality score of 4.0178 ($SD = 0.847$) and the paragraphs that were correctly classified by the DFA had a mean human quality score of 4.289 ($SD = 0.735$). This difference was significant, $t(180) = -2.300$, $p < .05$. For the human classification results, the paragraphs that were misclassified had a mean score of 3.965 ($SD = 0.879$) and the paragraphs that were correctly classified had a mean score of 4.226 ($SD = 0.765$). This difference was not significant, $t(180) = -1.736$, $p = .144$. We also analyzed the percentage of paragraphs that were correctly classified that received a score of 3.5 and lower or 4.5 and higher to assess if paragraphs scored higher by the human quality raters were classified to a greater accuracy by both the DFA model and the human classification raters than those paragraphs that were scored lower by the human quality raters. The DFA model correctly classified 63% of the paragraphs that were scored 4.5 or higher and 43% of the paragraphs that were scored as a 3.5 or lower. The human classification raters correctly classified 100% of the essays that were scored a 4.5 or higher and 81% of these essays scored 3.5 or lower. These findings support the notion that the quality of the paragraphs affects the classification results with higher quality paragraphs having higher classification accuracy.

4.2 Paragraph length and paragraph classification

T-tests were conducted to examine if the length of the paragraphs was related to classification accuracy for both the DFA model and human results. For the DFA model, the paragraphs that were misclassified had an average length of 107.080 ($SD = 46.309$) and the paragraphs that were correctly classified had an average length of 137.672 (*SD*

= 70.126). This difference was significant, $t(180) = -3.413$, $p < .001$ and demonstrates that longer paragraphs were classified to a greater accuracy than shorter paragraphs. For the human results, the paragraphs that were misclassified had an average length of 111.352 (SD = 63.301) and the paragraphs that were correctly classified had an average length of 128.932 (SD = 62.987). This difference approached significance, $t(180) = -1.467$, $p = .084$. We also examined the classification averages for those paragraphs that were above and below 110 words to test classification accuracy for longer and shorter paragraphs. The DFA model classified 72% of the paragraphs above 110 words and 46% of the paragraphs that were under 110 words. The human raters classified 88% of the paragraphs that were above 110 words and 74% of the paragraphs correctly that were under 110 words. These findings support the notion that the number of words in a paragraph affects the classification results with longer paragraphs having better classification results.

5. DISCUSSION

Our analysis has demonstrated that the paragraphs in our corpus can be distinguished from one another based on positioning (initial, middle, and final) using the linguistic indices reported by Coh-Metrix. These positions are proxies for rhetorical paragraph types (i.e. introduction, body, and conclusion paragraphs) and we therefore contend that paragraph types contain specific linguistic features that allow them to be distinguished from one another (at least for the genre and population that we focus on in this study). Thus, we argue that top-level rhetorical patterns in argumentative essays (i.e., paragraphs) can be distinguished from one another and that these patterns hold across a variety of prompts. Such a finding counters notions presented by Christensen (1965) and Haswell (1986) that paragraph types would vary widely based on prompts. Instead, we see that paragraph types within a specific genre, in this case argumentative essays, are classifiable. The performance of our model on our test set is well above chance and reports an accuracy of classification that is similar to human judgments of paragraph type (66% accuracy for human versus 65% accuracy for our model). The model reported increased accuracy when we examined longer paragraphs that provided more linguistic coverage and when we examined paragraphs judged by human raters to be of higher quality.

The results support the notion that paragraphs are recognizable units of linguistic structure and that the structure of different paragraph types are definable much like sentence types are definable (e.g., topic sentences; McCarthy et al., 2008). We, thus, argue against Stern's (1976) position that the paragraph is not a logical unit. However, we recognize that our notion of a paragraph as a logical unit may differ in description to that discussed by Stern. Stern argued that because paragraphs do not necessarily begin with topic sentences and do not handle distinct topics, they are not independent, self-contained wholes, but part of the discourse of the essay and, therefore, flexible, rhetorical instruments. While our study does not contradict the notion that paragraphs

are heterogeneous, the notion that paragraphs are integral to the rhetorical functions of compositions appears to compel writers to produce identifiable paragraphs that contain certain rhetorical attributes. In what follows, we describe in detail the linguistic differences between paragraphs types found in our analysis and how these differences produce unique structures that relate to the rhetorical purpose of the paragraph.[3] Table 7 contains an overview of these differences.

