Determining Paragraph Type From Paragraph Position

Kyle B. Dempsey, Philip M. McCarthy, John C. Myers,
Jennifer Weston, & Danielle S. McNamara

Institute for Intelligent Systems
University of Memphis
Memphis, TN 38152
{kdempsey, pmmccrth, jcmyers, jen.weston, dsmcnamr} @ memphis.edu

Abstract

Students must be able to competently compose essays in order to succeed in school and progress into the workplace. Current intelligent tutoring systems (ITS) attempt to provide individual training that is lacking in the current educational system. To provide efficient individual training through ITS, the systems must be able to effectively assess writing input from students. Necessary components for computer-based writing tutors are algorithms that mimic human judgments of writing. The current study attempts to establish a connection between paragraph position and human ratings of paragraph type through the use of computational measures provided by Coh-Metrix. We find that expert raters do not easily identify paragraph type and ratings of paragraph type do not map onto paragraph position.

Introduction

Writing is essential to being prepared for college and ready for job training (Light 2001; National Commission on Writing 2004). Nonetheless, students continue to fail to write at a proficient level. To exacerbate the problem, students are unable to get the one-on-one help that they need to improve their writing abilities, because tutoring is typically unavailable in public schools, and teachers are increasingly required to deal with larger class sizes. As the ratios of teachers to students change for the worse, students are required to spend more time in autonomous learning situations.

One approach to alleviating the problem of teacher time and availability is through the use of computer technologies. For example, Intelligent Tutoring Systems (ITS) such as iSTART (McNamara, Levinstein & Boonthum 2004) can provide students with one-on-one training, without requiring valuable teacher time. By interacting with these ITSs, students gain critical and otherwise unavailable knowledge and experience.

Our team at the University of Memphis is currently developing an ITS that targets writing strategy instruction, called the Writing-Pal (W-Pal; McNamara et al. 2007). W-Pal is an ITS designed to provide instruction on seven writing strategies to high school students. Writing strategies not only help writers alleviate working memory demands, but these strategies also help individuals activate long-term working memory (McNamara & Scott 2001). Aside from reducing working memory demands, writing strategies keep the writer focused on the writing process.

The seven W-Pal strategies comprise three phases of the writing process: Prewriting, Drafting, and Revising. The current research focuses on the W-Pal Drafting phase strategies related to Introduction building and Conclusion building. The purpose of the strategies in these modules is to train writers to develop appropriate introductions and conclusions through a process-driven approach.

Introductions and conclusions are of particular interest given their prominent organizational role in persuasive essays. Introductions play the role of bringing the reader into the essay's context and establishing the principal arguments of the essay. Conclusions remove the reader from the essay context while restating the main points of the essay. We are specifically interested in this strategy in the context of ITSs (such as W-Pal), because identifying when a student is composing specific types of paragraphs can be useful in shaping the content and dialogue in the same manner as a human tutor. By establishing computational markers of paragraph type, W-Pal will be able to provide more facilitative feedback to students.

In turn, these feedback algorithms and evaluations will further develop the field of *natural language assessment and understanding* (Rus et al. 2008). The interplay between writing assessment and computer technology leads to the development of computational algorithms for identifying and assessing writing. One such tool that facilitates real-time computational assessment is Coh-Metrix (Graesser et al. 2004). Coh-Metrix is a powerful computational tool that provides multiple measures of cohesion, readability, lexicon use, part-of-speech classifiers, syntactic parsers, and several others that are widely used in computational linguistics. Our goal is to use Coh-Metrix to identify cohesion markers of paragraph type. We intend not only to distinguish between paragraph types, but also to distinguish gradations of quality within those paragraphs.

One potential source for paragraph type identification is paragraph position. When writing persuasive essays, students are taught to write a prototypical five paragraph essay: an introduction followed by three supporting body-paragraphs, and a conclusion. If writers are indeed ordering their writing as introductions first, body-paragraphs second, and conclusions last, then we can predict that a computational algorithm for identifying paragraph order will also identify paragraph type. We will examine this possibility in two related experiments. First, we use Coh-Metrix to determine whether there is a set of variables that identifies paragraph position in persuasive essays. We then conduct a second experiment to determine whether the same variables are indicative of paragraph type. If the variables also predict paragraph type, then we have evidence that students are indeed ordering their writing as introductions first, body-paragraphs second, and conclusions last. However, if the variables are not indicative of paragraph type, students may be writing in a different manner.