5.1 Initial/Introductory Paragraphs

The results of our analysis suggest that introductory paragraphs, as compared to body and concluding paragraphs, are characterized by shorter length, lower content word overlap, the use of fewer positive, logical connectors (e.g., *and, also, then*), and words that are more specific, more meaningful, and more imagable. In comparison to concluding paragraphs, the words in introduction are also less familiar. Lastly, in comparison to body paragraphs, introductory paragraphs contain less given information. This combination of linguistic features produces a rhetorical structure that is syntactically less embedded than other paragraphs allowing for the production of a clear, direct main idea. Because the introduction paragraph is concerned with stating a main idea and providing supporting arguments in a general sense, the paragraph does not depend on the overla*p* of content words to produce a cohesive structure. The generality of the introduction paragraph also helps to explain it shortness. From a lexical perspective, the introductory paragraph provides more specific words related to the supporting arguments. These words prime more lexical associations (i.e., word meaningfulness values) likely to induce a variety of ideas in the reader. The general nature of the paragraph likely produces words that are less familiar while the need to write clear supporting arguments likely produces more imagable words. Characteristics such as these can be seen in the following introductory paragraph taken from our corpus:

> I believe that there is still imagination and dreaming in this modern world dominated by science, technology, and industrialization. Just look at the movies, books, television shows, buildings, and new inventions made today. They all have some imagination and dreaming.

This introductory paragraph is short, contains few logical, positive connectives (and thus little embedding of ideas), expresses a main idea and supports that main idea with specific, meaningful, and imagable supporting arguments that are not elaborated on and thus do not overlap with other words in the paragraph. It fits the description of introductory paragraphs as developed by Grady (1971) in that it starts with a main idea, provides supporting arguments that are immediately related, but not interrelated.

5.2 Middle/Body Paragraphs

Our analysis depicts body paragraphs as containing more words when compared to introductory and conclusion paragraphs. Body paragraphs also contain more given

information and contain words that are less imagable when compared to introductions, but more imagable than conclusions. When contrasted with conclusion paragraphs only, the words in a body paragraph are less familiar. These characteristics relate to the rhetorical purposes of paragraphs as developed by Grady (1971). For instance, more words and a greater amount of given information likely support the notion that body paragraphs contain tighter groupings of ideas (more given information) that expand on the supporting arguments found in the introduction (more words). Unlike an introduction paragraph, though, a body paragraph expands on ideas and likely does not need to rely on highly imagable words. The following paragraph taken from our corpus illustrates these characteristics:

> During the adolescence stage is when a person figures out who he is and what his plans are for the future. This period is very critical. Many people start dreaming of what they want to become and what type of life they want to pursue. For females, not often as many males, this is when they start dreaming of the career, spouse, and family. In my early adolescence, I dreamed of having a husband with a great career, two children, a pet, and a lot of money. As I have gotten older, it is mostly whom I fall in love with and who will treat me well. That is not saying that I do not dream about whom I am going to share my dreams with and be together for the rest of my life. I still have an ideal husband and family thought of in my head. However, the imagination is not as widely spread during adolescence as childhood.

This body paragraph is longer than the average introduction and concluding paragraphs in our corpus. Much of the information in the paragraph is given information (i.e., recoverable from preceding discourse) such as *family, husband, children, love, childhood* and there is more overlap of content words when compared with the introductory paragraph (e.g., *adolescence, dream, life*) creating a more coherent structure. The words are not highly imagable when compared with those in the introduction paragraph (i.e., *period, person, critical, life, dream*). Thus, the paragraph appears to function as a means of elaborating on a specific point using a coherent collection of related words (Grady, 1971).

5.3 Final/Concluding Paragraphs

Our findings illustrate concluding paragraphs as containing fewer words and less given information than body paragraphs. Concluding paragraphs contain more content word overlap, more positive logical connectives and less meaningful words that are less imagable when compared to introductions. Concluding paragraphs also contain words that are less specific and more familiar than introductory and body paragraphs. This linguistic description of concluding paragraphs fits well with the paragraph's rhetorical purpose: to summarize the information in the essay without presenting new information (Grady, 1971). The process of summarization likely produces shorter paragraphs containing more connectives. The words used in this summarization will be less specific in nature, but more familiar. These characteristics are illustrated in the following paragraph taken from our corpus.

> In conclusion, I think it is fair to say that there are some men in this world that are treated more equal than others. Whether it is based on one's status or material possessions, equality should have nothing to do with that. It should not be that way because all men are created equal and deserve as fair of a chance as anyone else. Maybe one day people in this world realize that, but we, unfortunately, have quite a long way to go.

This concluding paragraph acts as a summarization of a main idea and supporting arguments. The paragraph is shorter than the body paragraphs in our corpus but longer than the introductions. The words in this conclusion are also less specific and less imagable, but familiar (i.e., *men, world, status, possessions*) affording greater summarization. They also permit for fewer associations with other words (i.e., *equal, status, chance*) thus better closing the essays by invoking fewer ideas. While this example does not contain many positive, logical connectives, it does contain many connectives and sentence modifiers that embed ideas within the summary such as *in conclusion, because,* and *but*.

In general, our results demonstrate that significant differences in the linguistic features reported by Coh-Metrix exist among paragraph types in the argumentative essays found in our corpus. More importantly, as the quality and length of the paragraph increases, the linguistic differences among paragraph types appears to become more acute. This finding lends support to the notion that higher quality paragraphs types are more easily distinguishable from one another. Thus, we have increased confidence that our findings may extend to more advanced writers.

Although we have demonstrated that linguistic differences permit paragraph classification using statistical modeling, we note that not all paragraphs are equally classifiable. The findings of this study demonstrate that introductory and body paragraphs demonstrate higher rates of classification accuracy than concluding paragraphs. This finding is likely related to the summarization purpose of concluding paragraphs. Human raters were more likely to misclassify conclusions as introductory paragraphs, perhaps reflecting the commonality of restating the thesis and the supporting arguments in concluding paragraphs. Our statistical model was more likely to misclassify concluding paragraphs as body paragraphs, perhaps reflecting the linguistic similarities between the two paragraph types (incidence of connectives, word meaningfulness, and word imagability).

From a pedagogical perspective, our findings also illuminate patterns of discourse that may prove beneficial in the classroom, especially for less experienced writers who may benefit from a set form that allows them to scaffold into more rhetorically situated writing. Our study provides supporting evidence for the division of paragraph types based on individual linguistic features. These features could be developed into teaching heuristics that would permit students to better understand the rhetorical roles of paragraph types through the linguistic features contained within those paragraphs. In detecting linguistic relations that help bind paragraph types together, we are also able to provide a more complete description of writing. This description could be relayed to student writers to help them better organize essays and more fully understand the links

between linguistic features and paragraph types. Such organizational schemes serve the developmental needs of writers (Haswell, 1986) who benefit from a deeper understanding of paragraph types, their content, and their purpose. Lastly, automatic detection of paragraph types could inform feedback mechanisms in intelligent tutoring systems. The ability to computationally predict the type of paragraph produced by writers would prove beneficial in automatically assessing essays for expected structural and linguistic patterns and using this assessment to provide real-time feedback.

6. CONCLUSION

We argue, in a similar manner as D'Angelo (1974), that studies of form and structure in composition benefit from examinations of texts that are not impressionistic, but quantitative, revealing relations within texts that afford a more complete description of writing. Knowing that an essay has a beginning, a middle, and end is important, but understanding how these sections differ affords us the opportunity to not only understand the rhetorical outline of argumentative essays, but provides examples of paragraph differences for use in student instruction. Researchers, teachers, and students will thus have a better understanding of how discourse functions in extended units of text and how these units influence essay organization. Because our examination of paragraph types permits specific distinctions and supports these distinctions through statistical analysis, we can say with increased certainty that these distinctions are tangible and, as a result, practical.