Experiment 1

Experiment 1 examines the effect of paragraph position (first, middle, or last) on computational markers within persuasive essays. Specifically, certain computational markers are expected to be either more or less prevalent depending on the position of the paragraph in a persuasive essay. We predict differences between middle and other paragraphs as well as distinct differences between first and last paragraphs.

Corpus

The current study uses a corpus of essays written by students at a university in the southern United-States. The essays were written as responses to specific prompts designed to mimic College Board prompts. The prompts present a dichotomous argument and require the writer to take a position to complete the assignment. Writers were not given instructions on how to write an effective essay.

The 1348 paragraphs' serial position was tagged creating three groups: first paragraphs (n = 270), middle paragraphs (n = 785), and last paragraphs (n = 293). The paragraph in essays containing on a single paragraph were coded as last paragraphs. The paragraphs were processed using Coh-Metrix and the scores served as the dependent variables in a series of Analyses of Variance (ANOVA). Based on the results of the ANOVAs, a discriminant analysis was conducted to determine the computational properties of paragraph order in persuasive essays. Discriminant analysis is a statistical procedure that is used to find the predictability of a dependent variable (group) based on a body of independent variables (Coh-Metrix results). All discriminant analyses performed in the current study were conducted using *training* and *test* set data. The training set is a random two-thirds sample of the corpus used to create an initial model. The accuracy of the model created with the training set is assessed with the test set. By assessing the model through training and test sets, we reduce the likelihood of overfitting of the model and increase its generalizability.

Table 1: Discriminant Analysis 1

		First (0.20)	Middle (0.58)	Last (0.22)
Test set	R	0.57	0.65	0.53
	P	0.48	0.65	0.53
	F1	0.52	0.71	0.48
Training Set	R	0.45	0.63	0.52
	P	0.41	0.75	0.38
	F1	0.43	0.69	0.44
All	R	0.53	0.65	0.53
	P	0.46	0.77	0.42
	F1	0.49	0.70	0.47

Note: Baselines in parentheses; R=Recall; P=Precision.

Results and Discussion

Using the training set data, we conducted a series of ANOVAs to identify candidate variables that best distinguished the three paragraph groups (*first*, *middle*, *last*). Correlations were calculated to reduce the effects of colinearity. For variables with a correlation greater than or equal to $r = 0.7$, the variable with the lower F value was eliminated. After this process, 24 variables remained. These 24 variables were used to conduct three discriminant analyses. The four most highly significant variables, and their means, are presented in Table 1.

The first discriminant analysis was conducted to discover the predictability of the grouping variable as a function of the 24 Coh-Metrix indices. The results of the first discriminant analysis suggested that linguistic features within the *first*, *middle*, and *last paragraphs* differ sufficiently to distinguish them from one another (see

Table 2: Discriminant Analysis 1 (Across All Paragraph Type)

	First	Middle	Last	F(2, 896)
Type token ratio	1.70 (2.26)	0.94 (0.78)	1.89 (2.36)	31.33
Number of Word Types	48.31 (21.33)	68.49 (28.45)	56.95 (43.16)	31.28
3rd person pronouns	0.01 (0.02)	0.02 (0.03)	0.01 (0.02)	27.49
Content Word Frequency	0.30 (0.49)	0.10 (0.29)	0.28 (0.49)	25.59
First person pronouns	0.04 (0.04)	0.03 (0.04)	0.05 (0.04)	22.71

Note: SD in parentheses. All variables sig. at $p<.001$.

Table 1). Discriminant analysis results are interpreted using *recall*, *precision*, and *F1*. The most significant markers or paragraph position were type-token ratio and number of word types. These markers may be distinguishing differences in paragraph position consistent with more varied word choice in middle paragraphs, but not in last paragraphs. Essentially, writers are likely using more complex language in middle paragraphs to expand on information presented in initial paragraphs.