We see this study as a step forward in further clarifying the structure and function of paragraph types in argumentative essays. We note that while our findings were significant, we were only able to classify 72% of the paragraph types in the test set. Even though the accuracy of our model was on a par with human judgments of paragraph types, it was, of course, imperfect. Much of the missing accuracy is likely explained by our use of student writers who may not attend to expected rhetorical patterns or may not produce paragraphs of high enough quality to classify with a high degree of accuracy. We also leave open the possibility that linguistic features not considered in this study could also increase our classification accuracy and potentially identify other characteristics of paragraphs not revealed in this study. Thus, as computational tools develop and expand, replication studies of this approach are warranted to examine linguistic features other than those reported by Coh-Metrix. We also accept the notion that not all paragraph types may be classifiable based on linguistic features. Much like studies examining topic sentences (cf. Popken, 1987, 1988), it is likely that not all paragraphs conform to specific patterns that permit classification. Lastly, the study only investigated paragraph types in argumentative essays as produced by American students in a university located in the United States. We have no evidence that the paragraph distinctions developed in this study are not culturally or rhetorically specific and thus give no assurance that the distinctions are generalizable outside of the sampled population.

Notes

1. Our working definition of a paragraph for this study is strongly influenced by studies conducted in the United States. Thus, we do not claim that the notion of a paragraph that we examine in this paper is universal in nature, but rather specific to U.S. writing communities.
2. Grady used the term sequences rather than paragraphs.
3. Our analysis used linguistic indices computed by one tool only (Coh-Metrix). Future analyses using different indices may reveal other linguistic features that also discriminate paragraph types.

Acknowledgments

This research was supported in part by the Institute for Education Sciences (IES R305A080589 and IES R305G20018-02). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IES. We are indebted to both Art Graesser and Zhiqiang Cai for their help in developing Coh-Metrix. We are also indebted to our expert raters Jen Weston, Brad Campbell, and Daniel White and to Jamal Williams for his assistance in organizing the data. Lastly, we thank Philip McCarthy for his assistance in early versions of this study.

References

Albertson, B. R. (2007). Organization and developmental features of grade 8 and grade 10 writers: A descriptive study of Delaware Student Testing Program (DSTP) essays. *Research in the Teaching of English, 41* (4), 435-465.

Arnaudet, M. L., & Barrett, M. E. (1990). *Paragraph development: A guide for students of English*. New Jersey: Prentice Hall Regents.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (Release 2)*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Becker, A. L. (1965). A tagmemic approach to paragraph analysis. *College Composition and Communication, 16*, 237-242. doi: 10.2307/355200

Biber, D. (1988). *Variations Across Speech and Writing*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511621024

Biber, D. (1993). Representativeness in corpus design. *Computational Linguistics,* 19, 243-257. doi: 10.1093/llc/8.4.243

Bond, S. J., & Hayes, J. R. (1984). Cues people use to paragraph text. *Research in the Teaching of English, 18* (2), 147-167.

Brace N., Kemp, R., & Snelgar, R. (2003). *SPSS for Psychologists*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Breetvelt, I., van den Bergh, H., & Rijlaarsdam, G. (1996). Rereading and generating and their relation to text quality. An application of multilevel analysis on writing process data. In G. Rijlaarsdam, H. van den Bergh, & M. Couzijn (Eds.), *Theories, models & methodology in writing research* (pp. 10–21). Amsterdam: Amsterdam University Press.

Brindley, R., & Schneider, J. J. (2002). Writing instruction or destruction: Lessons to be learned from fourth-grade teachers' perspectives on teaching writing. *Journal of Teacher Education*, *53*, 328-341. doi: 10.1177/0022487102053004005

Chafe, W. L. (1975). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In C. N. Li (Ed.), *Subject and Topic* (pp. 26-55). New York: Academic.

Christensen, F. (1965). A generative rhetoric of the paragraph. *College Composition and Communication, 16*, 144 – 156. doi: 10.2307/355728

Cohan, C. (1976). Writing effective paragraphs. *College Composition and Communication*, *27*, 363-365. doi: 10.2307/356298

Coltheart, M. (1981). The MRC psycholinguistic database quarterly. *Journal of Experimental Psychology*, *33*, 497-505. doi: 10.1080/14640748108400805

Costerman, J. & Fayol, M. (1997). *Processing Interclausal Relationships: Studies in Production and Comprehension of Text*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Crismore, A, Markkanen, R., & Steffensen, M. (1993). Metadiscourse in persuasive writing. *Written Communication*, *10*, 39-71. doi: 10.1177/0741088393010001002

Crossley, S. A. & McNamara, D. S. (2009). Computationally assessing lexical differences in L1 and L2 writing. *Journal of Second Language Writing, 18,* 119-135. doi: 10.1016/j.jslw.2009.02.002

Crossley, S. A. & McNamara, D. S. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 984-989). Austin, TX: Cognitive Science Society.

Crossley, S. A., & McNamara, D. S. (in press). Predicting second language writing proficiency: The role of cohesion, readability, and lexical difficulty. *Journal of Research in Reading.*

Crossley, S. A., McNamara, D. S., Weston, J., & McLain Sullivan, S. T. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication, 28* (3), 282-311. doi: 10.1177/0741088311410188

D'Angelo, F. (1986). The topic sentence revisited. *College Composition and Communication*, 37(4), 431-441. doi: 10.2307/357913

Dufty, D., Hempelmann, C., Graesser, A., Cai, C., & McNamara, D. S. (2005). An algorithm for detecting causal and intentional information in text. *Presentation at the 15th Annual Meeting of the Society for Text and Discourse*, Amsterdam.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: The MIT Press.

Field, A. (2005). *Discovering Statistics Using SPSS*. London: Sage Publications.

Field, J. (2004). *Psycholinguistics: The Key Concepts*. New York, Routledge.

Gilhooly K. J., & Logie, R. H. (1980). Age of acquisition, imagery, concreteness, familiarity and ambiguity measures for 1944 words. *Behavior Research Methods and Instrumentation*, 12, 395–427. doi: 10.3758/BF03201693

Grady, M. (1971). A conceptual rhetoric of the composition. *College Composition and Communication, 22,* 348-354. doi: 10.2307/356208

Graesser, A.C., McNamara, D.S., Louwerse, M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*, 193-202. doi: 10.3758/BF03195564

Haberlandt, K., & Graesser, A. C. (1985). Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology: General*, *114*, 357-374. doi: 10.1037/0096-3445.114.3.357

Halliday. M. A. K. (1967). Notes on transitivity and theme in English. *Journal of Linguistics, 3*, 199-244. doi: 10.1017/S0022226700016613

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

Haswell, R. H. (1986). The organization of impromptu essays. *College Composition and Communication*, *37* (4), pp. 402-415. doi: 10.2307/357911

Hempelmann, C. F., Dufty, D., McCarthy, P. M., Graesser, A. C., Cai, Z., & McNamara, D. S. (2005). Using LSA to automatically identify givenness and newness of noun phrases in written discourse. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 941-946). Mahwah, NJ: Erlbaum.

Johnson, T. S., Thompson, L., Smagorinsky, P., and Fry, P. G. (2003). Learning to teach the five-paragraph theme. *Research in the Teaching of English*, *38*, 136-176.

Just, M. A., Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*, 329–354. doi: 10.1037/0033-295X.87.4.329

Just, M. A., & Carpenter, P. A. (1987). *The Psychology of Reading and Language Comprehension*. Boston: Allyn & Bacon.

Hillocks, G. (2002). *The Testing Trap: How State Writing Assessments Control Learning*. New York: Teachers College Press.

Hillocks, G. (2005). The focus on form vs. content in the teaching of writing. *Research in the Teaching of English, 40,* 238-248.

Karrfalt, D. H. (1968). The generation of paragraphs and larger units. *College Composition and Communication*, *19*, 211-217. doi: 10.2307/356044

Kinneavy, J., & Warriner, J. (1993). *Grammar and Composition*. Austin, TX: Holt, Rinehart & Winston.

Kintsch, W., & Keenan, J. (1973). Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology*, *5*, 257-274. doi: 10.1016/0010-0285(73)90036-4

Koen, F., Becker, A. L., & Young, R. E. (1969). The psychological reality of the paragraph. *Journal of Verbal Learning and Verbal Behavior, 8,* 49-53. doi: 10.1016/S0022-5371(69)80010-1

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-140.doi: 10.1037/0033-295X.104.2.211

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes, 2*, 259-284.doi: 10.1080/01638539809545028

Longo, B. (1994). Current research in technical communication: The role of metadiscourse in persuasion. *Technical Communication, 41*, 348–352.