Table 3: Discriminant Analysis 2

		First or Last (0.44)	Middle (0.58)
Test set	R	0.68	0.73
	P	0.65	0.76
	F1	0.67	0.74
Training Set	R	0.69	0.75
	P	0.64	0.78
	F1	0.66	0.77
All	R	0.68	0.74
	P	0.65	0.76
	F1	0.67	0.75

Note: Baselines in parentheses; R=Recall; P=Precision.

The second discriminant analysis was similar to the first; however, the first and last paragraphs were collapsed into a single category. The second analysis used 27 variables to distinguish *middle paragraphs* from all other paragraphs (*first*, *last*). This analysis examined the question of whether middle paragraphs were uniquely distinguishable from first and last paragraphs. First and last paragraphs tend to establish and summarize information, whereas middle paragraphs tend to elaborate on information previously presented. The results of this analysis, like the previous analysis, confirmed that linguistic features are significant distinguishers of paragraph order (see Table 3 and Table 4). Table 3 shows that the ability of the linguistic features to distinguish between middle and other paragraphs is quite high, and much better than when the three types are included in the same analysis (i.e., Table 2). Table 4 shows that reduced diversity in word choice appears in middle paragraphs.

Table 5: Discriminant Analysis 3

		First (0.44)	Last (0.58)
Test set	R	0.71	0.59
	P	0.61	0.69
	F1	0.66	0.64
Training Set	R	0.83	0.77
	P	0.78	0.82
	F1	0.80	0.80
All	R	0.75	0.65
	P	0.66	0.74
	F1	0.70	0.69

Note: Baselines in parentheses; R=Recall; P=Precision.

These analyses imply that it may be difficult to distinguish between first and last paragraphs. One possibility is that their structures are similar, and thus cannot be distinguished. To test this possibility, a third discriminant analysis (including 28 variables that emerged from the ANOVAS) was conducted to distinguish between the *first* and *last paragraphs*. The results indicate that indeed there are distinguishing linguistic characteristics between the *first* and the *last paragraphs* (see Tables 5 and 6). Whereas decreased diversity of word choice was a marker of middle paragraphs to first and last paragraphs, more specific markers distinguish first and last paragraphs. For example, first paragraphs tend to contain fewer modal verbs (e.g. *should*, *must*, *can*) and verb phrases than do middle paragraphs possibly indicating that first paragraphs contain more verbs and modified verbs. This may also arise due to hedging in predictive language in last paragraphs of persuasive essay writing (i.e. *should mean*, *could improve*, *might predict*, etc.).

Overall, the results across the three analyses indicate that students are producing computationally distinguishable

Table 4: Discriminant Analysis 2 (Middle Paragraphs versus Other Paragraphs)

	Middle	First or Last	F(1,447)
Type Token Ratio	0.92 (0.71)	1.50 (1.48)	30.22*
Number of Word Types	67.93 (29.38)	54.04 (27.35)	25.38*
Word Information	397.97 (29.06)	384.53 (27.00)	24.35*
LSA Sentence to Sentence adjacent SD	0.12 (0.08)	0.08 (0.08)	23.41*

Note: SD in parentheses. All variables sig. at $p<.001$.

Table 6: Discriminant Analysis 3 (First versus Last Paragraphs)

	First	Last	F(1,561)
Modal verbs	17.04 (17.95)	26.43 (22.03)	30.41
Verb phrase incidence	222.41 (58.37)	248.40 (60.47)	26.83
Noun phrases	363.10 (70.75)	335.11 (59.84)	25.81
LSA Sentence to Paragraph SD	0.11 (0.08)	0.08 (0.08)	21.89
Incidence of Positive Connectives	10.25 (15.96)	17.57 (21.52)	20.73

Note: Standard deviations are in parentheses; All results $p<.001$.

paragraphs from the first to the middle to the last paragraphs (Tables 1 & 2). Specifically, the first two discriminant analyses show that middle paragraphs have a significantly lower type-token ratio than first and last paragraphs where first and last paragraphs contained more varied word choice than did middle paragraphs. Although type-token ratio is confounded by number of words in the text (McCarthy & Jarvis 2007), writers are producing the paragraphs in a natural setting, therefore the metrics are indicative of valid differences.

We also demonstrated a difference between first and last paragraphs in their use of specific linguistic markers. We found, among other indicators, that the part of speech modality occurrences, verb phrase incidence, and positive connectives (e.g. *and*) are more common in last paragraphs, whereas part of speech occurrence per noun phrase and LSA sentence to paragraph measures are more common in first paragraphs. This is an expected trend of differences due to the need to establish information (noun phrase incidence) before manipulating the information later in the essay (verb phrase incidence; positive connectives).

Experiment 2

The results of Experiment 1 indicate that there are computationally distinguishable differences between first, middle, and last paragraphs. Thus, the second experiment was conducted to determine whether the first, middle, and last paragraphs can be differentiated by humans who are not provided with information concerning the order of the paragraphs. Specifically, we used the linguistic indices established in Experiment 1 to determine whether first, middle, and last paragraphs would be identified by humans as introductions, body-paragraphs, and conclusions.

Using the established predictive linguistic indices for paragraph order identified in Experiment 1, we examined whether paragraph order corresponds to the expected content within introductions, body-paragraphs, and conclusions. We expected paragraph type to be statistically similar to paragraph position (first, middle, last). Specifically, we predicted that paragraphs rated as introductions, body-paragraphs, and conclusions would be computationally identified by the same linguistic indices that successfully identified first, middle, and last paragraphs, respectively. We further predicted differences across paragraph type for cohesion indices and word frequency measures. Cohesion indices, such as *argument overlap*, would likely be higher in body-paragraphs because body-paragraphs have a smaller lexical and conceptual focus than do introductions and conclusions. We also predicted that lower frequency words would be more prevalent in body-paragraphs. We expected these results to emerge because the introduction is often more vague and often uses more generalized language (higher frequency words), whereas body-paragraphs would tend to include more specific (lower frequency) words to expand upon those ideas. In essence, we expect most writers to use a general to specific approach to writing the essays.

The paragraphs in Experiment 1 were coded for serial paragraph position within their respective essays. For further investigation, we examined the question of whether the effects found within the groups established in Experiment 1 held true when the paragraphs were separated into groups based on expert ratings of paragraph type. A sample of the first (n = 105), middle (n = 291), and last (n = 101) paragraphs used in Experiment 1 were rated as introductions (n = 65), body-paragraphs (n = 307), and conclusions (n = 91) based upon the coding scheme described below. The first, middle, and last paragraphs are typically believed to be introductions, body-paragraphs, and conclusions. To empirically study this comparison, a one-third random sample of the paragraphs used in Study 1 (n = 497) was coded by two expert raters.

Coding Scheme

The expert-rater coding scheme takes into account multiple characteristics of paragraph type (i.e. thesis statements, grammar, etc.) to make a decision on both a yes/no level (judged on a binary scale) and a quality level (judged on a 6-point scale). The current study is only concerned with the binary rating for paragraph type. Expert raters scored all paragraphs at an inter-rater reliability of r = .72 level of agreement. Rater differences were resolved through discussion. Ambiguous paragraphs that were coded into multiple categories were resolved through discussion. If expert raters could not detect a difference in type, the paragraph was excluded from the analysis.

Although the inter-rater reliability may seem to be low or moderate compared to other studies that have reported IRR, the reliability is a reflection of the difficulty of the task. The raters in the study received extensive training on the coding scheme and completed ratings of training sets before rating the paragraphs individually. Essentially, the reliability reported should be viewed as a benchmark of reliability in for this type of judgment rather than being compared to results for other types of judgments..

Results and Discussion

We conducted ANOVAs using the Coh-Metrix variables identified as acceptable paragraph type markers in the three Experiment 1 discriminant analyses. The analyses indicated that none of these variables were significant markers of introductions, body, or conclusion paragraphs. We found that the paragraph order predictors were not significant predictors of paragraph type. This finding likely indicates that what students are writing is something different from introductions, body-paragraphs, and conclusions. Because none of the variables were significant predictors of human’s judgments of paragraph type, we focus on answering the follow-up question of which variables *do* mark introductions, body-paragraphs, and conclusions.