Louwerse, M. M. (2001). An analytic and cognitive parameterization of coherence relations. *Cognitive Linguistics*, *12* , 291–315.

Mabry, L. (1999). Writing to the rubric: Lingering effects of traditional standardized testing on direct writing assessment. *Phi Delta Kappan, 80,* 673-679.

McCarthy, P. M., Renner, A. M., Duncan, M. G., Duran, N. D., Lightman, E. J., & McNamara, D. S. (2008). Identifying topic sentencehood. *Behavior Research and Methods, 40,* 647-664. doi: 10.3758/BRM.40.3.647

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication, 27,* 57-86. doi: 10.1177/0741088309351547

McNamara, D.S., Louwerse, M.M., McCarthy, P.M., & Graesser, A.C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes, 47*, 292-330. doi: 10.1080/01638530902959943

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). Five papers on WordNet (Tech. Rep. No. 43). Princeton, NJ: Princeton University, Cognitive Science Laboratory.

Nunnally, T. E. (1991). Breaking the five-paragraph-theme barrier. *The English Journal*, 80, 67-71. doi: /10.2307/818100

Oshima, A., & Hogue, A. (1997). *Introduction to Academic Writing*. London: Longman.

Paivio, A. (1965). Abstractness, imagery, and meaningfulness in paired-associate learning. *Journal of Verbal Learning and Verbal Behavior, 4,* 32-38. doi: 10.1016/S0022-5371(65)80064-0

Pearson, P. D. (1974-75). The effects of grammatical complexity on children's comprehension, recall, and conception of certain semantic relations, *Reading Research Quarterly, 10*, 155-192. doi: 10.2307/747180

Perfetti, C. A. (1985). *Reading Ability*. New York: Oxford University Press.

Perfetti, C. A., Landi, N. & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. J. Snowling & C. Hulme (eds.). *The Science of Reading: A Handbook* (pp. 227-247). Oxford, Blackwell. doi: 10.1002/9780470757642.ch13

Popken, R. L. (1987). A study of topic sentence use in academic writing. *Written Communication, 4,* 209-228. doi: 10.1177/0741088387004002006

Popken, R. L. (1988). A study of topic sentence use in scientific writing. *Journal of Technical Writing and Communication, 18,* 75 – 86. doi: 10.2190/0QUV-78W9-E86P-W3ML

Rayner, K. & Pollatsek, A. (1994). *The Psychology of Reading*. Englewood Cliffs, New Jersey: Prentice Hall.

Salsbury, T., Crossley, S. A., & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research, 27* (3), 343-360. doi: 10.1177/0267658310395851

Sanders, T., & Schilperoord, J. (2006). Text structure as a window on the cognition of writing; How text analysis provides insights in writing products and writing processes. In C. A. MacArthur, S. Graham & J. Fitzgerald (Eds.), *The Handbook of Writing Research* (pp. 386 - 402). NY: Guilford Publications.

Stern, A. (1976). When is a paragraph? *College Composition* and *Communication*, *27*, 253-257. Swales, J. M. (1991). *Genre Analysis*: *English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Tabachnick, B., & Fidell, L. S. (1996). *Using Multivariate Statistics*. New York: Harper Collins College Publishers.

Toglia, M. P., & Battig, W. F. (1978). Handbook of semantic word norms. Hillsdale, NJ: Erlbaum.

Van de Kopple, W.J. (1985). Some exploratory discourse on metadiscourse. *College, Composition and Communication*, *36*, 82-93. doi: 10.2307/357609

Warner, R. (1979). Teaching the paragraph as a structural unit. *College Composition and Communication*, *30*, 152-155. doi: 10.2307/356320

Warriner, J. E. (1958). *English Grammar and Composition 10.* New York: Harcourt Brace and World, Inc.

Witten, I. H. & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques (2nd ed.)*. San Francisco: Morgan Kaufman.

Young, R. E., & Becker, A. L. (1966). The role of lexical and grammatical cues in paragraph recognition. *Studies in Language and Language Behavior, 11* (2), 1-6.