To determine if any of the available variables mark paragraph type, human ratings based on the paragraph coding scheme previously described (*introduction*, *body*, *conclusion*) were compared with the full set of Coh-Metrix variables used in Experiment 1. We conducted ANOVA in order to determine whether humans use the linguistic information (Coh-Metrix indices) to judge paragraph type that the computational algorithm used to identify paragraph order. Of the 374 Coh-Metrix indices in this analysis, only four of the indices reached an F level of greater than 3.00 in the comparison between the human paragraph ratings (see Table 7). Variables with significant inter-correlation were omitted due to redundancy.

The results indicate that introductions contain more content words per sentence as well as a significantly higher scores on the Paivio Norm per sentence meaningfulness measure than both body-paragraphs and conclusions. Conclusions have a significantly higher word rating on the Colorado Norm per word meaningfulness measure than body-paragraphs and a significantly lower incidence of negative causal connectives (e.g. *although*) than both introductions and body-paragraphs. These findings indicate that these writers use fewer content words as well as less meaningful words in body-paragraphs, and use fewer negative causal connectives in conclusions.

Based upon the indices in Table 7, we can also address our earlier hypotheses about cohesion and word frequency. First, we did not find any cohesion indices to be significantly different across paragraph type. Second, as predicted, we found that introductions contain more general language than do body-paragraphs (p = .007). Based upon these findings, along with the findings from Experiment 1, we conclude that first, middle, and last paragraphs cannot be assumed to be introductions, body-paragraphs, and conclusions, respectively.

The results from Experiment 2 are also likely indicative of a larger trend. The finding that expert raters were able to agree on a holistic judgment of paragraph type leads us to believe that the paragraphs did have qualities of the paragraph type, but not consistent qualities. Instead, the results from Experiment 1 can be combined with the results from Experiment 2 to show that students are writing computationally distinguishable paragraphs as they progress through their writing, but those paragraphs are not consistently introductions, body-paragraphs, and conclusions. ITS and teaching in general can benefit from this finding by reducing heuristic reliance on paragraph position for identification of paragraph type, while also emphasizing to writers the importance of adhering to this heuristic.

The results from Experiment 2 indicate that there are specific word frequency differences across paragraph type, but no differences in cohesion across paragraph type. The differences in word frequency indicate that the language being used in body-paragraphs is more specific than the language being used in introductions. The finding that lower frequency words are being used in body-paragraphs supports the idea that generalized language is indicative of introductions. This finding also supports the ratings made by the expert judges. We expected to find that body-paragraphs had more cohesive text than both introductions and conclusions, but did not find a difference across paragraph type. The finding that there are no differences in cohesion across paragraphs, combined with the finding that body-paragraphs contain more specific language than introductions, suggests that while specific language use may be different, the cohesion remains consistent. As such, cohesion may be more a function of the writing style overall, rather than a function of the purpose of a specific paragraph.

As a final caveat, we note that some of the results based upon the means and standard deviations observed in both experiments suggest that results should be interpreted with caution. In multiple cases, the standard deviations are higher than the means, indicating a diverse sample. The differences observed are a concern, and will be the focus of future studies.

Conclusions

We can draw three main conclusions based on the observed results of Experiment 1 and Experiment 2. First, when given a persuasive essay prompt, paragraph position can be identified computationally, but the computational indices for paragraph position do not identify paragraph type. We interpret these findings as first, middle, and last paragraphs not necessarily being introductions, body-paragraphs, and conclusions, respectively.

These results are useful for three reasons. We have identified a set of variables that may account for paragraph position in essays written from a persuasive prompt. We have empirically determined that variables such as type token ratio and part of speech occurrence computationally

Table 7: Means and F Scores for Paragraph Type

	Introduction	Body	Conclusion	F(2,466)
Content word frequency CELEX	1380.74a (261.51)	602.46b (120.33)	652.84b (221.02)	3.72**
Meaningful Colorado word	106.73ab (1.36)	104.35a (0.62)	107.42b (1.15)	3.42**
Negative causal connectives	2.85a (0.67)	1.86a (0.31)	0.70b (0.56)	3.13**
Meaningful Paivio per sentence	16.65a (5.81)	2.07b (2.67)	0.00b (4.91)	2.96*

Note: Standard Errors in parentheses. **p=0.05,* **p<0.05.; Statistically different relationships indicated by differing letter annotation within each row.

distinguish paragraph position. Establishing this type of distinction is particularly useful in the automation of text identification tools, where tools that can identify text independent of the full essay context can potentially identify text that is out of place or poorly organized. We have also established a set of linguistic features for paragraph type from essays written based on a prompt. Specifically, something as simple as frequency of content word use is shown to be a computational marker of paragraph type. This finding can be used in future work to distinguish paragraphs for programs in need of algorithms that quickly identify paragraphs as they are written. By identifying the paragraphs in real time, programs such as W-Pal can provide more facilitative and appropriate feedback for the users without having to involve human raters. Instead, human interaction can be reserved for more deep level processes to enhance learning.

Second, we found that there is a difference between paragraph type in terms of word frequency, but not in terms of cohesion. Writers use more complex word choice in body-paragraphs than in introductions. The language complexity is indicative of the need to expand on ideas initially presented in introductions with more complex word choice coming from an expansion on general ideas. Also, there is no difference in cohesion across paragraph type. This similarity in cohesion across paragraphs contrasts with conclusion regarding word frequency (writers are not expanding on ideas previously presented. These two findings together indicate that writers are not linking ideas between introductions and body-paragraphs. Instead, the writers are simply making general statements about the prompt in the introduction and then presenting their more specific ideas in the body-paragraphs.

Third, we did not find a link between paragraph position and paragraph type. We see the absence of a link as a positive identification that students are writing two different styles of paragraphs. Notably, the essays were written from a persuasive essay prompt. This raises the possibility that students wrote their essays using a narrative style, which produces distinguishable paragraph order as concepts are presented and stacked as the essay continues, but does not involve the production of expected paragraph types as would be expected within a persuasive essay.

As a caveat, the quality of the students' writing not considered in this study. It is possible that the writers are producing poor introductions, body-paragraphs, and conclusions, but still being judged as a specific type of paragraph. This potential failure to produce acceptable paragraph types is further support for programs such as W-Pal that train basic writers to include essential aspects of an essay and prepare them to write both as a college entrance necessity and a job market necessity. Nonetheless, it is also possible that our expert raters simply have a very difficult task of categorizing paragraphs that have been removed from the context of the whole essay. We can ask the raters to categorize the paragraphs while reading the whole essay, but without the proper context, the ratings suffer. However, if the readers are simply not writing introductions, body-paragraphs, and conclusions, then the main finding is that students are not writing three different types of paragraphs in a systematic manner. Further research will explore this issue by examining the paragraphs in context and also by examining other potential markers of paragraph type. While much work remains to be done, the research presented here offers a major step towards identifying a link between paragraph position and paragraph type.

Acknowledgements

This research was supported by the Institute for Education Sciences (IES R305A080589; IES R305G020018-02). 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.

References

Ericsson, K. A., & Kintsch, W. 1995. Long-term working memory. *Psychological Review*, **102**: 211–245.

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

Light, R. J. 2001. *Making the Most of College: Students Speaking Their Minds.* Cambridge: Harvard University Press.

McCarthy, P. M., & Jarvis, S. 2007. A theoretical and empirical evaluation of *vocd. Language Testing.* **24**: 459-488.

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.*

McNamara, D. S., Levinstein, I. B., & Boonthum, C. 2004. iSTART: Interactive strategy training for active reading and thinking. *Behavioral Research Methods, Instruments, and Computers,* **36**: 222 - 233.

McNamara, D.S., McCarthy, P.M., Kim, L., & Graesser, A. 2007. The Writing Pal: An automated tutoring system that provides interactive writing strategy training. Grant proposal funded by the Institute of Education Sciences.

McNamara, D.S., & Scott, J.L. 2001. Working memory capacity and strategy use. *Memory & Cognition,* **29**: 10-17.

National Commission on Writing. 2004, September. *Writing: A ticket to work… or a ticket out: A survey of business leaders*. Available www.collegeboard.com

Rus, V., McCarthy, P.M., McNamara, D.S., & Graesser, A.C. 2008. A study of textual entailment. *International Journal of Artificial Intelligence Tools,* **17**: 659-685.